COMPARATIVE AND FUNCTIONAL ANALYSIS OF THE *Msx*-1 PROXIMAL REGULATORY REGION

David J. Townley

Ph.D.

University of Edinburgh

MRC Human Genetics Unit

1994

DECLARATION

I declare that:

a) this thesis has been composed by myself, and

b) the work contained within is my own unless otherwise stated

Acknowledgements

This project was funded by an MRC studentship, for which I am most grateful

I owe a great debt to many people, both past and present, at the MRC HGU (and elsewhere in Edinburgh) for making the time I have spent here both fruitful and enjoyable. In particular my thanks go to:-

Bob Hill, for the opportunity to work in his lab at the MRC HGU, for putting up with me *all* this time, for reading the final end product and helping to shape it and for all his help and advice and the many things that I have learned from him in my time as his student.

Nick Hastie, for help in obtaining an extension to my studentship and for providing further extension from his own resources. Also for his frequent advice and wisdom regarding many aspects of science and life.

Richard Meehan for considerable amounts of advice on *in vitro* DNA-binding studies, and for generous provision of protein preps.

Peter Budd for B16 cells and advice on the gel-retardation assay

Justin Grindley and Ailsa Townley for significant statistical advice

Duncan Davidson for answering many queries on embryology and for stimulating discussions on limb development

Norman, Sandy and Douglas in Photography for making professional looking figures from the most mediocre raw-data, and for invaluable assistance in poster construction in the past.

Several of the computer boys, in particular John Ireland, Stuart Wyles and Daniel Cunliffe, for endless help and advice that enabled me to make the most of excellent facilites at the MRC, and for introducing me to MRC and Inveresk hockey.

To Doreen Chambers and Agnes Gallagher for prompt oligo synthesis

To Alasdair MacKenzie for his enthusiasm for the project and his advice and generosity.

The many, and too numerous to mention, members of the Hill lab over the years who have been an endless source of advice, solutions, laughs and bad tapes. They have all contributed to my work, fun and enjoyment.

Above all of these, Vivienne for her support and encouragement (especially latterly), welcome distraction and for dragging me up hills!

Lastly I must thank my parents, Harry and Jeanette Townley for their endless patience and support, both emotional and financial, without which I wouldn't have made it.

This thesis is dedicated to the memory of my Dad, he was my hero.

Abstract

Metazoan embryogenesis is characterised by the large scale cellular movements of morphogenesis and the co-ordinated expression of genes specifying pattern in the resultant structures. In *Drosophila* one family of such genes, the homeobox genes, is involved in some of the key mechanisms effecting these processes. Homeobox genes play a similar role of patterning the body in vertebrate embryos. *Msx*-1 is one member of a family of three vertebrate homeobox genes homologous to the *Drosophila* homeobox gene *msh*. Little is known of the role of the *msh* gene in *Drosophila* however detailed analysis of the expression pattern of *Msx*-1 in the developing mouse has led to the proposal that it is involved, possibly through a role in inductive interactions, in development of the heart, eye, limbs and craniofacial structures. Detection of *Msx*-1 transcripts by RNA *in situ* hybridisation demonstrates that precise spatial and temporal regulation of *Msx*-1 expression is achieved at the transcriptional level. Experiments in the limb show that this regulation responds to positional cues as expected of a gene concerned with patterning the developing body.

Knowledge of the mechanisms by which *Msx*-1 transcription is regulated and identification of the gene products involved is vital to an understanding of the regulatory cascade that patterns the embryo, and to a view of the role of *Msx*-1 in such a system. To elucidate this problem I have analysed the 5'-flanking region of the gene and attempted to identify *cis*-acting DNA regulatory elements close to *Msx*-1. Non-functional parts of the genome are subject to a gradual 'drift' in nucleotide content. Constraints against such change are placed upon the functional regions in the form of selection pressure. I exploit this to identify sequences of potential functional significance within the 5'-flanking DNA of *Msx*-1 by comparing similar, non-coding regions from the mouse and human cognates in a search for conserved sequence. *In vitro* analysis of a short DNA sequence thus identified shows it to be capable of binding proteins in a sequence-specific manner. Comparison of this sequence with the control regions of other genes reveals a similar feature upstream of another homeobox gene, *Hox*d-9, also expressed in the developing limb. Gel-retardation and South-western assays show that both sequences have similar protein-binding properties. A protein bound by all homologous sequences appears to be ubiquitously expressed in the mouse embryo. *In vitro* functional assays provide evidence that the sequence identified is the binding-site of a transcriptional repressor.

It appears that a sequence located upstream of *Msx*-1 and within the *Hox*D cluster may be the binding site for a transcriptional repressor ubiquitously expressed during murine development.

Abbreviations

β-gal	β-galactosidase
°C	degrees Celcius
μM	micromolar
A	adenine
A-P axis	Anterior-Posterior axis
AER	Apical Ectodermal Ridge
AMP	adenine monophosphate
Amp	ampicillin
ATP	adenine triphosphate
BCL	Boehringer Mannheim
BMP	Bone Morphogenetic Protein
bp	base pairs
BSA	bovine serum albumen
C	cytosine
C-terminal	Carboxy-terminal
CAT	Chloramphenicol acetyl-transferase
cDNA	complementary DNA
Ci	Curies
CNS	central nervous system
dATP	deoxyadenosine triphosphate
dCTP	deoxycytidine triphosphate
dGTP	deoxyguanosine triphosphate
DMSO	dimethylsulphoxide
DNA	deoxyribose nucleic acid
DNase I	Pancreatic deoxyribonuclease I
dNTP	deoxynucleotide triphosphate
dpc	days *post coitum*
dsDNA	double stranded DNA
DTT	dithiothreitol
dTTP	deoxythymidine triphosphate
EDTA	ethylenediaminetetra-acetic acid
EMSA	Electrophoretic mobility-shift assay
ES cell	Embryonal Stem cell
ExoIII	Exonuclease III
FGF	Fibroblast Growth Factor
G	guanine

g	grams
HEPES	(N-[2-Hydroxyethyl]piperazine-N'-[2-ethanesulfonic acid])
hr	hours
HTH	Helix-turn-helix
IGF	insulin-like growth factor
IPTG	isopropyl β-D-thiogalactopyranoside
kb	kilobases
kD	kilodaltons
L-broth	Luria-broth
M	molar
min	minutes
ml	millilitres
mM	millimolar
mRNA	messenger RNA
N-terminal	Amino-terminal
ng	nanogram
nM	nanomolar
OD_n	optical density$_{\text{wavelength in nanometers}}$
oligo	oligonucleotide
ORF	open reading frame
P-D axis	proximo-distal axis
PAGE	polyacrylamide gel electrophoresis
PBS	phosphate-buffered saline
PCR	polymerase chain reaction
PEG	polyethylene glycol
PMSF	phenylmethylsulfonyl fluoride
PNK	Polynucleotide Kinase
Pol II	RNA polymerase II
RA	retinoic acid
RNA	ribonucleic acid
rpm	revolutions per minute
s	seconds
SDS	sodium dodecyl sulphate
ssDNA	single-stranded DNA
T	thymine
TFD	Transcription Factor Database
TGF	Transforming Growth Factor
tk	thymidine kinase (HSV)
TLC	Thin Layer Chromotography

uORF	upstream Open Reading Frame
UTR	Untranslated Region
UV	Ultra Violet
X-gal	5-bromo-4-chloro-3-indolyl β-D-galactopyranoside
ZPA	Zone of Polarising Activity

List of Figures

Chapter 4

List of Tables

Chapter 1

Introduction

1 Introduction

Faithful heritability of bodily form is one of the fundamental properties of multi-cellular organisms. Genetic instructions, inherited from the previous generation, direct development of the body during embryogenesis, ensuring structural homology. The nature of the instructions and the processes by which they are followed have been elucidated by the combined powers of classical genetics and molecular biology. Genetic analysis of the fruit fly, instigated by T.H. Morgan in the early part of this century, has led to an understanding of the organisation and deployment of genes involved in establishing a body-plan. The latter half of this century has seen the discovery of molecules involved in the interactions that lead to body-plan specification.

Many of the genes and mechanisms involved in embryological development are conserved among widely divergent species. This enables access to the problems of embryogenesis in organisms less suitable to genetic analysis than *Drosophila*. Certain examples of extreme conservation highlight the central role of embryogenesis in metazoan evolution. While it is the genome that evolves, selection acts upon the functions of the body. The body is formed during embryogenesis according to information encoded by the genome and any bodily change that is selected for must be accompanied by a change in the appropriate embryogenetic processes generating it. Thus, embryogenesis imposes a restraint upon evolution of morphology as the potential for change is limited by the flexibility of the embryological processes concerned. The high degree of conservation that has been found between mechanisms of development, at the morphological and now the molecular level, reflects this. If we wish to understand the processes by which metazoan variation came about then a greater knowledge of the way in which genetic information is converted into three-dimensional form is essential.

A century ago, Bateson wrote on the variation of form found in the animal kingdom (Bateson, 1894). He noted that many dismorphic forms showed identifiable structures (or partial structures) inappropriately located at a position usually occupied by an alternative structure. This "phenomenon...not that there has merely been a change, but that something has been changed into the likeness of something else", he termed HOMEOSIS. Such 'homeotic' changes are readily definable in organisms with an overtly metameric body plan, such as the arthropods, where structural identities are often altered to those of adjacent metameres. While Bateson showed great foresight - "I believe that in the future its [homeosis] significance and the mode of its occurrence will become an object of high interest" - he was interested in such 'discontinuous

variation' for its role in speciation and hence evolution, not for the insight that it might provide into developmental strategies.

Mutants in the fruit fly, *Drosophila melanogaster*, were first identified in 1910. Many mutants were collected, but as part of studies aimed at understanding the rules of heritability, not development. As a consequence, mutants disrupting the body plan were not examined until Ed Lewis started work on the *bithorax* mutant, in the early 1950s. Over the next thirty years, by means of classical genetics, Lewis discovered many of the features of the immensely complicated 'bithorax complex'. Recent molecular analysis of this region has confirmed and clarified many of his findings.

1.1 The Homeobox

Lewis proposed that the homeotic selector genes of *Drosophila melanogaster* were members of a clustered, multigene family (Lewis, 1978). He had shown that mutants affecting segmental identity mapped to a tightly linked group, with similar phenotypic effects, that he termed the bithorax complex (BX-C). Such a gene complex could arise by duplication events producing an array of tandemly repeated genes related to a single precursor gene. He pointed out that this duplication and the consequent redundancy would enable the evolution of a wider repertoire of segmental identities and an accompanying increase in functional potential or "level of development". Further mapping of mutations affecting segmental identity led to the identification of a second cluster around the *Antennapedia* (*Antp*) gene (Kaufman *et al.*, 1980). This was termed the Antennapedia complex (ANT-C). Molecular cloning of the BX-C and ANT-C later confirmed that the complexes are gene clusters (Bender *et al.*, 1983; Scott *et al.*, 1983). One test of the proposal that BX-C and ANT-C arose by tandem duplication was to examine them for sequences repeated along their length.

1.1.1 Discovery of the homeobox

The question of whether the homeotic complexes represented multigene families formed by tandem duplication was addressed by several groups of workers at approximately the same time. Their findings confirmed many of Lewis's predictions and opened the way for the wealth of current research into the homeobox-containing genes.

The first clue to the presence of repeated sequences came from Garber and colleagues who showed that a cDNA from the *Antp* gene hybridised to the adjacent locus, the *fushi tarazu* (*ftz*) gene (Garber *et al.*, 1983). McGinnis and colleagues went

further, demonstrating that certain 3' probes from the *Antp* cDNA hybridised to several restriction fragments within the *Drosophila* genome and that a probe from the 3' end of the *ftz* transcription unit hybridised to many of the same fragments (McGinnis *et al.* 1984a). McGinnis and colleagues went on to show that a similar sequence was present in the 3' exon of the *Ultrabithorax* gene (*Ubx*). The association of this sequence with three genes involved in the segmental development of the fruit fly prompted them to search for other genes carrying the same sequence. Clones isolated from two genomic regions, using *Antp* and *ftz* probes on duplicate filters, were shown to contain the repeat sequence. One clone mapped to the bithorax complex, the other to the Antennapedia complex; both were shown to be expressed in a temporally controlled manner and in specific segmental regions within the developing embryo. The short, homologous region, seemingly restricted to genes involved in segmental development, was termed the *homeobox.*

Sequence of the homeobox from *Antp*, *Ubx* and *ftz* revealed a region of high nucleotide sequence homology (75-79% identity) extending over approximately 180 bp (Scott and Weiner, 1984; McGinnis *et al.*, 1984b). It is this 180 bp that has subsequently become defined as the homeobox. All three genes had open reading frames running the length of the homeobox, coding for highly homologous proteins (75-87% identity). The presence of several conservative substitutions between the peptide regions and the predominance of silent first and third position changes in the codons was taken to indicate that the homeobox represents a family of closely related, protein-coding DNA sequences.

1.1.2 Phylogenetic distribution of the homeobox

Early studies on the phylogenetic distribution of the homeobox led to speculation that it was limited to metameric organisms and unique to their characteristic developmental strategy (McGinnis *et al.*, 1984c). Later work, however, showed that the distribution of *Antp*-like homeoboxes was wide and covered many phyla including arthropods, annelids, chordates, echinoderms and molluscs (McGinnis, 1985; Holland and Hogan, 1986; Holland, 1992). Homeobox-containing genes were cloned from non-segmented organisms such as sea-urchins and were found to be expressed during embryogenesis in a stage specific manner (Dolecki *et al.*, 1986). *Antp*-class homeoboxes are known to be in nematodes (Bürglin *et al.*, 1989; Kenyon and Wang, 1991) and cnidarians (Schierwater *et al.*, 1991), the latter being considered as among the simplest of metazoa.

Soon after its identification the homeobox was used as a probe to isolate genes from vertebrates including the frog, *Xenopus laevis*, mouse and human (Carrasco *et al.*, 1984; Müller *et al.*, 1984; McGinnis *et al.*, 1984c; Levine *et al.*, 1984). In all cases the homeoboxes were highly conserved when compared to *Drosophila Antp*. Homologies ranged from 70-80% at the nucleotide level and higher for the putative translation products. Homeobox genes have now been cloned from all model organisms used in the study of vertebrate development including the chicken and the zebrafish along with those mentioned (Wedden *et al.*, 1989; Molven *et al.*, 1990).

The *Antp*-like homeobox is a sequence conserved across phyletic boundaries and throughout evolution. Its presence spans the major structural divisions of the animal kingdom (e.g. diploblastic and triploblastic organisms) and is not particular to any one developmental strategy. It seems likely that while it may be involved in the same mechanism used by several phyla, i.e. antero-posterior diversification, it generally plays the more common role of defining positional values in all metazoans.

1.1.3 Divergent homeobox-containing genes

So far I have discussed the *Antp*-like homeoboxes of the HOM-C and *Hox* complexes. As one of the first homeoboxes discovered, the *Antennapedia* homeobox has become the standard to which all homeoboxes are compared. All homeoboxes isolated in the early days of their study were of a type closely related to *Antp*. In addition to these genes, numerous other homeobox genes have been identified that are not members of one of the large complexes. This has come about both by searching for divergent homeobox sequences and by detection of homeobox motifs among previously cloned genes. *Drosophila* has yielded several examples of divergent homeoboxes; in many cases cognates of these genes are now known in vertebrates (Scott, 1990). Table 1.1 shows a list of several divergent homeobox-genes from *Drosophila* and their vertebrate cognates, where cloned. In addition to the groups defined by homology to *Drosophila* genes there are homeobox groups discovered from genes in other organisms. The POU homeoboxes are one such group. They were recognised as homologous between the mouse Pit-1 gene, the mouse Oct genes and the *C.elegans* gene Unc-86; Pit/Oct/Unc, hence POU. In the case of the POU and *Pax* genes the homeobox is only one of two highly conserved domains within the protein. The 'POU-specific' and 'Paired-box' domains are also DNA binding structures. *Pax* genes are defined as containing this motif but a number of them also contain a homeobox of the *paired* type. Whether a molecule has one or both of these domains is likely to affect the DNA target sites to which it binds (Czerny *et al.*, 1993).

While the HOM-C and *Hox* genes appear to specify anterior-posterior axial values, the divergent homeoboxes show a wide variety of expression patterns probably indicating widely varied functions. Several of these divergent homeobox genes, such as *Emx*, *Otx*, *Nkx*, *Msx* and *Dlx*, are expressed in the developing brain. The *Drosophila* archetypes of two of these gene families, *empty spiracles* and *orthodenticle*, have been shown to have a role in segmentation of the head; the expression pattern of their vertebrate homologues (*Emx* and *Otx*) suggests that this function may have been conserved and that these genes may play a role in subdivision of the developing vertebrate brain (Cohen and Jurgens, 1990; Simeone *et al.*, 1990; Holland *et al.*, 1992a). It has been suggested that given the ancient origin of the *Hox*-axis system and the relative evolutionary novelty of the brain, non-*Hox* genes have been recruited to provide positional values in more recent times (Holland *et al.*, 1992a).

The first non-*Antp*-like homeobox discovered was that of *engrailed*, a *Drosophila* gene with two vertebrate homologues (Poole *et al.*, 1985; Fjose *et al.*, 1985; Kuner *et al.*, 1985). Vertebrate *En*-1 and *En*-2 have a variety of expression sites including the embryonic midbrain-hindbrain boundary and the developing limb-bud (Joyner *et al.*, 1985; Davidson *et al.*, 1988). The limb is a site of expression for several divergent homeobox genes. As we shall see, the *Msx* genes are expressed there as are *Dlx*-2, *Evx*-1 and *Cdx*-1 (section 1.4.2; Bulfone *et al.*, 1993; Niswander and Martin, 1993a; Meyer and Gruss, 1993, respectively)

Table 1.1

Drosophila GENE	VERTEBRATE HOMOLOGUES
engrailed (en)	*En*-1, *En*-2
muscle-segment homeobox (msh)	*Msx*-1, *Msx*-2, *Msx*-3
even-skipped (eve)	*Evx*-1, *Evx*-2
orthodenticle (otd)	*Otx*-1, *Otx*-2
empty spiracles (ems)	*Emx*-1, *Emx*-2
distalless (dll)	*Dlx*-1, *Dlx*-2
caudal (cad)	*Cdx*-1, *Cdx*-2, *Cdx*-3, *Cdx*-4
NK/*tinman*	*Nkx*-1.1, *Nkx*-2.2, *Nkx*-2.5, *Nkx*-3.1

The chromosomal location of the *Evx* genes is of interest in considering the evolutionary origins of all homeobox genes. *Evx*-1 is located at the 5' end of the *Hox*A cluster and *Evx*-2 at the same end of the *Hox*D cluster (Bastian *et al.*, 1992). The

location of homologous, divergent homeobox genes in a conserved position within two of the four mammalian clusters suggests that these two clusters are more recent relatives and were derived from a single *Evx*-associated cluster by a duplication event. The more interesting evolutionary question raised by the location of the *Evx* genes is whether all homeobox genes were once part of a single complex. In *Drosophila*, *eve* does not map near the HOM-C but examination of homeotic complexes from an increasing number of organisms may shed light on the relationship between the 'divergent' homeoboxes and the '*Antp*-like' homeoboxes. The homeobox can be considered to have a single evolutionary origin but it seems that while many homeobox genes remained linked (the HOM/*Hox* homologues) several others became scattered around the genome at an early stage, it is these that we now think of as the divergent homeoboxes.

1.1.4 The homeobox encodes a DNA-binding protein

It was noted that the high proportion of basic amino-acids encoded by the homeobox are consistent with a possible role in DNA or chromatin binding (Laughon and Scott, 1984). Weak homology was detected, in the 3' end of the 60 amino-acid region (the homeodomain), to the a1 and α2 proteins of the yeast MAT locus (Shepherd *et al.*, 1984). These gene products had been implicated in control of gene expression at the transcriptional level (Strathern *et al.*, 1981). The homeotic genes of *Drosophila* had been proposed as developmental switches controlling the large battery of genes responsible for elaborating segmental phenotypes (Garcia-Bellido, 1975; Lewis, 1978). The basic charge of the homeodomain along with homology to the yeast proteins suggested that it was a DNA-binding structure apparently common to developmental control genes (McGinnis *et al.*, 1984b). The homeotic 'switches' therefore looked likely be DNA-binding transcriptional regulators (Shepherd *et al.*, 1984). This proposal of function was greatly strengthened by observations made in the key paper of Laughon and Scott (1984). They realised that the homeodomain of the *Drosophila* genes had, near the 3' end, a pattern of residues highly conserved among several bacterial DNA-binding proteins. The crystallographic structures of proteins such as λ Cro, λ repressor and *Escherichia coli* CAP protein revealed a DNA-binding structure consisting of two α-helices connected by a β-turn (the so-called helix-turn-helix - HTH). The C-terminal α-helix fits neatly into the major groove of the DNA molecule while the N-terminal helix lies across it making contacts with the sugar-phosphate backbone of the DNA. Specific residues, known from genetic studies and model building to be involved in the maintenance of this structure, are highly

conserved between these proteins. Laughon and Scott recognised this pattern of conserved positions essential to the helix-turn-helix within the 3' end of the homeobox. They noted that the positions involved in sequence-specific binding by the prokaryotic proteins were perfectly conserved between *Antp*, *Ubx* and *ftz* suggesting that they bind the same sequence.

The DNA-binding capacity of the homeodomain was first demonstrated with the protein product of the *Drosophila engrailed* gene. Desplan and colleagues (Desplan *et al.*, 1985) demonstrated the ability of proteins with an intact homeodomain to bind fragmented phage-λ DNA under conditions only permitting specific interactions. They also showed that a cluster of three high-affinity binding sites for the *en* homeodomain were present in 900 bp of the 5'-flanking DNA of the *en* gene. These sites are likely to be functional *in vivo* as they are conserved in another species of *Drosophila* and the chance grouping of apparently rare sites is highly improbable.

1.1.5 Structure and specificity of the Homeodomain

Specificity of action by the products of the homeobox genes is demonstrated to reside in the homeodomain. This is undoubtedly a rather broad oversimplification of the situation but it has been shown to be true for a number of *Drosophila* homeotic genes (Kuzoira and McGinnis, 1989; Ekker *et al.*, 1992c; Dessain *et al.*, 1992). The homeodomain is, by definition, a highly conserved structure yet it must provide differential specificity for the many gene-products in which it is found. As mentioned, the homeodomain is structurally related to the prokaryotic helix-turn-helix motif. Despite this homology the homeodomain does not bind DNA in a manner identical to the HTH. Genetic studies have shown that the DNA-binding specificity of the HTH resides in the N-terminus of the recognition helix (Pabo and Sauer, 1984; Wharton and Ptashne, 1985). In contrast, the homeodomain recognition helix has a C-terminal extension responsible for the specificity of the homeodomain-DNA interaction (Treisman *et al.*, 1989; Hanes and Brent, 1989). Crystallography performed on homeodomain-DNA complexes confirms this, showing a C-terminus in contact with the DNA molecule while the N-terminus is at some distance (Qian *et al.*, 1989; Kissinger *et al.*, 1990; Otting *et al.*, 1990). Systematic mutation of the *bicoid* and *paired* homeoboxes revealed that residue 9 of the recognition helix (helix 3 of the homeodomain) confers the ability on the homeodomain to distinguish between sites normally bound by divergent homeoboxes (Hanes and Brent, 1989; Treisman *et al.*, 1989). Determination of consensus binding sites for a large number of homeodomains indicates that all bind a common core sequence of 4 bp, flanked 3' by an additional 2 bp

that is highly variable. In this sequence, TAATNN, the bases present in the variable positions vary with the residue found at position 9 of the recognition helix (Treisman *et al.*, 1992a). By comparison to previously characterised combinations it is possible to predict the sequence bound by a homeodomain from the amino acid sequence of this region: *Msx-1* is a good example. The recognition helix has a Glutamine residue at position 9 (Hill *et al.*, 1989, Robert *et al.*, 1989), similar to the homeodomains of *Antp*, *Ubx*, *ftz*, *en* and others (e.g. Ekker *et al.*, 1991). From this one would predict that *Msx-1* bound a sequence TAATTG, TG being the suffix to the core sequence recognised by these homeodomains. Precisely this sequence has recently been identified as an optimal binding site for *Msx-1* (Catron *et al.*, 1993). The specificity dictated by the identity of residue 9 does not, however, determine all differences in site recognition. This position is invariant in *Dfd* and *Ubx* but their homeodomains are not functionally interchangeable *in vivo* (Dessain *et al.*, 1992).

A large body of evidence suggests that homeodomain proteins usually bind DNA as monomers (Qian *et al.*, 1989; Otting *et al.*, 1990; Kissinger *et al.*, 1990 Florence *et al.*, 1991) however there are reports of co-operative binding, by dimeric complexes of a human homeodomain protein, to tandem copies of a DNA-site with greater affinity than that found for the monomer (Galang *et al.*, 1992). There appear to be regions of the protein distinct from the homeodomain that are essential for this cooperativity. Cooperativity was detected *in vitro*, in the absence of any accessory factors however, such factors have been implicated in the precise recognition of half-sites by homodimers of the yeast α2 homeodomain protein (Smith and Johnson, 1992). It was shown that an extension to the homeodomain of α2 interacted with a second yeast regulatory protein, MCM1 (Vershon and Johnson, 1993). This extension region functions independently as shown by linking it to the *engrailed* homeodomain which then binds co-operatively with MCM1. The extension also specifies interaction with the Serum Response Factor (SRF) a human protein related to MCM1, raising the possibility that this is a common mode of recognition by homeodomain proteins. Such higher order complexes may be necessary in some cases to discriminate functionally significant *cis* elements from fortuitously similar sequences (Grueneberg *et al.*, 1992).

In summary, the homeodomain is a peptide region capable of binding DNA in a sequence-specific manner. The specificity is largely encoded by the recognition helix, with the identity of residue 9 directly influencing the sequence bound.

The nature of the homeodomain raises the problem of how such a highly conserved structure can produce the variation of specificity required to perform the diverse operations involved in systems such as *Drosophila* segmentation (Hoey and

Levine, 1988). The answer is far from clear but it is likely to involve a higher level of sensitivity in site recognition *in vivo* than is observed *in vitro*. There may also be several modes of DNA-binding by the homeodomain, the co-ordinated action of which maintains the accuracy of these complex systems. The majority of interactions appear to be made by monomeric homeodomain proteins but additional levels of specificity may operate by dimerisation of the proteins or formation of heteromeric complexes with the products of other genes. *In vivo* studies have shown there to be determinants of specificity outside the homeobox and these may represent domains involved in protein-protein interactions (Chan and Mann, 1993).

1.1.6 Homeobox genes are transcriptional regulators

Demonstration that the homeobox-containing genes encode proteins involved in transcriptional regulation came from work performed in cell free, cell-culture, and embryonic systems. The assignment of this function to the genes controlling *Drosophila* segmental identity supported the view that the homeotic genes are 'selector genes' that control 'realisator genes' responsible for terminal cytodifferentiation (Garcia-Bellido, 1975; Lewis, 1978).

Transcriptional activation and repression by homeodomain proteins was demonstrated biochemically using cell free systems (Biggin and Tjian, 1989; Ohkuma *et al.*, 1990). Repression was shown to occur by different mechanisms; steric hindrance of activators and active repression from a distance. *Drosophila* cells in culture were used in a number of studies that showed the activation and repression capabilities of several homeobox-gene products (Jaynes and O'Farrell, 1988; Han *et al.*, 1989; Winslow *et al.*, 1989; Krasnow *et al.*, 1989). In many cases the genes regulated by these proteins are other homeobox-containing developmental genes. Work done in the embryo revealed that the product of the *bicoid* (*bcd*) gene was an activator of the *hunchback* (*hb*) gene (Driever and Nüsslein-Volhard, 1989). This activation was shown to be dependent upon the concentration of the bicoid protein (Struhl *et al.*, 1989a). *bicoid* is a maternally expressed gene, the transcripts of which become distributed along the antero-posterior axis of the *Drosophila* egg in a concentration gradient. Interaction of this gradient with a number of bicoid binding sites located upstream of the *hb* transcription unit ensures that *hb* is activated only within a specific section of the gradient. Recently, repression by the products of the *Ubx* and *abd-A* genes has been shown to be cell-type specific, revealing mechanisms for a wider range of controlling functions by a single gene-product (Appel and Sakonju, 1993).

1.2 Evolutionary conserved clusters: structure and function

1.2.1 The homeobox-gene cluster is an ancient structure

Genetic analysis of the BX-C and ANT-C of *Drosophila* revealed that the order of genes along the chromosome was 'colinear' with their domain of function along the antero-posterior axis (Lewis, 1978; Kaufman, 1980). The BX-C and ANT-C (known collectively as HOM-C, the homeotic complex) comprise several genes with an *Antp*-like homeobox; *labial* (*lab*), *proboscipaedia* (*pb*), *Deformed* (*Dfd*), *Sex combs reduced* (*Scr*), *Ultrabithorax* (*Ubx*) and *Abdominal-A* (*Abd-A*), along with *Abdominal-B* (*Abd-B*) which has a slightly divergent homeobox but with more similarity to that of *Antp* than to any other class. They are the so-called 'homeotic selector' genes (Garcia-Bellido, 1975; Akam, 1988). The expression pattern of these genes was studied by *in situ* hybridisation. Genes were found to be expressed in regions that directly coincided with their domain of genetic function, emphasising the colinearity rule as a specific phenomenon at the molecular level.

Antp-like homeoboxes are organised into similar types of clusters in all organisms from which they have been cloned (Graham *et al.*, 1989). This suggests that there are specific mechanisms associated with this type of organisation that have been strongly conserved during evolution. In mouse there are four clusters of genes that encode *Antp*-like homeoboxes. These gene complexes, *Hox*A, *Hox*B, *Hox*C and *Hox*D, are found on chromosomes 6, 11, 15 and 2 respectively (McGinnis and Krumlauf, 1992 and references therein). The number of genes in each complex varies, with *Hox*A containing eleven genes and *Hox*B, C and D nine genes each. Sequence homology between genes of different clusters and conservation of the position of homologous genes along the clusters has led to the proposal that all four clusters are related. The present four clusters are thought to have arisen by a series of large duplication events in which the single, ancestral cluster was duplicated followed by similar duplication of each of the 'daughter' clusters (Hart *et al.*, 1987; Graham *et al.*, 1989; Duboule and Dollé, 1989; Schugart *et al.*, 1989). It is possible to assign genes in different clusters to sub-families of paralogous genes with which they share especially high homology (figure 1.1). There are thirteen paralogue families in all.

As the structure of these clusters was determined, several groups involved in the work performed *in situ* hybridisations to facilitate a comparative study of gene expression patterns along the clusters (Gaunt *et al.*, 1988; Graham *et al.*, 1989; Duboule and Dollé, 1989). There turned out to be a colinear relationship between the anterior boundary of expression of a gene (within the neural tube and the somitic

Figure 1.1

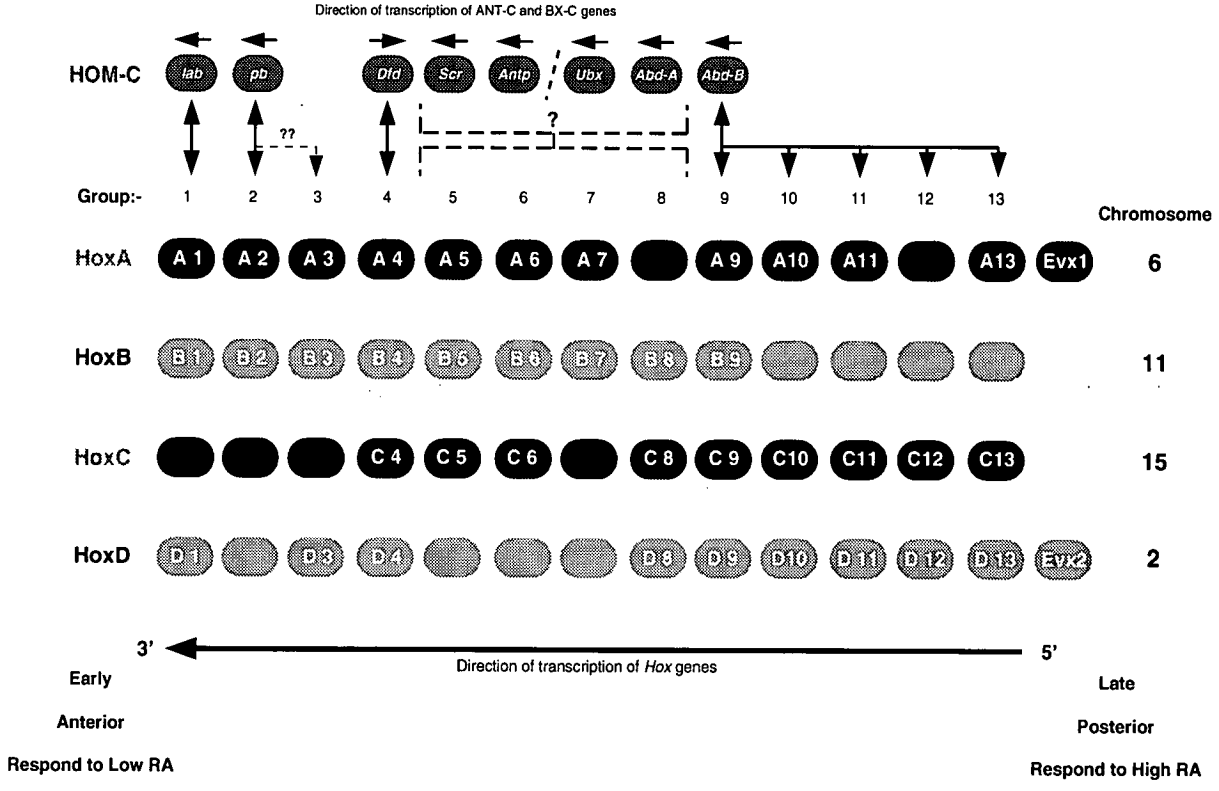


Figure 1.1 Diagramatic representation of map positions in the four mouse *Hox* clusters (bottom) and the HOM-C cluster of *Drosophila melanogaster*, derived by fusion of ANT-C and BX-C. The colinearity between homologues of mouse and *Drosophila* is clearly demonstrated. Below is an indication of the temporal and spatial colinearity known for the mouse genes and the differential response to Retinoic Acid seen by these genes in EC cells (see text).

mesoderm) and its position along the cluster, as with the genes of the HOM-C in *Drosophila*. In all cases a sharp anterior boundary of expression was noted with the signal extending posterior of this and in some cases fading slightly but never showing a clear posterior boundary. The colinear relationship was demonstrated directly for the entire *Hox*B cluster (Graham *et al.*, 1989) and for genes found in adjacent paralogous groups in different clusters (Gaunt *et al.*, 1988).

The observation that the murine genes are organised into evolutionarily conserved clusters, colinearly expressed along the A-P axis, led to the discovery that the paralogous sub-groups are themselves related to members of the *Drosophila* HOM-C that are arranged along the chromosome in a similar order, if the ANT-C and BX-C are viewed as a contiguous structure (Graham *et al.*, 1989; Duboule and Dollé, 1989; Figure 1.1). *Drosophila* HOM-C and vertebrate *Hox* clusters are therefore presumed to be homologous structures. Figure 1.1 shows how the four murine complexes and the *Drosophila* HOM-C can be aligned such that the genes lying at a similar position along the parallel complexes represent those with the highest homology. The problem of the discontinuity of the HOM-C (into ANT-C and BX-C), as compared to the vertebrate clusters, was illuminated by the discovery that the HOM-C is continuous in the beetle *Tribolium* (Stuart *et al.*, 1991; Beeman *et al.*, 1993). The split found in *Drosophila* may be specific only to a particular order of the insects, the dipterans.

It is possible to assign *lab*, *pb*, *Dfd* and *Abd-B* to groups of mouse genes with some confidence, based largely on homeobox sequence homology (McGinnis and Krumlauf, 1992 and references therein). For *Antp*, *Scr*, *Ubx* and *Abd-A* the extreme homology between their homeoboxes makes it difficult to say with certainty to which groups they belong. They are therefore assigned to four groups of murine genes with which they share equally high homology. The murine group aligning between *pb* and *Dfd* is sometimes linked with the *Drosophila* homeobox gene *Zerknüllt* (*Zen*) as it maps to this position in the ANT-C. *Zen* is not, however, a true homeotic gene involved in differentiation along the A-P axis. This group of murine genes shares equal homology with *pb* and may represent a new group produced since the split between vertebrate and arthropod lineages. It may be that early in vertebrate evolution there was a duplication of the *pb*-like gene in the single cluster with the 'new' gene evolving faster than its progenitor. Alternatively there has been a gene lost from HOM-C at some stage in evolution. There are examples of the vertebrate-specific duplication event, involving the *Abd*-B-like genes. The *Hox*B cluster contains a single *Abd-B*-like gene, *Hox*b-9. The other three clusters contain multiple examples of this type of gene. *Hox*A has four such genes while *Hox*C and *Hox*D both contain five. Clearly, during

vertebrate evolution the original *Abd-B*-like gene has been duplicated several times. It is not clear what the temporal relationship is between duplication of individual genes, along the chromosome, and duplication of the entire cluster to give the present structure but it would appear that the two processes have both occurred several times. Alternatively, the varying representation of paralogue groups in different clusters may result from differential gene loss from the duplicated copies of an original thirteen-gene vertebrate cluster (Schubert *et al.*, 1993). The recent discovery of a single homeobox-gene cluster in the chordate amphioxus (Garcia-Fernandez and Holland, 1994) suggests that multiple *Hox* gene clusters are a vertebrate-specific feature and that duplication may have been a significant step in the evolution of our own sub-phylum. A homeotic gene cluster has also been found in the nematode *Caenorhabditis elegans* and shown to be involved in patterning along the A-P axis (Wang *et al.*, 1993). It contains at least four genes that can be aligned to genes (or groups of genes) from the HOM-C on the basis of homeobox homology. The genes are found in the predicted order and their expression patterns obey the colinearity rule. These findings suggest a direct link between morphological elaboration and homeobox gene amplification.

It appears then that not only are the four murine clusters derived from a single progenitor but the *Drosophila* HOM-C is also related by some common ancestral cluster. This ancestral cluster must have been present in an organism alive some 600 million years ago and from which both vertebrate and arthropod lineages are descended. Based on the most strongly conserved sequences found in extant species it has been suggested that such a cluster contained genes structurally similar to *lab*, *pb*, *Dfd*, *Abd-B* and at least one of the *Scr/Antp/Ubx* type (Krumlauf, 1992). It seems that the arthropod/vertebrate cluster is a version of an ancient structure existing before the divergence of the nematode and arthropod/chordate lineages. It has been proposed that the original cluster from which nematode, arthropod and vertebrate clusters are derived contained three genes, a *lb/pb* type, an *Dfd/Scr/Antp/Ubx/Abd-A* type and an *Abd-B* type (Schubert *et al.*, 1993). Differential modification and elaboration of the structure of this cluster may have been significant to variations in body-plan as Lewis suggested (Lewis, 1978). The current role of the homeotic complex in the development of metameric organisms is likely to reflect the adoption of a pre-existing mechanism to a specific mode of development.

1.2.2 Mouse homeobox genes function in Drosophila

Considering the extreme conservation of sequence between the homeobox genes of insects and vertebrates, and the discovery that the gene complexes are

homologous structures, experiments were performed to determine the degree of any functional conservation.

It is known from work on *Drosophila* that ectopic expression of a homeotic gene will often lead to a phenotype in which additional parts of the body are transformed into those specified in the normal domains of expression. Indeed, some of the original homeotic mutants are gain of function mutations where ectopic expression leads to homeosis (e.g. Schneuwly *et al.*, 1987a & b). Ectopic expression of *Drosophila* genes is achieved using an inducible heat-shock promoter linked to the coding region of the gene in question. Similar ectopic expression of mouse genes in *Drosophila* showed that they gave a phenotype comparable to that produced by their endogenous cognate. Ectopic expression of the *Hox*b-6 gene produced a phenotype similar to that of ectopic *Antp*, the HOM-C gene to which it is most closely related (Malicki *et al.*, 1990). The phenotypic effects of ectopic expression of a homeotic selector gene are produced by inappropriate activation of the realisator genes that it controls. The assumption from the mouse-cognate experiment is that the vertebrate gene is able to activate many of the same genes as its fly counterpart. The *Hox*a-5 gene is capable of activating expression of the *Drosophila forkhead* (*fkh*) gene, a natural target of *Scr* of which *Hox*a-5 is a cognate (Zhao *et al.*, 1993). Experiments with human *Hox*d-4, a *Dfd* cognate, revealed that the two share similar regulatory specificities (McGinnis *et al.*, 1990). In this case the conservation of function has been shown to extend to the *cis*-acting elements associated with *Dfd* and *Hox*d-4. Not only is the mouse protein functionally equivalent to that of *Drosophila* but the autoregulatory element from the fly functions in a conserved manner when introduced into mice (Awgulewitsch and Jacobs, 1992) and a regulatory element from the human gene localises expression to the *Drosophila* head (Malicki *et al.*, 1992). This observation is not entirely unexpected as we might imagine that the great selective pressures that have preserved the homeobox genes during evolution have acted upon components of the regulatory circuits in which they function.

Conservation of the *Hox*/HOM genes during evolution is seen to include not only structure but also function. The specificity of action of the homeodomain proteins has been maintained over 600 million years of evolution. Identity between functionally equivalent molecules is limited to the 3' end of the homeodomain with some additional homology in the conserved hexapeptide. This restricted homology has led to suggestions that there are higher levels of conservation such as tertiary structure or that the position of a gene within a complex is more relevant to its functional specificity than sequence alone (Hayashi and Scott, 1990; Zhao *et al.*, 1993). Whatever

it may be, the molecular nature of this 'paralogue-specific' conservation must reflect a fundamental property of the *Hox*/HOM clusters

1.3 Homeobox genes regulate axial patterning

As mentioned, the murine *Hox* genes are expressed in positions that relate to their chromosomal location. This leads to a set of overlapping domains of which the anterior boundary is progressively more caudal the more 5' the gene is located. If the expression patterns are conceptually superimposed it becomes apparent that at any one level of the antero-posterior axis (somitic mesoderm or CNS) there is a specific combination of *Hox* genes expressed (Kessel and Gruss, 1991). This led to the proposal that position along the axis is defined by a combinatorial code of *Hox* genes, a model similar to that suggested for the homeotic genes of the BX-C (Lewis, 1978). Lewis' model states that a gene exerts its influence where it is the most posterior gene expressed. Therefore, it is the anterior boundary of the expression domain that is important. Clearly then, alteration of the expression domains should cause a change in pattern. A loss-of-function would result in transformation to a region with a more anterior identity (an anterior transformation) whereas a gain-of-function would cause a posterior transformation. These clear predictions of the '*Hox* code' model can be tested by studying the effects of changing the set of genes that are expressed in a given position. This can be achieved by causing either gain- or loss-of-function for specific *Hox* gene products.

1.3.1 Gain-of-function

Kessel and Gruss (Kessel and Gruss, 1991) proposed the vertebrate '*Hox* code' after studying the effects on axial patterning of the teratogen retinoic acid (RA). Retinoic acid is known to have teratogenic effects in a wide range of vertebrates (Tamarin *et al.*, 1984; Lammer *et al.*, 1985; Yasuda *et al.*, 1986; Ruiz i Altaba and Jessel, 1991). It was known that RA was capable of activating *Hox* genes in embryonal carcinoma cells (Colberg-Poley *et al.*, 1985; Hauser *et al.*, 1985) and more detailed studies had shown that genes were sequentially activated along the clusters in a 3' to 5' direction in response to an increasing concentration of RA (Breier *et al.*, 1986; Simeone *et al.*, 1990). Kessel and Gruss found that *in utero* application of RA caused homeotic transformations in the vertebral column. These transformations followed changes in the expression patterns of *Hox* genes along the A-P axis. In the cervical (anterior) region of the vertebral column RA produced posterior transformations that

followed an anterior shift in the anterior boundaries of *Hox* gene expression. This is consistent with the premature (more anterior) activation of more 5' *Hox* genes and the subsequent loss of the unique expression domains of the 3' genes; a gain-of-function in the domains of 3' genes caused by raising the RA concentration and resulting in posterior transformation of vertebral structures.

RA resulted in ectopic expression of *Hox* genes in anterior regions. The effects of ectopic expression of specific genes in defined regions were examined using transgenic mice expressing *Hox* genes under the control of heterologous promoters/enhancers. Ubiquitous expression of the *Hox*a-7 gene in mouse embryos resulted in complex cranio-facial abnormalities that were difficult to interpret (Balling *et al.*, 1989). In the trunk, however, ectopic expression of this gene caused transformations in vertebral identity interpreted as partial posterior transformations (Kessel *et al.*, 1990). Additional gain-of-function studies, using transgenic mice or injection of RNA into *Xenopus* embryos, have produced a variety of effects including alteration of axial identities but never a predictable set of transformations (Harvey & Melton, 1988; Wolgemuth *et al.*, 1989; Wright *et al.*, 1989; Cho *et al.*, 1991; McLain *et al.*, 1992). Over-expression of the human *Hox*c-6 gene in posterior regions of the mouse embryo, from which it is normally absent, causes *anterior* transformations (Jegalian & De Robertis, 1992). Such transformations are similar to those reported for the null allele of *Hox*c-8 (see below). This similar phenotype caused by both gain and loss-of-function suggests that the *levels* of homeobox-gene products are important in the specification of axial position.

1.3.2 Loss-of-function

Null alleles for the *Hox* genes have been created using homologous recombination technology in embryonic stem cells. Genes are inactivated by insertional mutagenesis in these cells which will then contribute to the germ-line of chimaeric mice. From these chimeras, homozygote mutants can be isolated in the F_2. *Hox*a-1 null mice show many cranio-facial abnormalities probably due to a disturbance in the patterning of the cranial neural-crest (Lufkin *et al.*, 1991). Similar effects are found in mice null for *Hox*a-3 (Chisaka & Cappechi, 1991). Both of these genes are early/anterior *Hox* genes and the phenotypes produced reflect their domain of expression. Disruption of *Hox*b-4 caused a partial homeotic anterior transformation in the cervical region, with the second vertebra, the axis, taking on certain characteristics of the first vertebra, the atlas (Ramírez-Solis *et al.*, 1993). Disruption of a more posterior gene, *Hox*c-8, gave a classic anterior transformation of the type predicted by

the combinatorial code (Le Mouellic *et al.*, 1992). The most anterior lumbar vertebra was transformed into a vertebra with the likeness of its thoracic neighbour, including the presence of a pair of ribs. Targeted mutation of *Hox*d-13 provides evidence that the *Hox* gene products are also involved in patterning the appendicular skeleton (Dollé *et al.*, 1993b). Apparently axial differentiation in alternate parts of the body is regulated by a common mechanism.

The presence of duplicated clusters in vertebrates complicates combinatorial models. Phenotypes may be rescued by functional redundancy among members of the same paralogue group. It is as yet unclear whether individual cells express multiple *Hox* genes at a given axial level, as the Lewis model requires. What is clear is that mutational analysis has the potential for elucidating the mechanisms whereby *Hox* genes provide axial information.

1.3.3 The Hindbrain

The hindbrain is divided into a series of units termed rhombomeres, recognised as bulges in the hindbrain neural epithelium. It was suggested that these may represent segmental units (Lumsden and Keynes, 1989) and subsequent grafting experiments confirmed this, showing each rhombomere to be a cellular compartment (Fraser *et al.*, 1990). There is a two-rhombomere periodicity to this compartmentalisation that ensures that cells will not mix between adjacent even and odd numbered segments. If two odd numbered rhombomeres are juxtaposed by grafting then their cells will mix freely. A similar periodicity of two rhombomeres is found in the differentiation of the branchial motor nerves from this region. The hindbrain represents the anterior limit of expression of the *Hox* genes. Anterior boundaries respect rhombomere boundaries with successive 5' genes expressed more posteriorly by a two rhombomere increment (Wilkinson *et al.*, 1989). One exception to this is *Hox*b-1 which is unique in having a defined anterior and posterior boundary limiting its expression to rhombomere 4 (Murphy *et al.*, 1989; Wilkinson *et al.*, 1989). It is proposed that the identity of individual rhombomeres is determined, at least in part, by the *Hox* genes that they express. In support of this idea, mutation of the 3' gene *Hox*a-1 causes alteration to hindbrain pattern including the reduction or loss of the anterior-most rhombomeres expressing the gene (Carpenter *et al.*, 1993; Dollé *et al.*, 1993a). Patterning of the branchial arches may also be associated with the *Hox* code in the hindbrain (Krumlauf, 1993; Wilkinson, 1993). The branchial arches grow ventro-laterally from the hindbrain region and are populated by cranial neural crest cells derived from this area of the

neural tube. These cells continue to express *Hox* genes corresponding to their position of origin in the hindbrain. The cranial neural crest forms the majority of mesodermally derived head structures in vertebrates with the branchial arch populations forming structures of the face and throat region (Noden, 1988). The disruption of normal expression patterns of the 3' *Hox* genes may cause the observed cranio-facial abnormalities by a perturbation of patterning in these neural crest populations (Lufkin *et al.*, 1991; Chisaka & Cappechi, 1991).

1.3.4 The Limb

The implication of *Hox* genes in axial patterning is extended to secondary axes such as those of the developing limb. Genes of the *Hox*D cluster are expressed in the limb bud in a series of partially overlapping domains with a common posterior-distal boundary and progressively restricted anterior-proximal boundaries for more 5' genes (Dollé *et al.*, 1989). This shows a marked similarity to the situation in the trunk. The limb is an accessible system in which regional interactions can be studied by grafting experiments. Regions of the limb bud have been identified as playing a role in defining the axial pattern of the structure, notably the Zone of Polarising Activity (ZPA) which determines antero-posterior polarity and the Apical Ectodermal Ridge (AER) which mediates outgrowth and proximo-distal patterning (Summerbell, 1974; Tickle *et al.*, 1975). *Hox*D genes expressed in the limb are from the 5' end of the complex and are also expressed in the posterior regions of the trunk (Dollé *et al.*, 1991). The gene at the 5' end of the *Hox*D complex is expressed at the posterior extremities of the trunk and in the most posteriorly restricted domain within the limb. This domain maps to the region of the ZPA. The ZPA has been suggested as a source of retinoic acid (Thaller and Eichele, 1987). The 'mirror image' digits produced by anterior grafting of the ZPA can be phenocopied by anterior application of retinoic acid (Tickle *et al.*, 1982). The similarity between trunk and limb *Hox* expression patterns, along with the recurring theme of RA, suggest that a common mechanism of *Hox* activation occurs during axial growth in the trunk and the limb.

The common model proposed is that *Hox* genes are activated progressively in a 3' to 5' direction along the cluster by a signalling centre (possibly using RA) moving in an anterior to posterior direction. A good candidate for this signalling centre in the trunk, given the that the *Hox* genes are first expressed during gastrulation, is the node (Hensens) at the cranial tip of the primitive streak (Hornbruch and Wolpert, 1986). As gastrulation proceeds the first cells passing through the streak form anterior structures.

The *Hox* genes are sequentially activated down the cluster, in cells entering the node, resulting in the observed spatial-colinearity. In the limb this signalling centre is the ZPA. It is significant that grafts of node into the limb provide polarising activity akin to that of the ZPA (Hornbruch and Wolpert, 1986). During limb outgrowth *Hox* genes are progressively activated by a signal from the ZPA. The interaction of growth and sequential activation produces the observed 'nested' domains of expression. It has been proposed that spatial-colinearity is the observed end-product of a mechanism of 'temporal-colinearity' due to the development of the vertebrate embryo in an anterior to posterior (anterior-proximal to posterior-distal in the limb) direction where the anterior of the embryo is always 'older' than the posterior (Dollé *et al.*, 1989; Duboule, 1992).

1.4 The Msh homeobox-gene family

The *Hox* genes clearly have a fundamental function in the specification of body plan. They have filled the role of determinants of anterior-posterior specificity throughout a large part of metazoan evolution, no doubt influencing the paths taken in the establishment of morphological complexity. However, as mentioned earlier (section 1.1.3), the *Hox* genes are but one class of the wider family of homeobox genes. In view, both of the importance granted to the *Hox* genes and the significant functions encoded by other classes of homeobox genes in *Drosophila* these divergent homeobox genes have also attracted much scrutiny.

One family of diverged homeobox genes is that related to the *Drosophila msh* gene. *Drosophila* carries a single gene of this type encoding a homeodomain with only 42% amino acid identity to that of *Antp* (Gehring, 1987). The *msh* homeobox encodes the nine positions invariant among all other homeoboxes but shows divergent features such as an Arginine→Threonine substitution at position 43 in comparison with the *Antp* homeobox. This is a rare change previously found only in the *labial* like homeoboxes. Differing approaches were adopted in cloning mouse cognates of the *Drosophila msh* gene. Direct library screening, with the *msh* homeobox and 3' end, enabled isolation of a mouse *msh*-like gene (Robert *et al.*, 1989). The same gene was isolated during a search for diverged homeobox sequences in which homeoboxes were sought that showed weak hybridisation to the *Hox*a-1 homeobox (*labial*-like) and none to the *ftz* homeobox (Hill *et al.*, 1989). This gene encodes a homeodomain that differs from the *msh* homeodomain at only 5 positions. Homology extends eleven amino acids 3' and at least seven amino acids 5' of the homeodomain. The gene was originally termed *Hox*-7.1 but has been renamed *Msx*-1 (Scott, 1992). Besides *Msx*-1 there are

two other *msh*-like genes in the mouse genome (Holland, 1991; Robert Hill, personal communication). *Msx*-2 encodes a protein with an overall 60% identity with *Msx*-1 and a homeodomain that differs at only two positions (Monaghan *et al.*, 1991). *Msx*-2 is the new name for the gene previously termed *Hox*-8.1. Characterisation of the third gene has not yet been reported with the exception of the homeobox sequence that has been determined by Holland (1991). The *Msx*-3 homeodomain differs at one position from that of *Msx*-1 and at three positions from that of *Msx*-2. The hexapeptide region upstream of the homeodomain, conserved among many *Antp*-like homeobox-genes, is poorly conserved in the *Msx* genes. *Msx*-1 has 3 of the 6 residues and *Msx*-2 has only 2. *Msx*-1 maps to mouse chromosome 5 and is not linked to any known homeobox-gene cluster (Hill *et al.*, 1989). *Msx*-2 is not linked to *Msx*-1 (Monaghan *et al.*, 1991) and maps to chromosome 13 (Bell *et al.*, 1993). A single intron is present in both *Msx*-1 and *Msx*-2 at the same position, 43 bp upstream of the homeobox; a genomic structure similar to many *Drosophila* HOM-C and murine *Hox* genes (Monaghan *et al.*, 1991). Cognates of the *Msx* genes have been cloned, or at least identified, from a variety of organisms. In vertebrates including chicken, quail, zebrafish, human and mouse there are 1-3 *Msx* genes cloned (Yokouchi *et al.*, 1991; Coelho *et al.*, 1992; Takahashi and Le Douarin, 1991; Akimenko *et al.*, 1991; Hewitt *et al.*, 1991; Hill *et al.*, 1989; Robert *et al.*, 1989). It is likely that full characterisation will reveal each of these organisms to have three genes as in mouse. A possible exception among vertebrates is zebrafish which has 5 *Msx* genes (Akimenko *et al.*, 1991; Monte Westerfield, personal communication). This unusual situation could result from the zebrafish genome being polyploidal (Ekker *et al.*, 1992a). Holland has engaged in an evolutionary study of the *msh*-like genes. In contrast to the 3 genes he found in mouse and zebrafish he has isolated only a single *msh*-like gene from both a Cephalochordate (*amphioxus*) and an Ascidian (*Ciona intestinalis*). Urochordates (including *Ascidiacae*) and Cephalochordates are considered the closest related organisms to the vertebrates. Vertebrates have many specific features not shared by these other members of the chordate phylum including the migratory neural crest, cranial and spinal ganglia, branchial arches and ectodermal placodes. The duplication of the *msh*-like genes in the vertebrate sub-branch may be significant to the development of some of these features. This seems especially likely when the expression patterns of these genes are considered.

Northern-blot analysis of mouse embryonic mRNA revealed that an *Msx*-1 transcript of 2.0-2.2 kb is expressed during embryogenesis, with a peak at ~9.0 days *post coitum* (Hill *et al.*, 1989). Detailed studies of expression patterns were undertaken

by RNA *in situ* hybridisation (Robert *et al.*, 1989; Hill *et al.*, 1989; Monaghan *et al.*, 1991; Davidson & Hill, 1991; MacKenzie *et al.*, 1991a & b; MacKenzie *et al.*, 1992). Unlike the *Hox* genes, *Msx* genes show no anterior-posterior specificity in their expression patterns. They are expressed along the full length of the embryo in the neural plate (subsequently in the neural crest) and the lateral plate mesoderm (Davidson & Hill, 1991). Neural plate expression is restricted to the tips of the neural fold which express high levels before and after neural tube fusion, when they mark the dorsal midline. Cells from this region constitute the migratory neural crest. They express *Msx*-1 at higher levels than *Msx*-2. Lateral plate expression is restricted to increasingly lateral positions with the final, stable domain of *Msx*-2 located more laterally and within that of *Msx*-1 (Davidson & Hill, 1991). It is possible that these genes play a role in patterning the medio-lateral axis of the early embryo.

Cells of the neural crest undergo extensive migration (Le Douarin, 1982), maintaining expression of both *Msx*-1 and *Msx*-2 as they do so. Organogenesis involving *Msx* expressing mesodermal populations, along with migration of the neural crest to several morphogenetically active regions results in elaborate expression patterns for the *Msx* genes. Expression then appears to enter a second phase in which maintenance and regulation are under the control of local influences. Only one site of *de novo* expression is known at present, that in the developing eye (Monaghan *et al.*, 1991). All other expression is derived from this early pattern. This diverse set of expression domains includes the developing head and facial structures, pituitary, ear, heart, genital structures and limbs (MacKenzie *et al.*, 1991a; Ekker *et al.*, 1992b; Hill *et al.*, 1989; Lyons *et al.*, 1992; Davidson *et al.*, 1991). The more closely studied of these systems are discussed in detail below.

1.4.1 Cranio-facial expression

In situ analysis of *Msx* gene expression in the developing mouse shows that early in the development of the head and face (8-9 dpc) *Msx*-1 is expressed in all neural crest cells that have migrated to this region (MacKenzie *et al.*, 1991a). *Msx*-1 is highly expressed in both the epithelium and mesenchyme of the medial and lateral nasal processes at 9.5 days *post coitum* (dpc) (Hill *et al.*, 1989; MacKenzie *et al.*, 1991a). Expression is high in the mesenchyme of the first branchial arch (mandibular arch), decreasing in a distal→proximal gradient. These sites of expression persist until fusion of the nasal, maxillary and mandibular processes during day 12 of development. By 11 dpc expression in the first branchial arches is restricted to the distal tips with localised expression persisting proximally in the mesenchyme surrounding the dental epithelium

(MacKenzie *et al.*, 1991a). Mesenchyme surrounding the developing brain, which forms bones of the skull, expresses highly during days 10 and 11 (Hill *et al.*, 1989; Robert *et al.*, 1989; MacKenzie *et al.*, 1991a). This expression has been implicated in the phenotypic effects of a human MSX2 mutation (Jabs *et al.*, 1993). The mutation causes a dominant defect in the sutures joining the skull bones resulting in craniosynostosis. During days 14-16 low levels of *Msx*-1 expression are detected in the bone and muscle anlage of the jaw, derived from neural crest and lateral plate mesoderm respectively (MacKenzie *et al.*, 1991a). Expression is high in the neuro-epithelium of the telencephalon that is destined to form the lateral choroid plexus. *Msx*-1 may be involved both in patterning and organogenesis of the choroid plexus. As the choroid plexus forms, *Msx*-1 expressing epithelium delaminates and is convoluted by an invasion of overlying dorsal mesenchyme (MacKenzie *et al.*, 1991b); this mesenchyme also highly expresses *Msx*-1. Rathke's pouch is another site of high *Msx*-1 expression during its formation on days 9.5-11.5 (MacKenzie *et al.*, 1991b). This small outpocketing of epithelium, in the roof of the oronasal cavity, forms the anterior pituitary.

Both *Msx*-1 and *Msx*-2 are expressed in the developing tooth. Until the cap stage *Msx*-2 is expressed in the epithelium of the enamel organ, with expression also in the underlying mesenchyme of the dental papilla by the late cap and bell stages (MacKenzie *et al.*, 1991a & 1992; Jowett *et al.*, 1993). By contrast *Msx*-1 is expressed solely in the underlying dental mesenchyme. *Msx*-2 expression persists until differentiation of epithelial cells into ameloblasts. The regulation of these genes during tooth development correlates with the developmental stage of the tooth rather than that of the embryo.

Expression of the *Msx* genes may be involved in patterning the developing eye (Monaghan *et al.*, 1991). *Msx*-2 expression marks the inner layer of the optic cup before invagination. Later, only the region of the inner layer destined to form the neural retina expresses. *Msx*-1 expression is specific to just below the tip of the optic cup from which the ciliary body develops. Expression marks this region up to two days before the ciliary body is morphologically distinguishable from the neural retina.

The inner ear is also a site of *Msx* gene expression (Ekker *et al.*, 1992b). In the zebrafish three *Msx* genes are expressed during the development of the semicircular canals and later in specific cell types within the structure.

Msx-1 has recently been shown to be essential for correct development of craniofacial structures (Satokata and Maas, 1994). Mice homozygous for a mutation in *Msx*-1 show defects in the secondary palate, mandible, maxilla and development of the teeth and middle ear.

1.4.2 Limb-bud expression

Both *Msx*-1 and *Msx*-2 are expressed in the developing limb (Davidson *et al.*, 1991). When the limb bud first protrudes from the flank *Msx*-1 is expressed throughout with *Msx*-2 confined to the ventro-distal ectoderm and a small region of underlying mesoderm. The expression pattern changes during outgrowth of the limb bud. By 11.5 dpc *Msx*-1 becomes restricted to the posterior-distal mesenchyme where it is expressed at high levels. *Msx*-2 is also present in the distal mesenchyme, at somewhat lower levels, and within the overlying ectoderm, the Apical Ectodermal Ridge (AER). The AER does not detectably express *Msx*-1. Following formation of the foot-plate and cartilaginous condensations of the digits, *Msx*-1 is expressed in the interdigital mesenchyme (Hill *et al.*, 1989). Cell death in this region is responsible for sculpting the limb. A chicken homologue of the *Msx*-1 gene is expressed in a similar domain under the AER (Coelho *et al.*, 1993a). Independent domains of expression are found in the proximal anterior and posterior regions of the limb-bud. These overlap but do not map exclusively to the anterior and posterior necrotic zones where cell death is again essential to the shaping of the limb-plate. The coincidence of *Msx*-1 expression with sites undergoing programmed cell death suggests that *Msx*-1 is involved in this process.

In the chicken embryo, grafting experiments and exploitation of certain chicken mutations affecting the limb have elucidated the mechanisms whereby the *Msx* genes are controlled during limb outgrowth. *Msx*-1 and *Msx*-2 respond to positional cues within the limb (Davidson *et al.*, 1991). This was shown by performing inter-specific grafts between mouse and chicken. A piece of proximal mouse limb-bud was taken from a region outside the distal expression domain of the *Msx* genes. This was grafted into a wing-bud of comparable stage in a more distal position; corresponding to the region in which the genes are expressed. Both genes were turned on in the graft in a manner dependent upon a distal position. Activation within the grafts closely matched expression in the surrounding tissue with a gradient of expression higher at the distal tip. These observations suggest that a diffusible factor emanating from the AER is up-regulating the *Msx* genes. This hypothesis is strengthened by examination of *Msx* gene expression in the chicken mutant *limbless* (Robert *et al.*, 1991). In *limbless* mutants there is a defect in the response of ectoderm to AER-inducing signals from the mesoderm (Carrington and Fallon, 1988). Consequently, initial budding is successful but outgrowth fails in the absence of a functional AER. Expression of the *Msx* genes is seen early in the budding flank mesoderm, prior to outgrowth, but is not maintained and soon disappears. The *limbless* phenotype can be rescued by grafting wild-type

ectoderm over the bud whereupon outgrowth continues. Such a rescue causes activation of the *Msx* genes as normal. Similarly, an additional outgrowth can be induced if an ectopic AER is grafted onto the dorsal side of a normal limb-bud (Robert *et al.*, 1991). Mesenchyme underlying the AER in this extra outgrowth activates *Msx* genes as seen in the endogenous bud. Two other chicken mutations causing limb abnormalities demonstrate the intimate relationship between *Msx* expression and the AER (Coelho *et al.*, 1993b). *talpid*^{2} and *diplopodia*-5 both cause polydactyly as a result of an expanded anterior-posterior axis in the limb bud. The broad limb bud found in both of these mutants is capped by an AER that extends the full width of the bud. In both cases the mesenchyme underlying this extended ridge expresses *Msx*-1 in a proximo-distal gradient similar to that of normal limbs. In addition *Msx*-2 is expressed throughout the ridge. Interestingly these polydactylous mutants lack necrotic zones in the anterior and posterior proximal regions of the bud. This coincides with an absence of *Msx* gene expression in these regions.

These data point to specific modes of regulation for the *Msx* genes in the vertebrate limb-bud. Study of the *limbless* mutation shows that there are two phases of *Msx* activation in the limb; an initial induction followed by an AER-dependent maintenance. The loss of part, but not all, of the expression pattern in the limbs of the polydactylous chickens implies that there are several regulatory influences acting upon these genes, possibly through different enhancers specific to each region (Coelho *et al.*, 1993b). The reliance upon a functional AER for *Msx* expression, and the close spatial relationship whereby underlying tissue expresses most highly, clearly suggests that the AER provides a factor essential for the maintenance of *Msx* gene expression. This factor is most likely a diffusible molecule such as a growth factor. Several candidates exist in this region of the limb; these include BMP-2A, BMP-4A, Wnt-1 and Wnt-5A, all members of the TGF-β family (Robert *et al.*, 1991; Gavin *et al.*, 1990; Parr *et al.*, 1993) and the FGF-like molecules FGF-4 and FGF-2 (Niswander and Martin, 1992; Munaim *et al.*, 1988). Recent work has suggested that FGF-4 may indeed be a diffusible factor produced by the AER and acting on the underlying mesenchyme (Niswander *et al.*, 1993b). Appropriate response of tissue grafted from mouse limb bud to that of chicken suggests that both the signalling pathway and the positional nature of the signal responsible for regulating these genes is highly conserved between birds and mammals. The expression pattern of *Msx*-1 coincides well with a region defined as the 'progress zone' (Robert *et al.*, 1991; Summerbell *et al.*, 1973). This is a zone of undifferentiated mesenchyme at the distal tip of the bud thought to be the site of positional specification along the proximo-distal axis. *Msx*-1 may have a role in maintaining the embryonic state of the progress-zone cells. This idea is supported by

the observation that over expression of *Msx*-1 in myogenic cells blocks terminal differentiation of these cells into myotubes (Song *et al.*, 1992). Overexpression of *Msx*-2 in the same system fails to prevent differentiation.

Recent generation of mice lacking *Msx*-1 function (Satokata and Maas, 1994) has cast doubt upon the role played by this gene in development of the limb, as the limbs of these mice are apparently normal. This conflict with the expression and functional data described may be explained by functional redundancy with *Msx*-2. Generation of mice homozygous for null alleles of *Msx*-2 and of both genes will address this possibility.

1.4.3 Msx genes and epithelial-mesenchymal interactions

A common feature of several sites expressing the *Msx* genes during embryogenesis is their involvement in epithelial-mesenchymal interactions. Such secondary interactions are characteristic of morphogenesis in many systems including the heart, ear, limb, tooth and facial processes; all sites of *Msx* gene expression. In many cases (tooth, limb, face) *Msx*-1 and *Msx*-2 are expressed in the epithelium and mesenchyme in a complementary pattern suggesting that they may define differences between the two or be involved in their alternative responses to signals. There does not seem to be a consistent rule for which tissue type expresses which gene.

Interaction between the flank mesoderm and the overlying ectoderm induces the formation of a region of thickened ectoderm, the AER, which in turn interacts with the underlying mesoderm inducing outgrowth of the limb bud (Saunders, 1948; Summerbell *et al.*, 1974). Outgrowth of the facial primordia has also been shown to require epithelial-mesenchymal interactions (Wedden, 1987; Richman & Tickle, 1989). Recombination experiments have shown that the limb and face share similar signals responsible for mediating these interactions (Richman & Tickle, 1989). The distribution patterns of the *Msx* genes make them ideal candidates as molecules involved in these signalling pathways. Recent work has shown that *Msx* genes respond to positional signals from both the face and the limb (Brown *et al.*, 1993). Inter-specific grafts show that *Msx* genes are capable of responding to local signals in the maxillary process regardless from which facial process they originate. Similarly, mesenchyme grafted from the face can respond to local signals in the limb by expressing *Msx* genes appropriately. However, reciprocal grafts show that limb mesenchyme is incapable of responding to signals in the face.

It has been proposed that a similar system regulates outgrowth in the facial primordia and the limb bud (Richman and Tickle, 1992). Outgrowth of the tail may

also be regulated by the same system; the ventral tail ridge, a thickened ectodermal structure essential to tail development, also expresses *Msx*-1 (Lyons *et al.*, 1992). The grafting experiments suggest that *Msx* genes are involved in such a system and that the epithelial to mesenchymal signals in face and limb are very similar, despite the lack of an ectodermal ridge on the facial processes. The inability of limb mesenchyme to respond to facial signals may represent a quantitative difference in the extra-cellular signals required or an absence of additional limb-specific signals (Brown *et al.*, 1993). *In vitro* recombination assays involving dental epithelium and mesenchyme have elucidated the relationship between epithelial-mesenchymal interactions and *Msx* gene activation (Vainio *et al.*, 1993). These assays enable examination of the effects of candidate signalling molecules and their role in pathways leading to gene activation.

In summary, the *Msx* genes are a small family of diverged homeobox genes related to the *Drosophila msh* gene. They show expression patterns distinct from the *Hox* genes. Duplication of a single ancestral gene appears to be vertebrate specific and may have had a role in evolution of several vertebrate-specific features. Expression pattern studies support this hypothesis with the genes activated in regions such as the neural crest, ectodermal placodes and several cranial sense organs. Many sites of expression coincide with morphogenetically significant epithelial-mesenchymal interactions and the complementary patterns of different *Msx* genes in tissues implicate them in signalling or signal response. Expression in the limb correlates with a region of undifferentiated mesenchyme and along with cell transfection studies this suggests that the *Msx* genes are involved in maintaining cells in this state. *Msx* genes are also expressed in several locations undergoing programmed cell death, indicating a possible role in this process.

1.5 Transcriptional regulation

The original paradigm for our understanding of the mechanisms of gene regulation, at the transcriptional level, was provided by the prokaryotic operon model of Jacob and Monod (1961). This basic model, in which gene activity is regulated through adjacent *cis*-acting sequences by *trans*-acting regulatory proteins, also applies to eukaryotic systems (reviewed in Gluzman, 1985).

1.5.1 cis *regulation*

Genes encoding messenger RNA's are transcribed by RNA polymerase II, one of three RNA polymerases in the cells of higher eukaryotes. Analysis of such genes has provided a view of the organisation of *cis* sequences in a typical pol II eukaryotic gene (Maniatis *et al.*, 1987). This organisation is summarised in figure 1.2A&B. Two classes of DNA elements function in the regulation of these genes; promoters and enhancers. Promoters are situated immediately upstream of the transcriptional start site from which they mediate accurate and efficient initiation of transcription. Analysis of a number of promoters reveals the presence of a common set of features typical to this region (Dynan and Tjian, 1985; McKnight and Tjian, 1986; Myers *et al.*, 1986; Figure 1.2A). An AT-rich region termed the TATA box (or Golberg-Hogness box) is located approximately 30 bp upstream of the initiation site. This is the site at which assembly of the initiation complex (the polymerase and its associated factors) occurs; it may also be essential for strand separation (Buratowski *et al.*, 1989; Roeder, 1991; Klug, 1993). Upstream of the TATA box are binding sites for a variety of proteins. Several ubiquitously expressed proteins are known to bind to these 'upstream promoter elements' where they act to promote the level of transcription. These elements include a sequence containing the CCAAT motif which binds several factors including the ubiquitous CTF (Dorn *et al.*, 1987; Santoro *et al.*, 1988), and the sequence GGGCGG, the so called GC-box, that binds the ubiquitous factor Sp1 (Dynan and Tjian, 1985). These proteins are involved in the stimulation of basal levels of transcription. Upstream promoter elements can bind both ubiquitous and cell-type specific factors, for example the CCAAT sequence interacts with the liver-specific factor C/EBP as well as CTF (Friedman *et al.*, 1989). The β-globin gene promoter has elements bound by both ubiquitous (CCAAT) and specific (CAC) factors (Mantovani *et al.*, 1988). Specific promoter factors appear to co-operate with the ubiquitous proteins to stimulate transcription (deBoer *et al.*, 1988). Recent X-ray crystallography studies of the

Figure 1.2

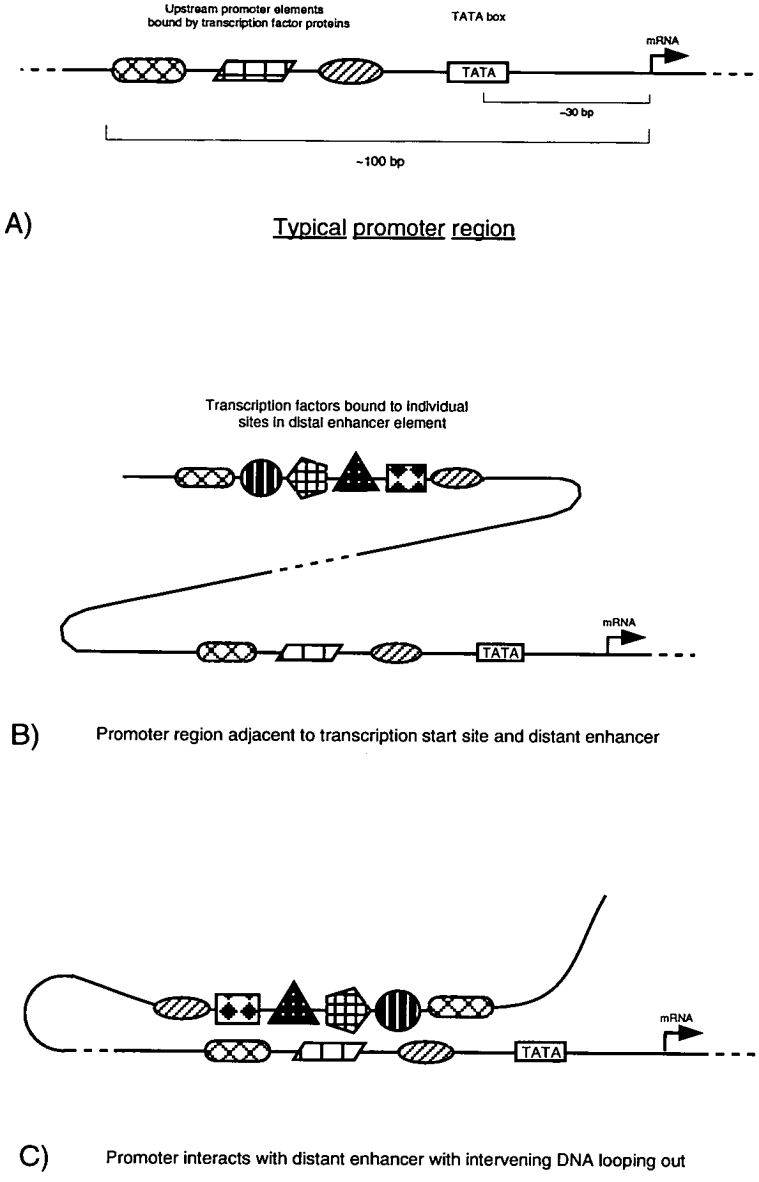


Figure 1.2 Diagrammatic summary of regulatory elements elements in a canonical eukaryotic gene. A) shows the organisation of the typical proximal promoter region. B) shows the binding of transcription factors to a distal enhancer element in which there are binding sites for specific factors and some of those usually associated with the proximal promoter. C) figure depicting the looping out of intervening DNA in one proposed mechanism of enhancer-promoter interaction.

interaction between the TATA box and the TATA-binding protein (TBP) suggest that the initiation complex may play a role in distorting the DNA such that the complex and the upstream element-binding proteins closely interact (Kim *et al.*, 1993a; Kim *et al.*, 1993b).

The use of multiple promoters by a single gene adds a further level of variability to transcriptional regulation (Schibler and Sierra, 1987). A number of *Drosophila* developmental genes are known to be transcribed from multiple promoters, including *Antp* (Schneuwly *et al.*, 1986; Bermingham *et al.*, 1990) and *caudal* (Mlodzik and Gehring, 1987). The alcohol dehydrogenase (*Adh*) gene of *Drosophila* also uses tandem promoters in a differential manner during its developmentally regulated expression in the larva and adult (Lockett and Ashburner, 1989). At the simplest level of control, one promoter may be significantly more efficient than the other, dictating the levels of transcript produced, independent of enhancer function. The use of different promoters provides several alternatives in the production of both transcript and protein product. The arrangement of promoters in relation to exons and start codons can provide diverse 5' leader sequences with a similar open reading frame (with or without the use of additional 5' exons) or variable translation products by using alternative initiation codons (Schibler and Sierra, 1987).

Tissue-specific, region-specific and temporal regulation of transcription is largely mediated via the second class of *cis* elements. Enhancers are regions of DNA capable of acting at a distance from the promoter and in an orientation independent manner (Serfling *et al.*, 1985; Ptashne, 1986; Hatzopoulos *et al.*, 1988). Their position relative to the coding region varies greatly, being either upstream, downstream or within introns (figure 1.2B shows an enhancer element upstream of the gene). Enhancers are further defined as capable of conferring their specificity upon a heterologous promoter. They were first identified from studies on DNA tumour viruses such as SV40 and Polyoma (de Villiers and Schaffner, 1981; Banerji *et al.*, 1981; Herbomel *et al.*, 1983). Sequences close to the viral origin of replication were found to be capable of enhancing transcription from a heterologous promoter by several hundred-fold. Analysis of the immunoglobulin heavy chain gene revealed a sequence with similar properties within an intron (Gillies *et al.*, 1983; Mercola *et al.*, 1983). This enhancer specifically directs transcription within myeloma cell lines. Activation of genes at specific times during development is regulated through enhancers, as shown for the Mid-Blastula-Transition-specific enhancers of *Xenopus laevis* genes (Krieg and Melton, 1987). Heterologous genes linked to this are activated at the same stage of embryogenesis. Enhancers within exon 3 and downstream of the β-globin gene specifically boost expression within erythroid lineages (Behringer *et al.*, 1987; Kollias

et al., 1987). A master enhancer is present at the β-globin locus that exerts an influence across the whole gene family on chromosome 11 (Grosveld *et al.*, 1987). This 'Dominant Control Region' is also involved in providing erythroid-lineage-specific function (van Assenfeldt *et al.*, 1989). Enhancers have been identified in many organisms and with a wide variety of specificity (Walker *et al.*, 1983; Garabedian *et al.*, 1986).

Detailed analysis of enhancers reveals that they have many similarities with promoter regions. Enhancers vary in their level of complexity from as few as one protein binding site to a multi-level organisation in which many individual sites (enhansons) are grouped into 'modules' (Ondek *et al.*, 1988; Dynan, 1989). Enhancers are found comprising one or several such modules. Enhancer activity can be produced by creating an array of adjacent, identical sites, a phenomenon demonstrated by early work on the SV40 major enhancer: mutants with a truncated enhancer, that were unable to produce high levels of transcription, often reverted by means of tandem duplications of the remaining, shortened enhancer region (Herr and Clarke, 1986). Despite the differences between promoters and enhancers, the same proteins are often found to bind to the two classes of *cis* element (Sen and Baltimore, 1986).

The mode of action for an enhancer sequence located some distance from its point of influence, the promoter, has been the subject of much theorising. Enhancers are capable of acting not only at considerable distance on the same molecule but function in *trans* has been demonstrated where the two DNA molecules are either concatenated (Dunaway and Dröge, 1989) or connected by a protein bridge (Müller *et al.*, 1989). These observations discount one proposed model, the scanning model, in which the enhancer serves to recruit proteins that then migrate along the DNA to the promoter. They support, however, the view that action at a distance occurs by means of direct physical interaction between the two sites. The model now most widely accepted for enhancer function is one where an enhancer recruits regulatory protein complexes which then interact with the promoter by means of a looping out of the intervening DNA (Figure 1.2C). The DNA-looping model is demonstrated to be correct by electron-microscopic analysis of DNA-protein complexes (Su *et al.*, 1991; Mastrangelo *et al.*, 1991). An interaction between two protein complexes, bound at proximal and distal sites, is seen to unite the two DNA sites resulting in an intervening DNA loop. This is proposed to be a mechanism for the maximal recruitment of regulatory molecules to the promoter. The ubiquitous activator Sp1 stimulates transcription to a far greater degree when present at the promoter in multiple copies (Anderson and Freytag, 1991). This activity is synergistic and may explain the strong effects exerted by enhancers as they recruit large numbers of regulatory molecules to

the promoter region. Enhancers can be said to act by 'sensing' the state of the cell, i.e. binding the relevant factors when present, by integrating the multi-factorial signal and then presenting it at the promoter (Dynan, 1989). The specificity of an enhancer, therefore, depends upon the sites comprising it, and its function relies upon the presence of the factors binding at those sites.

The enhancers described all have a positive effect upon transcription. There is a class of enhancer-like elements, first discovered in yeast, that confer negative control: these 'silencers' also function at a distance and in an orientation independent manner (Brand *et al.*, 1985). Negative regulatory elements are known from a large number of eukaryotic genes, some having silencer properties, some with stricter positional requirements; e.g. mouse Ig heavy chain gene (Imler *et al.*, 1987); chicken ovalbumin gene (Gaub *et al.*, 1987); human retinol-binding protein gene (Colantuoni *et al.*, 1987); rat myosin heavy chain (Bouvagnet *et al.*, 1987); human α-interferon (Kuhl *et al.*, 1987); *Drosophila Ubx* (Qian *et al.*, 1991). A characteristic of negative regulation in many systems is that it acts by repressing expression in inappropriate cell types (Atchison, 1988). An example is the expression of the myosin heavy chain gene which is repressed in all cell types but muscle cells (Bouvagnet *et al.*, 1987). Complex expression patterns may be produced by a combination of positive and negative regulatory elements as is found in the *Drosophila decapentaplegic* gene involved in dorso-ventral axis specification (Huang *et al.*, 1993). Negative regulation is often found in circumstances where rapid induction of gene expression is required, such as in the activation of survival pathways in response to insult (Lee *et al.*, 1992). Negative regulation can be achieved in a variety of ways; any step along the pathway leading from polymerase assembly at the promoter to enhancement and elongation is a potential target for repression (Herschbach and Johnson, 1993).

1.5.2 trans *regulation*

Both upstream promoter elements and enhansons bind *trans*-acting regulatory proteins in a sequence-specific manner. As mentioned, there are in general two classes of proteins, in some cases binding similar sites. There are ubiquitous, general transcription factors such as Sp1, Oct-1 and CTF (Kadonaga *et al.*, 1987; Sturm *et al.*, 1988; Mermod *et al.*, 1989) and there are specific factors restricted to particular cell types, tissues or regions such as B-cell-specific Oct-2, the pituitary-specific Pit-1 or the *Hox* proteins (Clerc *et al.*, 1988; Ingraham *et al.*, 1988; McGinnis and Krumlauf, 1992). The situation is complicated somewhat by the existence of factors that whilst ubiquitously expressed are only *active* in a subset of cells, e.g. *dorsal*, NF-κB. These

are grouped with the specific factors. These two classes of transcription factor act together to bring about high levels of appropriate gene expression.

Co-transfection assays are used to study the function of a transcription factor: both the gene encoding the factor and a reporter gene, linked to an appropriate binding site, are introduced into the same cell where the factor binds to its *cis*-element thereby regulating transcription of the adjacent reporter gene. Assays of this type are employed in dissecting the transcription factor and determining which regions of the protein are responsible for its various activities (e.g. Ali and Bienz, 1991). In this way specific protein domains have been identified as domains of DNA-binding, transcriptional activation or protein-protein interaction. Transcription factors are often classified by the type of DNA-binding domain that they possess (Struhl, 1989b). I have extensively discussed the homeodomain proteins and mentioned that homeodomains themselves exist in several subclasses such as *Antennapedia*-like, *paired*-like and POU. In addition to the homeodomain some of the best characterised DNA-binding motifs are the zinc-finger (Miller *et al.*, 1985; Kadonaga *et al.*, 1987; Neuhaus *et al.*, 1992), basic helix-loop-helix (bHLH; Murre *et al.*, 1989a; Jiang and Levine, 1993; Gibson *et al.*, 1993), basic-leucine zipper (bZIP; Johnson *et al.*, 1987; Saudek *et al.*, 1991) and forkhead domains (Weigel and Jäckle, 1990; Monaghan *et al.*, 1993). Other, lesser studied, DNA-binding motifs include the MADS box (Ma *et al.*, 1991), the HMG domain (Laudet *et al.*, 1993) and the helix-span-helix (Williams and Tjian, 1991b).

Positive and negative modes of regulation are most often distinguished by the transcription factors bound by a particular regulatory region; there are, however, situations where the simple relationship between activator and repressor does not apply and both functions may reside in the same molecule (Diamond *et al.*, 1990). Functional dissection of transcription factors has led to a view of them as modular structures, comprising a number of domains of largely independent function (Keegan *et al.*, 1986; Giguere *et al.*, 1986; Evans, 1988; Mitchell and Tjian, 1989; Ransone *et al.*, 1990). As mentioned, the DNA binding domain is distinguishable as an independent unit of varying nature. Additional domains such as those mediating transcriptional activation have not been so well characterised (Johnston and Dover, 1988; Hollenberg and Evans, 1988). Specific structures have not been determined for these motifs and many are defined by general amino-acid content. The best known examples are the so-called 'acidic-blob', proline-rich and glutamine-rich activation domains (Courey and Tjian, 1988; Mermod *et al.*, 1989; Courey *et al.*, 1989). A functional mechanism has not yet been described for these regions but it has been proposed that the negatively charged region acts as an interface for interaction with the general transcriptional machinery (Ptashne, 1988; Sigler, 1988). Recent work has questioned the significance of charge-

interactions, suggesting that the clustering of acidic amino acids in activation domains may be required for structural conformation (Leuther *et al.*, 1993). Further types of activation domains have been defined from a variety of factors and it seems that they may function by diverse mechanisms (Tasset *et al.*, 1990; Sutherland *et al.*, 1992; Quong *et al.*, 1993; Attardi and Tjian, 1993).

Induction of gene expression at the transcriptional level can be mediated through inducible transcription factors acting in either a positive or negative fashion. Increase in transcription can be a result of direct activation by the induced factor or by derepression upon induction. The serum response element found in the *c-fos* gene, the glucocorticoid response element (GRE) and the heavy metal response elements of the metallothionein genes are examples of inducible enhancer elements (Jantzen *et al.*, 1987; Culotta *et al.*, 1989; Treisman, 1990). These sequences interact with proteins whose activity in some way depends upon an inducing agent. The glucocorticoid response element binds the glucocorticoid receptor, a member of a large group of related molecules comprising the steroid hormone receptor super-family (Evans, 1988). These are the products of a group of related genes that encode inducible transcription factors. The steroid-hormone nuclear receptors are characterised by the ability to directly interact with a ligand, their inducing agent, thereby modulating their DNA-binding properties (Wahli and Martinez, 1991). Specific receptors have ligands including oestrogen, thyroid hormones and various retinoids (Evans, 1988; Giguere and Evans, 1990). Other inducible transcription factors have a less direct relationship with their inducing agent; they may be phosphorylated (or modified in some other way), by way of a signal transduction cascade from a cell surface receptor, eliciting responses such as a change in their DNA-binding affinity (Rivera *et al.*, 1993). Alternatively they may be transported to the nucleus in response to a receptor signal as is the case for the *dorsal* gene product of *Drosophila* (Govind and Steward, 1993). Derepression as a mechanism of activation may be brought about by either ligand-dependent loss of DNA-binding activity by a protein imposing a negative effect (Thompson *et al.*, 1992), or by *de novo* synthesis or activation of an activating molecule competing for the same site as a repressor (Kimura *et al.*, 1993).

1.5.3 Transcriptional regulation of the Homeobox genes

Homeobox-containing genes encode transcriptional regulators involved in positional specification during embryogenesis, as I have described. Their role in providing positional information relies upon the precise spatial and temporal distribution of active forms of their products. Region-specific activity may be regulated

at several levels including transcription, translation, post-translational modification and nuclear localisation. A large amount of work on embryonic patterning genes in *Drosophila* has provided examples of all such mechanisms, often in combination with one another, however the primary mechanism for a large majority of genes appears to be transcriptional regulation.

Analysis of the pattern formation process in the early *Drosophila* embryo reveals cascades of gene activation in which the products of many genes are transcriptional regulators that directly control expression of 'downstream' genes, i.e. those genes acting next in the cascade: many of the downstream genes are themselves transcription factors (Ingham, 1988; Thisse and Thisse, 1992). Genetic studies enabled the identification of interacting genes and in many cases these interactions have now been characterised at the molecular level. *In situ* expression studies enable identification of candidate downstream and upstream genes where genetic studies provide no clues or are insufficiently advanced as in many other organisms.

The early *Drosophila* embryo comprises a syncitium; a multi-nucleate, non-cellularised space in which molecules are free to diffuse. In this 'syncitial blastoderm' the sequential expression of three classes of genes generates the basic segmental divisions of the embryo: these are the maternal, gap and pair-rule genes. Maternal genes such as *bicoid*, a homeobox gene, and *nanos* are transcribed from messenger RNA located in the egg at the anterior and posterior ends, respectively. A concentration gradient of bicoid protein is established along the anterior-posterior axis and has been shown to provide cues for the correct expression of the gap gene *hunchback* in a defined region of the A-P axis (Tautz, 1988; Driever and Nüsslein-Volhard, 1988). The spatially restricted expression of *hunchback* (*hb*) is achieved by concentration dependent transcriptional activation by bicoid protein. The requirement for a particular threshold level of bicoid is due to a series of bicoid-binding *cis* elements upstream of *hb* (Driever and Nüsslein-Volhard, 1989). Both the number and the affinity of these sites determine the level of the threshold and therefore the position of the *hb* expression domain within the bicoid gradient (Struhl *et al.*, 1989a). In many cases it is known that the complex expression patterns of developmental genes are regulated at the transcriptional level. Enhancers with varying function can co-operate in the control of a single gene to regulate its expression in different domains within the embryo and at different times of embryogenesis. A good example of such a phenomenon can be seen in the transcriptional regulation of the *Drosophila ftz* gene, various aspects of which are co-ordinated by different enhancers (Dearolf *et al.*, 1990). Analysis of the *cis* regulatory elements influencing HOM-C/*Hox* genes is complicated

by the tight clustering of these genes which may enable the use of individual enhancers by multiple genes. The *cis* regulation of several mouse *Hox* genes has been studied revealing a situation similar to the multi-component regulation of *Drosophila* homeotic genes. Transgenic mice have been used extensively to reproduce endogenous expression patterns with reporter gene constructs. The most commonly used reporter is the bacterial lacZ gene encoding β-galactosidase. Localised expression of β-galactosidase can be visualised by staining embryos with the chromogenic substrate X-gal. An example typical of much of this work is the study of *Hox*a-7 (formerly *Hox*-1.1) expression (Püschel *et al.*, 1991). In this study correct regulation of *Hox*a-7 was shown to be possible outwith the context of the *Hox*A cluster. Dissection of the DNA flanking the gene identified elements regulating position-specific expression in all tissues, confining expression to a domain limited by anterior and posterior boundaries. Further elements restricted expression to sclerotomal cells and to prevertebrae within this broad domain. It was shown that different elements were required at different stages of embryogenesis "reflecting different developmental decisions". Similar studies have been performed on the regulation of several *Hox* genes (Bieberich *et al.*, 1990; Kress *et al.*, 1990; Whiting *et al.*, 1991). Elucidation of the genetical hierarchy of vertebrate development, equivalent to the known cascade of interacting genes in *Drosophila*, is at the earliest stage. Recent identification of the putative 'segmentation' gene *Krox*-20 as a regulator of the homeobox gene *Hox*b-2 is the first step in an attempt to tie together the increasing number of known developmentally expressed genes into a 'body-building' system.

1.6 Aims

Msx-1 has advantages over many other homeobox genes in regard to the study of its transcriptional regulation. As a single gene, not part of a multi-gene complex as are the *Hox* genes, it is unlikely to have an especially complex arrangement of *cis* regulatory elements and may be structured much more like the canonical eukaryotic gene of figure 1.2. Regulatory elements found in the vicinity of the gene are almost certain to regulate *Msx*-1 and not several adjacent genes which is a distinct possibility with the genes of the *Hox* clusters. The modular paradigm of *cis* regulatory control, discussed in section 1.5.3, in which a gene develops a complex expression pattern by responding to the activity of individual enhancer elements, functioning at different times and locations within the embryo, provides a conceptual model for *Msx*-1 transcriptional regulation. *Msx*-1 has a complex expression pattern in the developing embryo but one which evolves from more simple origins. The early expression pattern

of *Msx*-1 appears to be generated in response to position along the medio-lateral axis. Consequently one might predict the presence of *cis* elements responsive to notochord or neural-tube induced pathways in the *Msx*-1 regulatory regions. Early expression is also in the neural crest and is probably defined by interaction between lineage-specific factors and particular regulatory regions. In the second phase of *Msx*-1 expression there are many examples of response to epithelial-mesenchymal interactions. It is likely that all are effected by way of the same regulatory elements and possibly along a common pathway modulated in a manner appropriate to the system concerned, be it tooth, limb, heart or face. Both positive and negative regulatory influences are likely to interact in the generation of these patterns.

The 5' proximal promoter region of *Msx*-1 was made the subject of this study in an attempt to elucidate the regulatory mechanisms controlling expression of this gene. The aim of this work is to characterise features of the transcriptional regulatory machinery acting to generate the spatially and temporally diverse expression pattern of the *Msx*-1 gene in the developing mouse embryo. I hoped to exploit the conservation of function seen in other homeobox-gene regulatory systems and considered likely for *Msx*-1, given the widespread phylogenetic distribution of *Msx*-1 cognates. Principally the *cis* regulatory elements are the subject of enquiry, with the view that the discovery of such will lead to identification of *trans* regulatory factors controlling *Msx*-1 expression and encoded by genes upstream of *Msx*-1 in the genetic hierarchy regulating embryogenesis.

Chapter 2

Materials and Methods

2.1 Bacterial cell culture

2.1.1 Bacterial media

L-Broth (Luria-Broth) and agar - Per litre: 10g bacto-tryptone (Difco), 5g bacto-yeast extract (Difco), 10g NaCl, pH to 7.5. L-agar contains 15g agar (Difco) per litre in addition.

H-agar - Per litre: 10g bacto-tryptone (Difco), 8g NaCl, 12g agar (Difco) pH to 7.3. H-agar plates are used for the *lac*Z blue/white test described (section 2.1.3). IPTG is added to the agar as an inducer for the β-Galactosidase gene (0.024%). The agar is also supplemented with the chromogenic substrate X-gal (5-bromo-4-chloro-3-indolyl-β-D-galactoside) which provides the blue colour (0.02% from 2% stock in dimethylformamide).

Terrific Broth (Tartoff and Hobbs, 1987) - Per litre: 15g bacto-tryptone (Difco), 30g bacto-yeast extract (Difco), 5 ml glycerol, 1/10 volumes of 1M K_2HPO_4 added immediately before use. This medium produces high density cultures and was routinely used for plasmid preparations.

TYP-Broth - Per litre: 16g bacto-tryptone (Difco), 16g bacto-yeast extract (Difco), 5g NaCl, 2.5g K_2HPO_4 . This medium is used in the production of ssDNA by phagemid rescue.

Media supplements - media were supplemented with antibiotics where described. Concentrations for each antibiotic are: Ampicillin (Sigma), 100 μg/ml; Kanamycin (Sigma), 30 μg/ml; Tetracyclin (Sigma), 25 μg/ml. Stocks were made at 1000x and stored at -20°C. IPTG and X-gal were added as described above.

All media and glassware were sterilised by autoclaving, supplements were filter sterilised. Bacterial cultures were grown at 37°C with constant agitation on an orbital shaker. Permanent stocks of bacteria containing particular plasmids were maintained by freezing cultures in 20% glycerol at -70°C.

2.1.2 *Bacterial strains*

JM83 - *ara*, Δ(*lac-pro* AB), *rsp*L, ϕ80*lac*ZΔM15, (rk^+, mk^+) (Yanish Perron et al, 1985). This strain was used as a host for pUC based plasmids. The *lac*ZΔM15 gene is integrated into the host genome.

XL1-Blue - *recA1*, *endA1*, *gyrA96*, *thi-1*, *hsdR17*, *supE44*, *relA1*, *lac*, [F' *pro*AB, *lacI*^{q}Z*ΔM15*, Tn*10* (tet^r)] (Bullock et al, 1987). This strain was also used as a general host for pUC based plasmids. The presence of the F' plasmid permits infection by M13-based helper phage, via the sex pili, utilised during phagemid-rescue ssDNA production.

2.1.3 *DNA vectors*

Vectors used were pTZ-18R (Pharmacia) a general purpose vector; pGEM-7Zf(-) from Promega, for its restriction sites suitable for unidirectional exoIII deletion and for production of single-stranded DNA by helper phage rescue; pGEM-5Zf(-) from Promega, as a cloning vector with a NcoI site in the multiple cloning site; pBluescript II SK(+) from Stratagene which was used as an general-purpose vector with a large choice of restriction sites.

2.1.4 *Preparation of E.coli cells competent for DNA transformation*

The conditions under which *E.coli* cells are competent to take up plasmid DNA molecules have been well studied and optimised (Hanahan, 1985). *E.coli* cells (JM83, XL1-Blue, etc.) were made competent for DNA transformation by the following method. 100 ml of L-Broth media plus 10 mM $MgCl_2$ were inoculated with a single bacterial colony and incubated at 37°C; on an orbital shaker, until the optical density of the culture at 560 nm (OD_{560}) was approximately 0.5. The cells were chilled on ice for 10 minutes and all subsequent steps performed at 4°C. Cells were pelleted at 2500 rpm for 10 minutes then gently resuspended in 33 ml of cold, filter-sterilised FSB (10 mM $CaCl_2$, 10 mM potassium acetate, 100 mM RbCl, 45 nM $MnCl_2$, 3 mM hexamine cobalt chloride, 10 % Glycerol, pH 6.4). After 10 minutes the cells were again pelleted, as before, and resuspended in 8 ml FSB plus 280 µl DMSO (spectroscopic grade). After 10 minutes a further 280 µl of dimethyl sulfoxide (DMSO) was added and 200 µl aliquots of the cells were quick frozen in liquid nitrogen. The aliquots were then stored

at -70°C. This method gave transformation efficiencies of up to 4 x 10^7 colonies/μg plasmid DNA.

2.1.5 Transforming competent cells

A 200 μl aliquot of competent cells was thawed slowly on ice and the DNA to be transformed was added (10-50 ng in approx. 10 μl). Cells and DNA were mixed and placed on ice for 30 minutes. After this period the cells were heat shocked for exactly 2 minutes at 42°C then returned to ice for a further 5 minutes. 500 μl of 2 x TY medium was added and the cells incubated at 37°C for 30 minutes. They were then plated on media supplemented with appropriate antibiotic and indicators and incubated, with the dishes inverted, at 37°C overnight.

2.2 Isolation and purification of DNA

2.2.1 Large Scale preparation of plasmid DNA

A single colony was picked and used to inoculate 500 ml of Terrific Broth supplemented with the appropriate antibiotic. This was incubated, on a shaker, at 37° C overnight. The cells were pelleted at 6000 rpm and resuspended in 20 ml GTE (50mM Glucose, 25mM Tris Cl pH 8.0, 10 mM EDTA pH 8.0) plus lysozyme (10mg/ml). 40 ml alkaline SDS (0.2M NaOH, 1% SDS) was added and the lysate placed on ice for 5 minutes. 30 ml of 3M Sodium acetate (pH 5.0) was added, the solution mixed by inversion and the cell debris pelleted by centrifugation at 12000 rpm for 30 minutes. The supernatant was strained through muslin and the DNA precipitated by addition of 0.6 volumes of isopropanol. DNA was pelleted by centrifugation at 8000 rpm for 20 minutes. The supernatant was discarded and the pellet dried under vacuum. The dried pellet was resuspended in 22 ml T.E. (10 mM Tris.Cl, 1 mM EDTA) to which was added 24g of caesium chloride (CsCl) and 2 ml Ethidium bromide (10mg/ml). The CsCl was dissolved by shaking to give a solution of refractive index 9.3-9.45. The plasmid DNA was banded in a polyallomer tube by centrifugation at 40000 rpm, 20°C, overnight in a vertical rotor. The less dense upper band of plasmid DNA was visualised under UV light (300nM) and removed from the tube with a hypodermic syringe. The ethidium bromide was removed by repeated extractions with water-saturated butan-2-ol until the aqueous phase is colourless. The DNA was then precipitated by addition of 2.5 volumes of 75% ethanol and pelleted by centrifugation at 12000 rpm for 15 minutes.

2.2.2 *Small scale preparation of plasmid DNA*

Typically, 3 ml of Terrific-Broth, supplemented with the appropriate antibiotic, was inoculated with a single colony and incubated at 37°C overnight on a rotary shaker. 1 ml of this culture was transferred to an eppendorf tube and the cells were spun down by centrifugation in a bench top centrifuge for 1 minute. All centrifugation in this method was performed at 12000 rpm in a bench top centrifuge. The supernatant was discarded and the pellet resuspended in 100µl of GTE. 200µl of a fresh solution of alkaline SDS was added and mixed by inversion of the tube several times. The tube was placed on ice for 5 minutes. 150 µl of 3M potassium acetate (pH 4.8) was added and again mixed by inversion. The tube was then placed on ice for a further 5 minutes. The preparation was centrifuged for 5 minutes followed by transfer of the supernatant to a clean tube. 450µl of 100% ethanol was added to the supernatant and the tube was briefly vortexed. The preparation was centrifuged for 15 minutes. The supernatant was discarded and the pellet washed in ice-cold 70% ethanol. The ethanol was removed and the pellet dried under vacuum. The dried pellet was resuspended in an appropriate volume of water or TE (10mM Tris.HCl, 1mM ethylenediaminetetra-acetic acid (EDTA)).

2.2.3 *Single-stranded DNA rescue from phagemid carrying cells*

A 1-2 ml culture was set up in TYP-broth plus the appropriate antibiotic (usually 50 µg/ml ampicillin) and grown at 37°C overnight. 100 µl of this culture was used to inoculate 5 ml TYP (in a 50 ml tube) which was then incubated at 37°C for 30 minutes. The helper phage M13K07 (Vieira and Messing, 1987) was added at a mutiplicity of infection of 10-20 (with the assumption that the 30 minute culture has 5 x 10^7 - 1 x 10^8 cells/ml) and incubation continued for a further 30 minutes. At this point Kanamycin was added (to a concentration of 25 µg/ml) to select for cells infected by M13K07 which carries a Kanamycin resistance gene. Incubation was continued for a further 5 hours. The cells where then pelleted by centrifugation at 12000 rpm for 15 minutes. The supernatant was removed and subjected to a repeat of the centrifugation. 0.25 volumes of PEG precipitation solution (20 % polyethylene glycol, 3.75M ammonium acetate) was added to the supernatant which was then stored on ice for 30 minutes. The single-stranded DNA (ssDNA) was pelleted by centrifugation at 12000 rpm for 15 minutes and care was taken to remove all traces of the supernatant. The ssDNA was resuspended in 400 µl of T.E., 400 µl of

chloroform:isoamyl alcohol (24:1) was added and the mixture vortexed for 1 minute. The preparation was then centrifuged at 12000 rpm for 5 minutes. The upper (aqueous) phase was transferred to a clean tube (taking care not to transfer any of the interface) and 400 µl phenol:chloroform (1:1) (saturated with T.E.) was added. This was vortexed and centrifuged as before. This extraction was repeated until there was no material visible at the interface. The aqueous phase was extracted with an equal volume of chloroform, vortexed, centrifuged and repeated. The aqueous phase was transferred to a fresh tube. 200 µl 7.5M ammonium acetate and 1.2 ml ethanol were added, mixed by inverting the tube and stored at -70°C for 15 minutes. The preparation was centrifuged at 12000 rpm for 15 minutes, the supernatant removed and the pellet washed in 80% ethanol and dried under vacuum. The resulting ssDNA pellet was resuspended in 20 µl T.E.

2.2.4 DNA purification and precipitation

For use as a substrate in many reactions DNA must be purified by removal of proteins and salts that would otherwise interfere with the action of the enzymes used. The common way to do this is by a combination of Phenol:Chloroform extraction and ethanol precipitation. Extraction with Phenol:Chloroform (repeated if necessary) removes proteins and the DNA can be recovered from the aqueous layer, into which it partitions, by precipitation with ethanol. Typically, 2 volumes of Phenol:Chloroform:Isoamyl alcohol (25:24:1) are added to the preparation. This is vortexed for 1 minute followed by 2 minutes centrifugation in a microcentrifuge (12000 rpm, used for all centrifugation steps in this protocol). The upper, aqueous layer is transferred into a clean eppendorf tube and 2 volumes of diethyl ether are added. Care should be taken when transferring the upper phase not to remove any of the interface. The tube is thoroughly vortexed then the upper layer discarded and the tube heated to 65°C for 1 minute to remove traces of diethyl ether. Sodium Acetate is added to a final concentration of 0.3 M (pH 5.2). Next, 2.5 volumes of 100% ethanol is added and the tube vortexed. The tube is placed at -70°C for 15 minutes, to allow the DNA to precipitate, and then centrifuged for 15 minutes to pellet the DNA. The supernatant is discarded and the pellet washed once or twice with 80% ethanol. The pellet is then dried under vacuum and suspended in TE., or a buffer appropriate to the next step. For solutions of DNA with little or no protein the addition of sodium acetate and ethanol precipitation is sufficient.

2.3 Enzymatic manipulation of DNA

2.3.1 Restriction endonuclease digestion

The routine digestion of plasmid DNA by site-specific endonucleases was performed using enzymes purchased from Amersham International plc., Boehringer Mannheim GmbH and New England Biolabs. Reactions were performed according to manufacturers instructions using reaction buffers supplied.

2.3.2 DNA ligation

Construction of recombinant DNA molecules requires the joining of fragments, most frequently vector and insert. This is achieved using the DNA Ligase enzyme encoded by bacteriophage T4. T4 Ligase was purchased from Boehringer Mannheim GmbH and used in the buffer provided (20 mM Tris-HCl, 10 mM $MgCl_2$, 10 mM DTT, 0.6 mM ATP, pH 7.6). In a typical ligation reaction approximately 25 ng of vector DNA was combined with insert DNA to give an insert:vector molar ratio of 1:1, 2:1 or 5:1. Each reaction was performed in a total volume of 10 µl with 2 units of ligase and incubated at 16°C overnight. Multiple reactions were used with a variety of vector:insert concentrations along with reactions containing vector alone and fragment alone. This provides increased likelihood of optimisation of the reaction and enables assessment of the level of any vector re-ligation and vector background within the fragment preparation.

2.3.3 Generation of double-stranded nested deletions

The Pharmacia double-stranded nested deletion kit was used to produce such deletions to facilitate sequencing of large inserts. This kit is based on the method of Henikoff (1984). In brief, deletions are generated using exonuclease III (exoIII), a 3' exonuclease, in conditions favouring a controlled digestion rate. Unidirectional deletions can be performed as exoIII requires a double-stranded 3' end as a start point for digestion, whether a blunt ended molecule or one with a 5' overhang. 3' overhanging ends are resistant to digestion. A plasmid carrying the insert to be deleted is linearised with two restriction enzymes cutting at closely positioned sites. One restriction site is used to give a 3' overhanging end to the plasmid sequence and one to give a 5' overhang (or blunt end) to the insert sequence. A deletion reaction is set up in which the 3' end of the insert sequence is progressively digested and samples are

removed at regular time intervals. The remaining single-stranded overhang of these nested deletions is removed by digestion with S1 nuclease. The extent of deletion can be analysed at this point by agarose gel electrophoresis and samples corresponding to the desired extent of deletion circularised by T4 ligase and transformed into competent bacteria to produce a 'deletion library' from which individual clones can be sequenced.

ExoIII deletion of the pDEL3 plasmid carrying the EcoRI-SmaI fragment from the 5' flanking region of *Msx*-1 was performed in 50 mM NaCl (a departure from the manufacturers recommended conditions) at 37°C. Conditions were established to give a suitable rate of deletion by several trial experiments. 2μg of DNA were added to the reaction mixture. Samples were removed every 5 minutes after addition of the exonuclease and a total of 8 time points were taken. After S1 treatment and ligation the deletions were transformed into XL1-blue cells and plated on L-amp agar. Single colonies were then picked and streaked onto L-amp/tet agar. Tetracyclin selection was used to ensure maintenance if the F' episome, essential for single-stranded rescue.

2.3.4 Polymerase Chain Reaction on plasmid templates

The Polymerase Chain Reaction (PCR) was employed to amplify regions from plasmid inserts. A typical reaction contained 30 ng plasmid template, 100 ng of each oligo, 50 nM for each dNTP (A,T,G & C), 1 x reaction buffer (500 mM KCl, 100 mM Tris.Cl (pH 9.0), 1% Triton X-100) and 0.2 units Taq DNA polymerase (Promega) in a total volume of 50 μl. The various components were added and mixed, with the template added last. The reaction was overlaid with paraffin oil to prevent evaporation and placed on the thermal cycler. A program of 25 cycles was used, each one consisting of 30 seconds denature at 94°C, 30 seconds oligo annealing at the calculated temperature and 1 minute strand extension at 72°C. The annealing temperature was calculated by allowing 2°C per A or T and 3°C per G or C in the oligo, then summing the total. Oligonucleotides of 18-22 bases were used giving annealing temperatures in the order of 35-55°C.

2.3.5 Agarose-gel electrophoresis

Electrophoretic separation of DNA through agarose gels is a standard technique in molecular biology. It was used to analyse the products of enzymatic manipulations such as restriction digestions, ligations, deletions and PCR. Gels of 0.5-3 % agarose (Sigma, Type II) in 0.5 x TBE (45 mM Tris-borate, 1 mM EDTA) were cast with a depth of ~5 mm and in a variety of sizes: mini-gel - 5 cm x 7 cm; midi-gel -

11 cm x 14 cm. Larger, less concentrated gels were used to separate larger fragments while small gels were used to separate smaller fragments. Typically 1 % agarose was used except where particularly small fragments were separated, when 3 % agarose was used. DNA samples were loaded with addition of 1/10 volume of loading buffer (15 % Ficoll, 0.25 % Orange G, 0.25 M EDTA in 10 x TBE). Orange-G acts as a visible marker for the migration front of small (200-500 bp) DNA fragments. A voltage of 10V/cm (maximum) was applied for the requisite period after which the gel was stained with ethidium bromide (1-2 drops of a 10 mg/ml solution). The ethidium bromide stained DNA could be visualised by exposure to ultraviolet light (254 nm) whereupon gels were photographed using a UVP video camera and Mitsubishi video copy processor.

2.3.6 DNA recovery from agarose gel

Specific fragments identified by agarose-gel separation are often required for further steps such as cloning or use as probes. In such cases purification of the fragments from the gel can be achieved.

Typically a preparative gel of 0.8% SeaKem GTG agarose (FMC Bioproducts) was cast and separation performed as described (section 2.3.5). DNA restriction fragments were recovered from preparative agarose gels using one of two methods. In both, the bands were visualised by ethidium bromide staining of the gel and UV illumination. The required bands were excised using a clean scalpel blade. Method 1) used the COSTAR® SPINEX filter centrifuge unit according to manufacturers recommendations. The excised agarose slice containing the chosen DNA fragment was finely chopped and placed in the SPINEX tube. It was centrifuged in a benchtop microcentrifuge at 12 000 rpm for approximately 10 minutes after which time the liquid component of the gel, including the DNA in solution, had passed through the filter into the lower tube. The DNA was then phenol extracted and ethanol precipitated as described. Method 2) used the 'Geneclean Kit®' from BIO 101 according to manufacturers recommendations. This protocol exploits the DNA-binding ability of fine glass beads under appropriate conditions to isolate the DNA from the melted agarose. Both techniques provided clean DNA suitable as a substrate for the further enzymatic manipulations of the various cloning steps.

2.4 Radiolabelling DNA

2.4.1 Nick translation

Nick translation reactions were performed by the method of Rigby *et al.* (1977) using a kit from Gibco-BRL. This labelling method exploits the properties of two enzymes, *E.coli* DNA polymerase I and bovine pancreatic DNAse I. The DNAse creates intermittent nicks in one strand of the double-helix which act as a point of entry for the polymerase. Exonuclease activity of the polymerase removes nucleotides before it and re-synthesises the strand in its wake, incorporating α-dCT^{32}P radiolabelled nucleotide as it processes along the molecule. The reaction is performed at 16°C to provide a balance between sufficient strand extension speed by the polymerase and low enough levels of DNAse activity to prevent degradation.

This technique was employed for labelling the concatamerised, double-stranded oligonucleotide probes used in South-Western hybridisation. Typically, 500 ng of DNA and 65 µCi [α-^{32}P]dCTP were included in a standard reaction (according to manufacturers' instructions). A mixture of the two enzymes is added last and incubated for 1 hour. After this time the incorporation was tested using trichloroacetic acid (TCA) precipitation onto glass-microfibre filters (Whatman GF/B) and a Cerenkov counting protocol on a liquid scintillation counter.

2.4.2 Random-Primer labelling of DNA

This method was used as an alternative to nick translation for labelling double-stranded DNA molecules to be used for probes on Southern blots. The method used was based on that of Feinberg and Vogelstein (1983). The basic principle is one of strand separation by high temperature melting followed by cooling in the presence of a mixture of all possible hexanucleotides. Second strand extension from the hexanucleotide primers was performed in the presence of radiolabelled dCTP which was incorporated into the nascent strand. A Random Prime kit from BCL was used for this reaction. In brief; 50-100 ng of DNA was denatured in 10 µl of DNA by heating to 100°C; then chilled on ice, 5 µl of random prime buffer/dNTPs (4 x Klenow buffer, 200 µM dATP/dTTP/dGTP) was added followed by 4 µl of [α-^{32}P]dCTP and 1 µl of DNA Polymerase I. The reaction was then incubated for at least 1 hour at 37°C and unincorporated radio-nucleotides removed as described (section 2.4.4).

2.4.3 *Oligonucleotide labelling*

Synthetic oligonucleotides were radio-actively labelled for a variety of purposes including cycle sequencing and Grunstein-Hogness screening for sub-clones. The protocol used T4 polynucleotide kinase (PNK) which transfers the γ-^{32}P from [γ-^{32}P]ATP to the phosphate-less nucleotide at the 5' end of the oligonucleotide. Typically 5 µl reactions were performed containing 5-10 pmoles of oligonucleotide (in 1 µl), 1 µl of 5X PNK buffer (250 mM Tris.Cl (pH 7.6), 50 mM $MgCl_2$, 25 mM DTT, 0.5 mM spermidine HCl, 0.5 mM EDTA (pH 8.0)), 1 µl of [γ-^{32}P]ATP (5000 Ci/mmol), 1 µl of PNK (10 units/µl; Boehringer Mannheim) and 1 µl of water. This reaction mix was incubated at 37°C for 30 minutes after which time it was heat inactivated by 5 minutes at 65°C. Larger amounts of oligo were labelled by scaling up this basic reaction. The levels of incorporation were crudely assessed (where a more accurate measure was not required) by chromatography of a small sample of the reaction on DEAE ion exchange paper (Whatman DE81) with a solvent of 0.3 M ammonium formate. This separates unincorporated label, which travels at the solvent front, from labelled oligo which remains at the origin. A hand held counter was then used to gain an approximate measure of the percentage of incorporation.

2.4.4 *Removal of unincorporated radio-nucleotides*

After radio-labelling reactions it is desirable to remove radioactive nucleotides that have not been incorporated into the DNA. This cuts down on background signals and enables simple detection of the DNA as it is the only labelled molecule used; e.g. allows one to determine whether DNA had pelleted successfully simply by monitoring. Removal of free radio-label is achieved by size exclusion chromatography. The labelling reaction is passed over a Sephadex G-50 column (Bio-Rad Nick Column) that has been equilibrated with 1 x TE. Single nucleotides are sufficiently small to be retained by the column whereas the DNA is eluted in the second of two 400 µl fractions of 1 x TE.

2.4.5 *Detection of radioactive signals*

Two methods were employed for the detection of signals from the radioactive probes used in sequencing, southern hybridisations, mobility-shift assays and south-western hybridisations.

The first, autoradiography, was performed by placing the filter, or dried gel, next to a piece of Kodak X-OMAT™ AR X-ray film. The two were placed in an autoradiography cassette. After the appropriate exposure time the film was removed under safe-light and processed in an automatic developing machine. For the detection of weak signals, the performance of the fluorescence screens in the cassette was boosted by exposure at -70°C. Sequencing gels were examined by autoradiography exclusively, as were library screens. All other detections used a combination of autoradiography and phosphor-imaging.

Phosphor-imaging was performed using a Molecular Dynamics Phosphor Imager (Model 400B). Images scanned using this machine were analysed on a personal computer running the windows based ImageQuant™ system. Briefly, radioactive filters or gels are placed in a Storage Phosphor Screen cassette for a short exposure (approximately one tenth that required for autoradiography). The screen stores energy from the radiation that is later 'read' by laser scanning and the information retrieved and converted into an image on the computer screen. The major advantages of this system over autoradiography are reduced exposure times and sensitivity over a far greater range of values; approximately 1 x 10^4 fold as compared to 200 fold for X-ray film.

2.5 Identification of specific sequences

2.5.1 Colony hybridisation screening

Colony hybridisation is a rapid method of simultaneously screening several hundred transformed bacterial colonies to identify the plasmid or cosmid clone of interest. The method involves the removal of the components of the bacterial cell walls and the immobilisation of the remaining DNA onto a membrane on which the colonies were grown. This membrane can then be exposed to a radioactive DNA probe in conditions promoting selective hybridisation to complementary DNA sequences on the filter, enabling the identification of a colony containing the clone of interest.

The method used is based on that found in Sambrook *et al.* (1989; section 1.98) which is itself a version of the original method of Grunstein and Hogness (1975). Bacteria are plated onto nitro-cellulose (or nylon) filters and grown to give colonies of a suitable size (small if the density is high). Whatman 3MM paper was cut to size and fitted into the bottom of three plastic trays, often the lids of 20 cm x 20 cm plates. The filter paper was then saturated with one of three solutions, with any excess liquid being poured off. The solutions were 1) Denaturing solution - 0.5 M NaOH, 1.5 M NaCl; 2) Neutralising solution - 1.5 M NaCl, 0.5 M Tris.Cl (pH 7.4); 3) 2 x SSC (20 x SSC -

3M NaCl, 0.3M Sodium citrate, pH 7.4). Filters were lifted from the plate, using Millipore blunt-nosed forceps, and placed colony side up on the filter paper of the first dish, that containing denaturing solution. To ensure even treatment of the colonies it was essential to avoid bubbles between the filters and the 3MM. Filters were left on the denaturing dish for 5 minutes then transferred, with care taken to avoid any excess liquid carry-over, to the neutralising dish. After 5 minutes neutralising and 5 minutes on the 2 x SSC dish the filters were immersed in 2 x SSC and remaining bacterial debris removed by gently stroking the filter in one direction with a gloved finger. Filters were placed on dry 3MM paper and left to dry completely. DNA was fixed to the filters by sandwiching them between two sheets of 3MM and baking them at 80°C in a vacuum oven for 1 hour.

Prior to addition of the radiolabelled probe, filters were prehybridised in 'Hyb-Mix' (5 x SSC, 0.1% pyrophosphate, 0.1% SDS, 0.05% ficoll, 0.05% polyvinylpyrolidine, 0.05% bovine serum albumin, 100 µg/ml denatured salmon sperm DNA) for 1 hour at 68°C. Hybridisations were carried out in bottles in a bench-top rotary oven (Hybaid®) with the filters between sheets of nylon mesh. Labelled probe was added to the prehybridisation solution and the hybridisation continued at this temperature (or a lower temperature suitable for the probe used) overnight. The following day the radioactive hybridisation solution was removed and the filters were washed twice (30 minutes each wash) in 1 x SSC, 0.1% SDS at 68°C. Filters were then dried on 3MM paper, wrapped in Saran Wrap and autoradiographed.

2.5.2 Cosmid library screening

Screening of the cosmid library was done essentially as described in Sambrook *et al.* (1989). The library was plated on 20 cm x 20 cm plates onto Hybond nylon membrane overlaid on L-agar plus kanamycin (25 µg/ml). Plates were incubated overnight at 32°C, rather than 37°C, to produce small colonies. Duplicates were taken by placing a fresh filter on top of the one overgrown with colonies and applying gentle pressure evenly across its area. Originals and duplicates were marked using a hypodermic needle immersed in permanent ink. The duplicate was placed on a fresh agar plate and grown at 37°C for 4 hours. The membrane was processed in the same way as described for screening of plasmid colonies by hybridization with radioactive DNA probes (section 2.5.1). Colonies corresponding to the signals on the autoradiographs were picked by cutting a square of the nylon membrane ~½ cm^2 carrying the positive colony (or several colonies if it was impossible to identify an individual). This small piece of nylon was immersed in 1 ml L-broth, 15% glycerol and

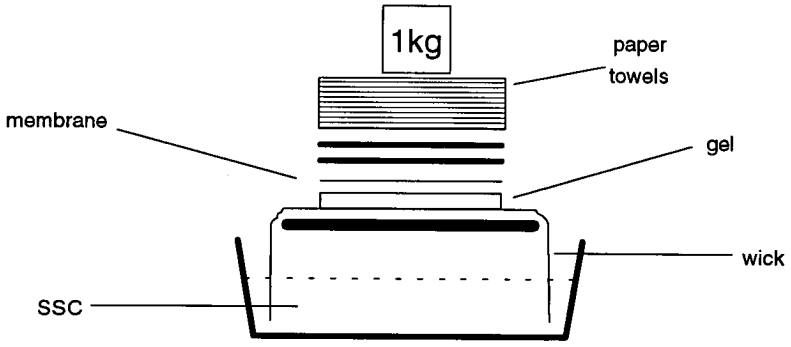
vortexed for several minutes. The resulting bacterial suspension can be stored indefinitely at -70°C. A dilution of this suspension was made and an equivalent of 0.1µl of original concentration was plated onto 82 mm nitro-cellulose discs (Schleicher and Schuell) overlaid on L-agar plus kanamycin, on each of a series of four 9 cm plates. A similar series was plated for two additional concentrations, 5 fold greater and 5 fold less than the first. These plates were grown at 32°C overnight, as before, and replica filters made. After 4 hours recovery at 37°C the replicas were processed and screened with the radioactive SacI-NcoI fragment as before. Clones were taken to an additional, tertiary screen to ensure single colony purity.

2.5.3 Southern-blot hybridisation

Transfer

Digested DNA was separated by horizontal agarose gel-electrophoresis (section 2.3.5) and denatured prior to transfer to a membrane. The gel was soaked, with gentle shaking, in 0.5M NaOH, 1.5M NaCl for 45 minutes. It was then neutralised by gentle shaking in 1M Tris, 2M NaOH for 45 minutes. The gel was inverted onto a wick of Whatmann 17MM paper soaking in a reservoir of 20 x SSC (20 x SSC = 3 M NaCl, 0.3 M Na Citrate, pH 7.4): Figure 2.1. Saran wrap was used

Figure 2.1



to surround the gel to prevent transfer of the SSC other than through the gel. A piece of nitrocellulose (Schleicher and Schuell) or nylon membrane (Hybond-N, Amersham), cut to size, was soaked in distilled water and placed on top of the gel with care taken to exclude bubbles from between the two. Two pieces of 3MM Whatmann paper of similar size were placed over the membrane and covered with several layers of paper towels. A weight (~1 kg) was placed on top of the towels and transfer allowed to procede overnight. Transfer of the DNA to the membrane occurs by capillary action. The membrane was washed briefly in 5 x SSC. DNA was fixed to the membrane in one

of two ways. Nitrocellulose membranes were air dried and baked in a vacuum oven, at 80°C for 1 hour. Nylon membranes were irradiated by UV in a Stratalinker (Stratagene).

Hybridisation

Hyridisation was performed in rotating bottles in a temperature-controlled oven (Hybaid). Each membrane was wetted in distilled water, sandwiched between two sheets of nylon mesh which were rolled up and placed in the bottle, unrolling to line the inside of the bottle. Approximately 30 ml of hybridisation mix were added (0.05% BSA, 0.05% polyvinylpyrolidine, 0.05% Ficoll, 0.1% SDS, 0.1% sodium pyrophosphate, 5 x SSC, 100 µg/ml denatured salmon sperm DNA) and the membrane was incubated at 68°C for 1 hour. This temperature was used for all southern hybridisations performed. Probes added had a specific activity of approximately 10^9 cpm/mg, they were boiled for 5 minutes and added to the hybridisation solution giving 3-6 x 10^3 cpm/ml. The hybridisation was incubated overnight at 68°C after which the radioactive solution was discarded and the membrane rinsed 3 times in fresh wash solution (2 x SSC, 0.2% sodium pyrophosphate, 0.2% SDS) for 30 minutes at 68°C. After washing, the background non-specific radioactivity should have been largely removed leaving the signal. This was assessed using a hand held monitor and if no further washing was deemed necessary the excess liquid was drained from the membrane and it was wrapped in Saran wrap. The signal was detected by autoradiography.

2.6 Sequencing

The chain termination method of Sanger (1977) was used in all sequence determination. Two applications of this method were used, the standard chain-termination technique using Sequenase® (USB) and a 'Cycle sequencing' method using Taq polymerase (GIBCO BRL; dsDNA Cycle Sequencing System).

2.6.1 Sequenase sequencing

Standard chain-termination sequencing was performed using Sequenase (United States Biochemicals) according to manufacturers recommendations. Termination products were labelled by incorporation of [α-^{35}S]dATP. Two types of template were used; 1) single stranded DNA produced from phagemid vectors; 2) double stranded DNA denatured with NaOH.

Single-stranded templates were generated by helper phage rescue of phagemids carrying the f1 origin (section 2.2.3). Primer annealing, labelling and chain-termination reactions were performed according to the Sequenase manual. Primers were artificially synthesised and 5-10 ng was used per reaction.

Double-stranded templates were CsCl banded plasmids denatured by incubation of 4 µg DNA with 100µl of denaturing solution (200 mM NaOH, 0.2 mM EDTA) at 37°C for 30 minutes. Denatured DNA was precipitated by addition of 1/10 volume 3M Sodium Acetate and 2.5 volumes ethanol. The precipitate was washed with 70% ethanol and dried under vacuum. DNA was resuspended in 7 µl H_2O to which was added 2 µl reaction mixture (Sequenase kit) and 1 µl primer (10 ng). The mix was incubated at 37°C for 30 minutes to permit oligo annealing. Labelling and chain-termination reactions were performed according to Sequenase manual.

2.6.2 Taq polymerase 'Cycle sequencing'

Regions amplified from plasmid clones by polymerase chain reaction (PCR; section 2.3.4) were sequenced using a 'dsDNA Cycle-sequencing' kit from GIBCO-BRL. This technique sequences by the chain-termination method but uses Taq polymerase and a thermal cycling protocol to amplify the signal in a linear fashion. An oligonucleotide, either one of those used for PCR amplification or an internal one, is added together with the linear template to a reaction mix containing deoxynucleotides and dideoxynucleotides. Thermostable Taq polymerase was added and several rounds of strand melting, annealing and extension were performed on a thermal cycler. The initial chain-terminated extension product was melted from the template and a new product synthesised in the next cycle. Very small amounts of template DNA can be sequenced in this way as the product signal is amplified by the number of cycles performed. The labelling for this sequencing was achieved by labelling the 5' end of the oligonucleotide with [γ-^{33}P]ATP using polynucleotide kinase (section 2.4.3). This alternative isotope of phosphorous, ^{33}P, was used as it gives sharp intense bands after several hours autoradiography. This is due to a higher specific activity than that of ^{32}P, but a lower emmision energy than the other phosphorous isotope, more like that of ^{35}S. Sequencing reactions were set up by following manufacturers instructions. The recommended 'quick cycle' was used with linear PCR-product DNA as template. Briefly, template DNA was melted at 94°C for 2 minutes followed by 30 cycles of sequencing/amplification (5 seconds at 94°C, 5 seconds at 55°C, 5 seconds at 72°C). The annealing temperature of 55°C was found to be suitable for most oligos used but was altered slightly in some cases.

2.6.3 *Electrophoresis of sequencing products*

Sequencing reaction products were separated by electrophoresis through a 6% polyacrylamide, 6 M Urea, 1 x TBE denaturing gel. The gel was run at 27 watts (power limiting) in 1 X TBE buffer. To obtain maximum information from a single reaction it was usually necessary to run two gels, one for approximately two hours and the second for approximately four. After electrophoresis the gel was fixed in 10% methanol/10% acetic acid for 15 minutes then transferred to Whatmann 3MM paper and dried on a vacuum gel drier. The dried gel was autoradiographed overnight.

2.6.4 *Sequence analysis by computer*

Computer analysis and manipulation of sequence data was used for a variety of purposes; management of sequencing projects; sequence comparison, sequence-library searches and searches for known *cis*-elements.

Individual gel-reads of sequence data were assembled in one of two ways. Sequence obtained from the deletion series of the mouse 5' region was input into the Staden automatic Contig-Assembly package (Amersham Staden-Plus; Release 1, Version 6) where overlapping ends were joined to form large stretches of contiguous sequence. This approach, although based on the ordered set of nested deletions, allows for a random input of sequences thereby accommodating the varied population of deletions found at any one time point. The second method employed the GCG program LINEUP. This program allows the user to input individual sequences one on top of the other and to slide them along relative to one another to obtain the optimum match by eye. This method is only appropriate when the relationship between each gel-read is clear, e.g. when they are known to be overlapping due to use of custom-made oligonucleotides to prime the sequencing reactions. This was the approach adopted in the sequencing of the majority of the human 5' sequence and LINEUP was used in the assembly of this sequence.

The sequence-databases accessed were updated versions of the GenBank and EMBL databases maintained at the UK Human Genome Mapping Project Resource Centre (UK HGMP-RC) at Harrow in Essex. They were searched using the BLAST program available on site. Transcription factor binding sites were searched for by comparison with the Transcription Factor Database (TFD) (Ghosh, 1992) using the SIGNAL SCAN program (Prestridge, 1991), also at the UK HGMP-RC.

All computing was performed on Sun MicroSystems Sparc workstations running Unix operating system. Postscript files of plot output from GCG programs were printed on an Apple Laserwriter.

2.7 Protein extracts

2.7.1 Tissue culture conditions

B16 and H3M cells were cultured to provide protein for *in vitro* binding studies and to enable transfection with reporter gene constructs. Both cell lines were grown under similar conditions. RPMI 1640 culture medium (Flow Laboratories) was used supplemented with 10% fetal calf serum (Gibco Bio-Cult) and cells were incubated at 37°C with 5% CO2. Cells were grown in plastic culture flasks (Nunc: typically 175cm^2) as monolayer cultures and split when confluent. To split, cells were first detached from the flask by trypsinisation using 1:10 trypsin:versene. After 5 minutes incubation at 37°C the cells were dislodged by repeated pipetting of 10ml of medium and the cell suspension distributed among new flasks to which fresh medium was added.

2.7.2 Preparation of Protein from Mammalian cell culture

Protein extracts from cultured mammalian cells was performed using the method of Andrews and Faller (1991).

Cells from a confluent 175 cm^2 flask were scraped into 1.5 ml of ice cold PBS. The suspension was transferred to an eppendorf tube and the cells pelleted by 10 seconds of bench-top centrifugation. The pellet was resuspended in 400ml of ice cold buffer A (10 mM HEPES-KOH pH 7.9 at 4°C, 1.5 mM $MgCl_2$, 10 mM KCl, 0.5 mM DTT, 0.2 mM PMSF) by flicking the tube. The cells were allowed to swell on ice for 10 minutes and then pelleted as above. The pellet was resuspended in 20-100ml (the volume depending on the number of cells) of ice-cold buffer C (20 mM HEPES-KOH pH 7.9, 25% Glycerol, 420 mM NaCl, 1.5 mM $MgCl_2$, 0.2 mM EDTA, 0.5 mM DTT, 0.2 mM PMSF) and placed on ice for 20 minutes. Cell debris was pelleted by bench top centrifugation for 2 min, at 4°C, and the supernatant stored at -70°C as the protein extract. B16 cells were kindly donated by Peter Budd. HeLa cells extracts were purchased from Promega (HeLascribe™). ES and F9 cell extracts were a generous gift from Dr. Richard Meehan.

2.7.3 Harvesting Mouse embryos

Proteins were extracted from whole mouse embryos and from dissected embryo parts. Pregnant mice were provided by the animal facility at the Western General Hospital. Embryos were harvested from Swiss mice at 10 or 11 days *post coitum* (dpc). The mother was killed by cervical dislocation followed by opening the abdominal cavity and removing the two horns of the uterus. These were immersed in ice cold phosphate-buffered saline in order the kill the embryos. All subsequent dissection was performed in ice-cold PBS (phosphate buffered saline). Individual embryos were removed from the decidua by creating a small cut in the wall with a pair of spring-scissors. Slight pressure was then enough to cause the embryo within the amniotic sac to emerge. The extra-embryonic membranes were teased away using two pairs of fine forceps and the embryos processed further.

When sub-regions of the embryos were required for region-specific protein extracts the embryos were dissected under a binocular dissection microscope while immersed in PBS.

2.8 Gel retardation assay

The gel retardation assay involves the use of radiolabelled double-stranded oligonucleotides, protein extracts and native polyacrylamide-gel electrophoresis.

2.8.1 Oligonucleotides

Oligonucleotides were synthesised as described. Oligonucleotides used for this assay are required at a higher level of purity than for other purposes and were therefore gel purified before further use. Purification was performed by electrophoresis through a denaturing polyacrylamide gel. This served the dual purpose of removing impurities, during the process of gel-migration and elution, that may interfere with protein-binding, and size fractionation as a band representing the full-length oligo could be excised. The oligo purification protocol is essentially that of Sambrook *et al.* (1989 - page 11.23). Typically, a 15 % polyacrylamide gel (separating oligonucleotides of 45-65 bp) in 1 x TBE was cast to a thickness of 1mm in a water cooled vertical gel apparatus (LKB). An equal volume of formamide is added to the oligo solution along with Orange-G to 0.2 % and it is heated to 55°C for 5 minutes to disrupt any secondary structure. ~ 2 OD_{260} units of oligo are loaded in each well. In another well, alongside those loaded with oligonucleotide, an equal mixture of formamide and

tracking dyes (0.05 % xylene cyanol, 0.05 % bromophenol blue) is loaded. The migration of the dyes in varying percentages of acrylamide was assessed according to the table on page 11.26 of Sambrook *et al.* (1989). The gel was run at ~200 V until the bromophenol blue band had run three-quarters of the way down the gel. The gel apparatus was then dismantled and the gel sandwiched between two sheets of Saran wrap. A gel of this concentration is robust and readily handled. The oligo bands were then visualised by ultra-violet shadowing. A TLC plate with a fluorescent coating was used (Whatman Silica Gel 60 A). Under ultra-violet light at 254 nm the coating fluoresces green. DNA absorbs light at this wavelength resulting in the oligonucleotides casting a shadow on the plate when it is illuminated through the gel. This enables the careful excision of the oligo band with a clean scalpel blade. The gel slice is placed in a clean eppendorf tube and eluted. Elution was performed by the 'crush and soak' method. The gel slice is crushed against the sides of the tube using the disposable tip from a P1000 Gilson pipette. 500 µl of oligo elution buffer (0.1 % SDS, 0.5 M ammonium acetate, 10 mM magnesium acetate) is added and the tube is incubated on a rotary shaker at 37°C overnight. The eluate was recovered by centrifugation of the gel pulp through a SPINEX column. The oligonucleotides are recovered from this eluate using a 'Mermaid kit' from BIO 101® according to manufacturers instructions.

Concentration of the gel-purified oligonucleotides was determined by measurement of optical density at 260 nm. The extinction coefficient (ε) for each oligonucleotide was calculated, according to Sambrook *et al.* (1989; section 11.21), and the OD_{260} divided by it to give the concentration ($OD=\varepsilon \times C$). Equimolar concentrations of the two complementary oligos to be used in the assay were then mixed, heated to 90°C and cooled slowly to room temperature. The double-stranded oligos were recovered by ethanol precipitation at -20°C overnight.

2.8.2 Binding Reactions

Binding reactions were performed in multiwell plates enabling the simple and rapid processing of several reactions for one or two gels. Each reaction has several components: binding buffer which specifies the salt conditions etc., crucial to the protein-DNA interaction; non-specific competitor DNA to 'mop up' the excess of non-specific DNA-binding proteins and provide a high stringency for the reactions; specific-competitor DNA, either unlabelled probe or an alternative competitor to determine specificity; protein; radiolabelled double-stranded DNA probe. The copolymer poly-d(I-C) was used as a non-specific competitor. Typically 3µg of poly d(I-C) was used

per reaction. Reactions were performed in a total of 20µl. All components except the labelled probe were added, mixed and incubated on ice for 15 minutes. The ^{32}P labelled probe was added finally (typically 10^4-10^5cpm/lane) and mixed followed by a further 15 minute incubation on ice. Loading buffer was then added and the reaction mixture loaded onto the gel immediately.

2.8.3 Native acrylamide gels

Separation of the protein-bound and free oligonucleotide probes was achieved by electrophoresis through non-denaturing polyacrylamide gels. Typically 5% gels were cast by mixing 6.25ml acrylamide (40% stock 19:1 acrylamide:bisacrylamide. Northumbria Biologicals Ltd) with 2.5ml 20 x TBE (final concentration of 0.5x), and 300µl 10% ammonium persulphate in a total volume of 50ml. Just prior to casting 30µl of TEMED was added. A vertical, water cooled apparatus (LKB) was used with gels cast to a 1mm thickness. Binding mixtures were loaded upon addition of 10µl of loading buffer (0.1% bromophenol blue, 10% glycerol). A current of ~15V/cm was applied until the bromophenol blue had migrated close to the bottom of the gel. The gel was dried under vacuum on a heated slab-gel drier and detected by phosphor-imager.

2.9 Southwestern blotting

2.9.1 dsDNA probes from synthetic oligonucleotides

dsDNA probes for use in South-western analysis were produced by concatenation of complementary synthetic oligonucleotides after Sambrook *et al.* (1989; section 12.32). 2 µg of each of the oligonucleotides was phosphorylated at the 5' terminus. This reaction was performed in a kinase/ligase buffer (50mM Tris.Cl pH 7.6, 10mM $MgCl_2$) in which both enzymes will function. The phosphorylated oligonucleotides were then annealed together by heating to 85°C and cooling slowly. The resulting double stranded oligonucleotides were ligated at 16°C overnight. Following ligation the concatenated product was phenol extracted and ethanol precipitated. The pellet was resuspended in 40µl of T.E. (pH 7.6) and 5µl of this (500ng) was used in each labelling reaction. Probe was labelled by nick translation (see above) to a minimum specific activity of 1 x 10^8 cpm/µg.

2.9.2 *SDS - Polyacrylamide Gel electrophoresis*

Proteins were separated according to size by SDS-Polyacrylamide gel electrophoresis (Laemmli, 1970). Denatured (linear) protein is coated with negatively-charged sodium dodecyl sulphate (SDS) thus acquiring a charge approximately proportionate to the length of the polypeptide. In this form the molecule will migrate through a polyacrylamide gel towards an anode with a mobility relative to its size.

Proteins separated prior to South-western blotting were run on gels of varying concentrations dependent upon the degree of separation required for proteins of a particular molecular weight. It was determined that a gel of 7% acrylamide gave good separation in the 30-200 kD range when run at 6 volts/cm for approximately 16 hours.

A 7% separation gel was used with a 4% stacking gel. Stock acrylamide (19:1 acrylamide:bisacrylamide) was mixed with the relevant buffer (Stacking:- 0.5 M Tris pH 6.8, 8 mM EDTA, 0.4% (w/v) SDS; Separating:- 1.5 M Tris pH 8.8, 8 mM EDTA, 0.4% (w/v) SDS) and polymerised by addition of approximately 0.01 volumes of 10% ammonium persulphate and 0.001 volumes TEMED. The gel was cast to a thickness of 1mm in a water cooled vertical gel apparatus (LKB). Samples to be separated were mixed with 10 µl (or an equal volume if the sample was larger than 10 µl) of SDS sample buffer. SDS sample buffer was made by mixing 900 µl stock dissociation buffer (10% stacking buffer, 2% SDS, 10% Glycerol) with 50 µl β-Mercaptoethanol and 50 µl 0.1% Bromophenol Blue (in stacking buffer). The mix of sample and sample buffer was incubated at 100°C for 2 minutes and loaded onto the stacking gel. An electrode buffer of 24.8mM Tris pH 8.3, 192mM Glycine, 2mM EDTA and 0.1% SDS was used and the gel run at 65 volts (6 volts/cm) for 16 hours.

2.9.3 *Electrotransfer of Proteins*

After SDS-PAGE separation, the proteins must be transferred to a nitro-cellulose membrane prior to South-western blotting. Transfer was performed in a vertical electrotransfer apparatus (LKB).

The gel was disassembled and the stacking gel cut away. The gel was soaked in transfer buffer (50mM Tris, 380mM Glycine, 0.1% (w/v) SDS, 20% (v/v) Methanol) for 1 hour. Nitro-cellulose membrane (0.45 µm) was cut to size, along with two pieces of Whatmann 3MM paper, and soaked in transfer buffer for 5 minutes. The membrane was placed on the gel, with care taken to avoid any bubbles between, and the two were placed between the 3MM paper in the transfer apparatus. The set-up was oriented with

the membrane adjacent to the anode and transfer was performed at 50mA for 3 hours in buffer at 4°C.

2.9.4 Filter processing and Hybridisation

The method used was based upon that of Dikstein *et al.* (1992) which is itself an adaptation of the method used in Singh *et al.* (1988).

Proteins were separated by SDS polyacrylamide-gel electrophoresis as described. The proteins were then electro-transferred to a nitro-cellulose support as described. Following transfer the nitro-cellulose filter was first incubated for 1 hour at room temperature (on an orbital shaker) in a blocking solution containing non-fat dried milk (Marvel™) as a blocking agent (5% non-fat milk, 50mM Tris pH 7.5, 100mM NaCl, 1mM EDTA, 1mM DTT). The filter was then washed twice in TNE-100 (10mM Tris pH7.5, 100mM NaCl, 1mM EDTA, 1mM DTT). This was followed by 1 hour at room temperature in a protein-denaturing solution (7M guanidine HCl, 50mM Tris pH 8.0, 50mM DTT, 2mM EDTA, 0.25% non-fat milk). Finally the proteins were slowly renatured by incubation at 4°C overnight in a renaturing solution (50mM Tris pH 7.5, 100mM NaCl, 2mM DTT, 2mM EDTA, 0.1% NP-40, 0.25% non-fat milk).

Filters treated as above were incubated for 1-2 hours at 4°C in binding buffer (0.25mM DTT, 2mM, 10mM HEPES pH7.9, 100mM KCl, 100mM NaCl, 4mM spermidine, 2.5% glycerol) plus 0.1% Triton X-100, in the presence of 10 μg/ml of non-specific competitor DNA (poly dI-C). The radiolabelled probe was added (5 x 10^6 cpm/ml) and the incubation continued for a further 30 minutes. Filters were then given three 5 minute washes in binding buffer plus 0.05% Triton X-100. The nitro-cellulose was removed from the washing solution and allowed to dry on a piece of Whatman 3MM paper. The filters were wrapped in Saran wrap and autoradiographed.

Chapter 3

Comparative Sequence analysis

3 Introduction

The mechanisms regulating gene expression comprise two components, *cis*-acting elements and *trans*-acting factors (section 1.5). In the absence of prior information, from genetic studies for example, the *cis* elements are more directly accessible to the molecular biologist. *Cis*-regulatory elements are features of the DNA sequence near to the gene (on a genomic scale) making it possible to clone them as adjacent DNA from the locus and to determine their location by a variety of means. Whatever method one chooses to identify *cis* elements, at some stage in the work it will be necessary to determine the nucleotide sequence of the DNA. One method used to define the position of *cis* regulatory elements is to apply a functional assay, such as reporter gene fusion, to sub-sections of the DNA with the aim of defining the minimum required stretch of DNA. Sequence of this region can then be determined. Alternatively the sequence itself can be exploited to identify which regions of it are functionally significant. This type of approach is taken in this study.

As the number of model organisms used in the study of development grows, genes important to the processes of embryogenesis are isolated from a diverse range of animals with varying phylogenetic relationships to one another. It is becoming increasingly clear that many of the genes regulating development have been highly conserved during evolution, few more so than the homeobox genes (sections 1.1.3 & 1.2.1). As transcription factors, the products of the homeobox genes interact with *cis* regulatory elements flanking genes downstream in the regulatory cascade. Given the evolutionary conservation of DNA binding domains, such as the homeodomains, it is not unreasonable to expect that the *cis* sequences with which they interact will also have been tightly conserved. On the basis of such functional constraint upon evolutionary change, *cis* elements may be identified by comparing the flanking, non-coding DNA of cognate genes from related species. Functional sequences may be expected to stand out due to their higher level of homology.

In addition to homology studies there are increasing bodies of information regarding the *cis* control regions of many genes, and any putative regulatory region (flanking sequence) can be assessed for the presence of known elements or characteristic features associated with regulation of gene expression.

This chapter describes the use of comparative sequence analysis to search for conserved sequences possibly functioning as *cis* regulatory elements of *Msx*-1. It also describes a detailed examination of other characteristics of *Msx*-1 flanking DNA in reference to potential regulatory function.

3.1 Sequencing 5'-flanking DNA of the mouse *Msx*-1 gene

As a first step toward analysing the 5'-flanking region of the mouse *Msx*-1 gene a 1.7 kb genomic fragment of this region was subcloned into a suitable vector and sequenced.

3.1.1 Subcloning Msx-*1 5' fragment*

The 5'-flanking DNA to be sequenced was a sub-fragment of a cosmid isolated by Robert Hill. The subclone provided by Dr. Hill was a plasmid, pCosλ2-22/Sma#4, carrying a 1.7 kb EcoRI-SmaI insert in the vector pTZ18. This vector was unsuitable for the required purposes so the insert was cloned into an alternative vector. The insert was excised by digestion with EcoRI and XbaI (thereby leaving a small portion of pTZ at the SmaI end of the insert) and gel purified. It was ligated into the polylinker of pGEM-7Zf(-) (Promega) which was digested with the same two enzymes. The resulting plasmid was termed pDEL3 as it allowed for deletions to be made into the 3' end of the insert (figure 3.1). The pGEM-7 vector was chosen as it enabled production of single-stranded DNA, using the f1 origin, and possessed convenient restriction sites to allow unidirectional deletion with exoIII (section 2.3.3).

Figure 3.1

Figure 3.1 Map of clone pDEL3. Subclone derived from a *Msx*-1 cosmid into pGEM-7Zf, a vector suitable for the generation of unidirectional deletions. Restriction sites shown: Sp=SphI; X=XbaI; Sm=SmaI; E=EcoRI; C=ClaI

3.1.2 Production of a deletion library of pDEL3

To facilitate sequencing, a series of nested deletions was created spanning the length of the pDEL3 insert. Exonuclease III (exoIII) can create unidirectional deletions with use of specific restriction sites (see section 2.3.3). Deletions were made into the SmaI end of the EcoRI-SmaI *Msx*-1 fragment in pDEL3 (the 3' end with respect to transcription of *Msx*-1; see Fig. 3.1). The 3'-overhanging, exonuclease-resistant end was generated adjacent to the vector DNA by digestion with SphI at a site in the polylinker. The 5'-overhanging, exonuclease-digestible end was generated adjacent to the insert DNA by digestion with XbaI at the site into which the insert was cloned from pTZ. Deletions were generated as described with seven time points selected over a range which spans the whole insert. The time points have an approximate interval of 300 bp as seen in figure 3.2. This interval allows for sequence information from one time point to overlap with that from another. There is heterogeneity in the extent of deletion within the population of DNA molecules in a given time point and this may lead to individual clones not having a deletion representative of their time point. This problem is overcome by analysing individual clones on an agarose gel and sequencing a representative selection.

3.1.3 Sequencing of pDEL3

Single-stranded DNA was prepared as described (section 2.2.3) and sequencing was performed using the Sequenase protocol for single-stranded DNA (section 2.6.1). Maximum information was collected from each reaction by using short (~2 hours) and long (~4 hours) gel-runs. Following assembly of the sequence, synthetic oligonucleotides were synthesised, complementary to the sequence at intervals of 200-300 bp, and used to sequence double-stranded templates. The information from the complementary strand served to confirm the data produced from the deletion series and negated the need for a second set of deletions in the opposite direction. The STADEN (Amersham Staden-Plus; Release 1, Version 6) contig assembly computer program was used to match overlapping sequences and assemble them into longer stretches of contiguous sequence. Once the individual sequencing files had been assembled into a small number of large contigs, the sum length of which approximated the size of the insert, synthetic oligonucleotides were made to enable sequencing from the ends of these contigs across the remaining gaps.

The sequence information generated from the pDEL3 insert is shown in figure 3.3: in total 1683 base pairs were sequenced.

Figure 3.2

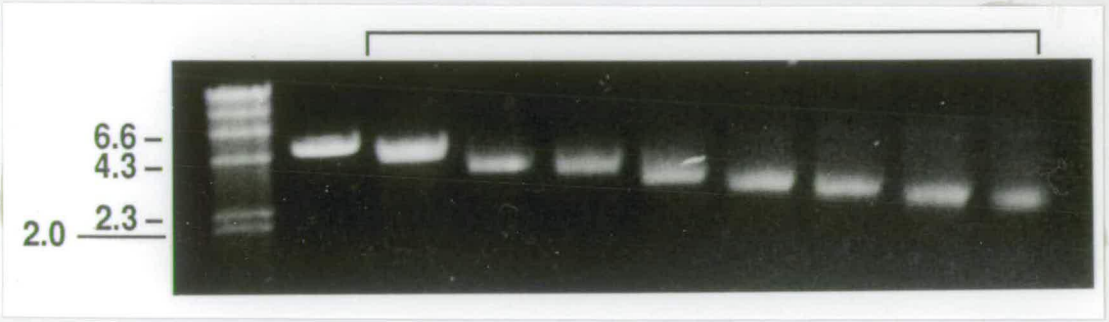
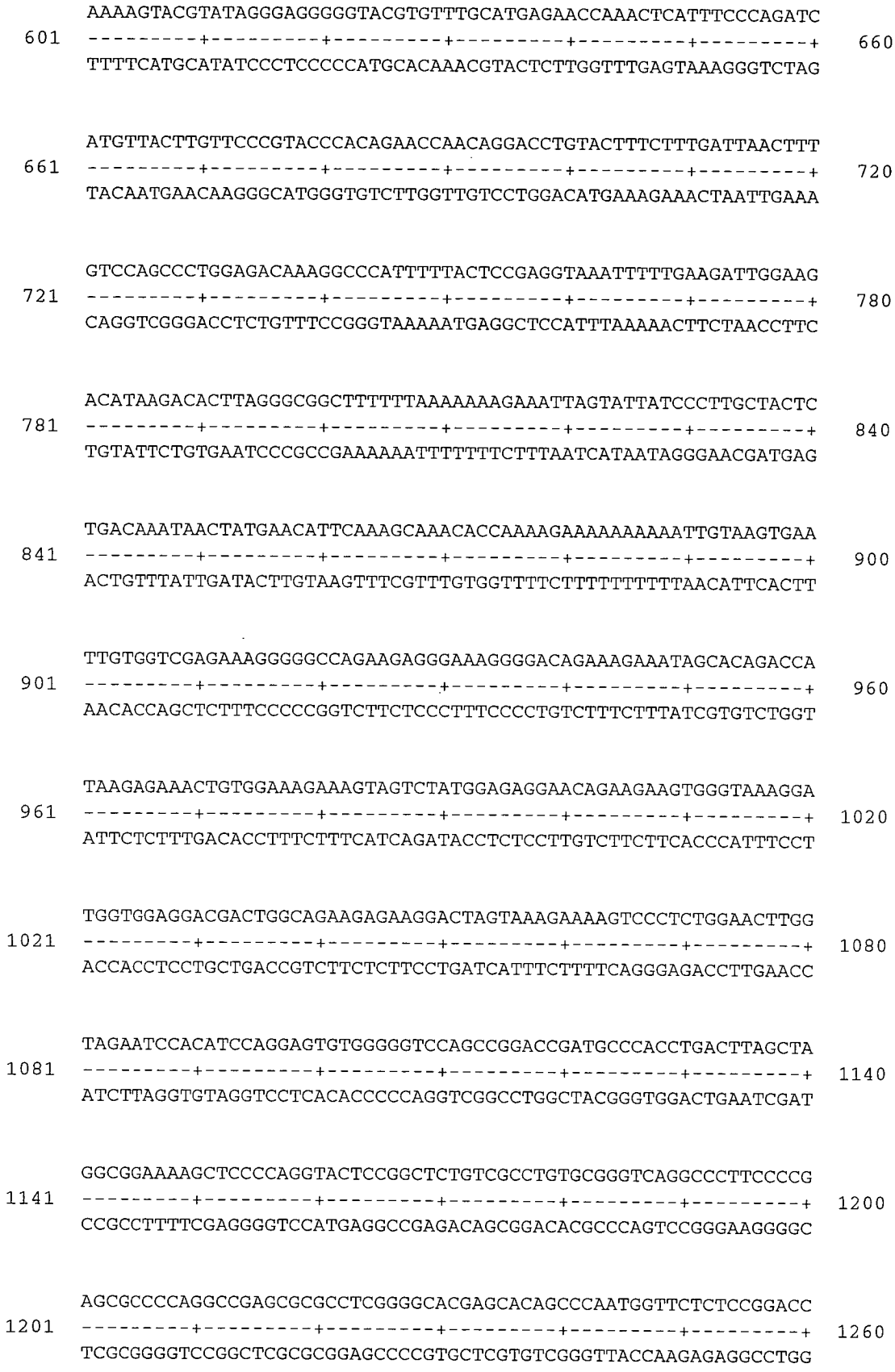


Figure 3.2 Ethidium bromide stained agarose gel of the products from the unidirectional ExoIII deletions of pDEL3. Molecular weight markers demonstrate a total range of deletion of ~2.5 kb. Lanes 2-8 (bracketed) were selected for sequencing.

Figure 3.3


GAATTCGCTATCCCTCCGTGCTCGGATAATTACCCCTAAATGGCCACGGCAGCCCCTTGT
1 Eco RI---+---------+---------+---------+---------+---------+ 60
CTTAAGCGATAGGGAGGCACGAGCCTATTAATGGGGATTTACCGGTGCCGTCGGGGAACA
GTTTCCTGGAGATTAGAGCCCGGCAGTCATCAATGCACAGTCCCGGTGAGCCGCCAATCA
61 ---------+---------+---------+---------+---------+---------+ 120
CAAAGGACCTCTAATCTCGGGCCGTCAGTAGTTACGTGTCAGGGCCACTCGGCGGTTAGT
CCTCCTCCACCTCCCCCGGAGCCGCGAGCCTGGGCCCTGGGATAAGAGGCCATATAAAGA
121 ---------+---------+---------+---------+---------+---------+ 180
GGAGGAGGTGGAGGGGGCCTCGGCGCTCGGACCCGGGACCCTATTCTCCGGTATATTTCT
AAAGCCCCCCCCCCCCCAACTCCCCAACAGCCTGTTCGAACCCAGTTTACAAGGTCTTCT
181 ---------+---------+---------+---------+---------+---------+ 240
TTTCGGGGGGGGGGGGGTTGAGGGGTTGTCGGACAAGCTTGGGTCAAATGTTCCAGAAGA
GTGGGTCCCCACTCCAACTTTCCTTCTTTTTGTATCAGTCAAACAAAACAAAACAAAACA
241 ---------+---------+---------+---------+---------+---------+ 300
CACCCAGGGGTGAGGTTGAAAGGAAGAAAAACATAGTCAGTTTGTTTTGTTTTGTTTTGT
AAACAAAACTCGTTTCCTTCGAAATCGGGCAAACAAAACTTCTTAAGCCTTGCACCAAAG
301 ---------+---------+---------+---------+---------+---------+ 360
TTTGTTTTGAGCAAAGGAAGCTTTAGCCCGTTTGTTTTGAAGAATTCGGAACGTGGTTTC
ATATACCTCCAGGAACTGCTTATTTTTCCTACTGGCCCATCAAATGGGCCTCCACTCTGC
361 ---------+---------+---------+---------+---------+---------+ 420
TATATGGAGGTCCTTGACGAATAAAAAGGATGACCGGGTAGTTTACCCGGAGGTGAGACG
ACCAGCAGACAGACCCACAGTTAGTGAAGTGTGATTGTTTTTGAGGATTCATTGTCTGAG
421 ---------+---------+---------+---------+---------+---------+ 480
TGGTCGTCTGTCTGGGTGTCAATCACTTCACACTAACAAAAACTCCTAAGTAACAGACTC
GTCATGTTGAGGCCAGATTGTTGGTGGGTTCTGTTGGTCTTATTTCGGTCTTCACCCAAG
481 ---------+---------+---------+---------+---------+---------+ 540
CAGTACAACTCCGGTCTAACAACCACCCAAGACAACCAGAATAAAGCCAGAAGTGGGTTC
GCATCCAGTGGCTTCTTTGGCTAGTAGCAGTTCTCTGGTAGAGGTGTAACTTTCTGGAGA
541 ---------+---------+---------+---------+---------+---------+ 600
CGTAGGTCACCGAAGAAACCGATCATCGTCAAGAGACCATCTCCACATTGAAAGACCTCT


AAAAGTACGTATAGGGAGGGGGTACGTGTTTGCATGAGAACCAAACTCATTTCCCAGATC
601 660
TTTTCATGCATATCCCTCCCCCATGCACAAACGTACTCTTGGTTTGAGTAAAGGGTCTAG
ATGTTACTTGTTCCCGTACCCACAGAACCAACAGGACCTGTACTTTCTTTGATTAACTTT
661 720
TACAATGAACAAGGGCATGGGTGTCTTGGTTGTCCTGGACATGAAAGAAACTAATTGAAA
GTCCAGCCCTGGAGACAAAGGCCCATTTTTACTCCGAGGTAAATTTTTGAAGATTGGAAG
721 780
CAGGTCGGGACCTCTGTTTCCGGGTAAAAATGAGGCTCCATTTAAAAACTTCTAACCTTC
ACATAAGACACTTAGGGCGGCTTTTTTAAAAAAAGAAATTAGTATTATCCCTTGCTACTC
781 840
TGTATTCTGTGAATCCCGCCGAAAAAATTTTTTTCTTTAATCATAATAGGGAACGATGAG
TGACAAATAACTATGAACATTCAAAGCAAACACCAAAAGAAAAAAAAAATTGTAAGTGAA
841 900
ACTGTTTATTGATACTTGTAAGTTTCGTTTGTGGTTTTCTTTTTTTTTTAACATTCACTT
TTGTGGTCGAGAAAGGGGGCCAGAAGAGGGAAAGGGGACAGAAAGAAATAGCACAGACCA
901 960
AACACCAGCTCTTTCCCCCGGTCTTCTCCCTTTCCCCTGTCTTTCTTTATCGTGTCTGGT
TAAGAGAAACTGTGGAAAGAAAGTAGTCTATGGAGAGGAACAGAAGAAGTGGGTAAAGGA
961 1020
ATTCTCTTTGACACCTTTCTTTCATCAGATACCTCTCCTTGTCTTCTTCACCCATTTCCT
TGGTGGAGGACGACTGGCAGAAGAGAAGGACTAGTAAAGAAAAGTCCCTCTGGAACTTGG
1021 1080
ACCACCTCCTGCTGACCGTCTTCTCTTCCTGATCATTTCTTTTCAGGGAGACCTTGAACC
TAGAATCCACATCCAGGAGTGTGGGGGTCCAGCCGGACCGATGCCCACCTGACTTAGCTA
1081 1140
ATCTTAGGTGTAGGTCCTCACACCCCCAGGTCGGCCTGGCTACGGGTGGACTGAATCGAT
GGCGGAAAAGCTCCCCAGGTACTCCGGCTCTGTCGCCTGTGCGGGTCAGGCCCTTCCCCG
1141 1200
CCGCCTTTTCGAGGGGTCCATGAGGCCGAGACAGCGGACACGCCCAGTCCGGGAAGGGGC
AGCGCCCCAGGCCGAGCGCGCCTCGGGGCACGAGCACAGCCCAATGGTTCTCTCCGGACC
1201 1260
TCGCGGGGTCCGGCTCGCGCGGAGCCCCGTGCTCGTGTCGGGTTACCAAGAGAGGCCTGG

```
      CGCCCCCTCGCGCTCTGATTGGCCGCTGCCACGCTGGCCTTGCCTTATTAACAAGTTCTC
1261  ---------+---------+---------+---------+---------+---------+  1320
      GCGGGGGAGCGCGAGACTAACCGGCGACGGTGCGACCGGAACGGAATAATTGTTCAAGAG

      AGGGGAGCGGCGGCGGACCCGGAGCCGGCGAGTGCGCCTGGGAACTCGGCCTGAGCGGCG
1321  ---------+---------+---------+---------+---------+---------+  1380
      TCCCCTCGCCGCCGCCTGGGCCTCGGCCGCTCACGCGGACCCTTGAGCCGGACTCGCCGC

      CAGGGATCCAGGCCCCGCTCGCTCGAGTTGGCCTTCTGGGGAAGCCGCAGGAGGCTCGCG
1381  ---------+---------+---------+---------+---------+---------+  1440
      GTCCCTAGGTCCGGGGCGAGCGAGCTCAACCGGAAGACCCCTTCGGCGTCCTCCGAGCGC

      CGCGAGAGCCGGCCGGGCCAGGAACCCAGGAGCTCGCAGAAGCCGGTCAGGAGCTCGCAG
1441  ---------+---------+---------+---------+---------+---------+  1500
      GCGCTCTCGGCCGGCCCGGTCCTTGGGTCCTCGAGCGTCTTCGGCCAGTCCTCGAGCGTC

      AAGCCGGTCGCGCTCCCAGCCTGCCCGAAACCCATGATCCAGGGCTGTCTCGAGCTGCGG
1501  ---------+---------+---------+---------+---------+---------+  1560
      TTCGGCCAGCGCGAGGGTCGGACGGGCTTTGGGTACTAGGTCCCGACAGAGCTCGACGCC

      CTGGAGGGGGGGTCCGGCTCTGCATGGCCCCGGCTGCTGCTATGACTTCTTTGCCACTCG
1561  ---------+---------+---------+---------+---------+---------+  1620
      GACCTCCCCCCCAGGCCGAGACGTACCGGGGCCGACGACGATACTGAAGAAACGGTGAGC

      GTGTCAAAGTGGAGGATCCGCCTTCGCCAAGCCTGCTGGGGGAGGCGTTGCCAAGCCCCC
1621  ---------+---------+---------+---------+---------+--------Sm  1680
      CACAGTTTCACCTCCTAGGCGGAAGCGGTTCGGACGACCCCCTCCGCAACGGTTCGGGGG

      GGG
1681  a I  1683
      CCC
```

Figure 3.3: Sequence of the insert in clone pDEL3 corresponding to the 5'-flanking region of mouse *Msx*-1. The restriction sites used in the cloning of the insert, EcoRI and SmaI, are marked at the ends of the sequence (section 3.1.1). The C underlined at position 1283 marks the transcriptional start site published by Kuzuoka *et al.* (1994).

3.2 Sequencing the 5'-flanking region of human MSX1

To enable a comparative study between the upstream regions of the human and murine genes it was necessary to obtain a clone for the human cognate of mouse *Msx*-1 (not cloned at this time). A cosmid library was screened to isolate the human gene and various fragments subcloned for sequencing.

3.2.1 Isolation of a human MSX1 *cosmid*

A cosmid library derived from a normal human male lymphoblastoid cell-line, constructed in the Lawrist-4 vector, was kindly donated by Dr Wendy Bickmore. The library was constructed by the cloning of Sau3A partially digested DNA into the BamHI site of Lawrist-4. Four amplified pools of clones were provided and ~2.5 x 10^5 colonies from each pool were screened (section 2.5.2) using the 215 bp SacI-NcoI fragment from the 5' region of the *Msx*-1 cDNA as a probe (figure 3.4). This probe fragment was excised from plasmid pHox-7XS, a cDNA clone containing the full coding region inserted into the pTZ18R vector (R. Hill). The DNA probe was labelled by the random priming method (see section 2.4.2). A primary screen of the four pools of clones comprising the library (amplified) identified several positive clones in one particular pool. Six colonies corresponding to the signals on the autoradiographs were picked and taken through secondary and tertiary screens. The result of one such secondary screen can be seen in figure 3.5A which shows the autoradiograph produced from a high concentration plating of one particular primary positive suspension. One colony from this filter was picked and taken to a tertiary screen, performed as for the secondaries. The tertiary screen was performed at a range of lower concentrations to ensure that an individual colony could be selected. Figure 3.5B shows a tertiary screen filter: the positives on this filter represent the vast majority of colonies on the plate. The cosmid isolated from the individual tertiary colony picked was termed pCosHumH7.

3.2.2 Subcloning from MSX1 *cosmid*

Subclones were derived from pCosHumH7 to enable sequencing of human MSX1 5'-flanking DNA. A fragment from the correct region, and of suitable size, was identified and cloned. The cosmid was digested with a panel of restriction enzymes and the fragments separated on an agarose gel (figure 3.6A). The gel was Southern blotted

Figure 3.4

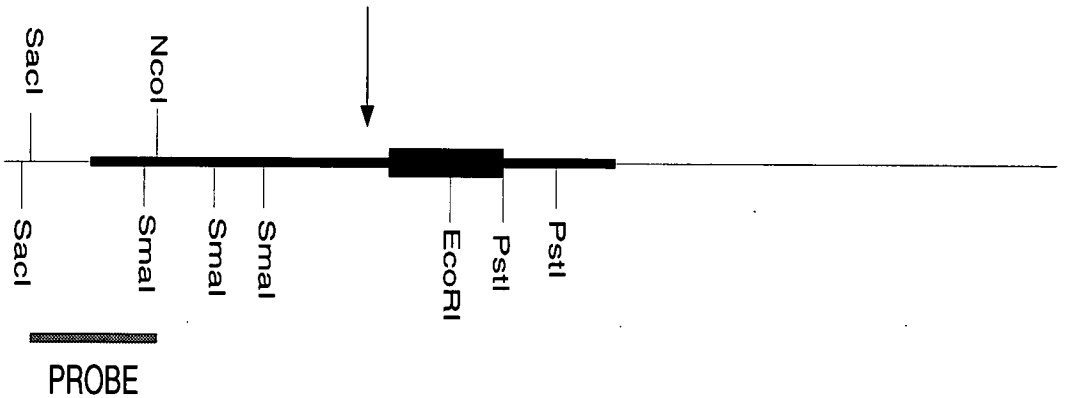


Figure 3.4 Map of the mouse *Msx*-1 cDNA showing key restriction sites, the probe used in cloning the human cognate and the position of the intron (arrow). The thickened line represents the open reading frame with the homeobox shown as a yet thicker region.

Figure 3.5

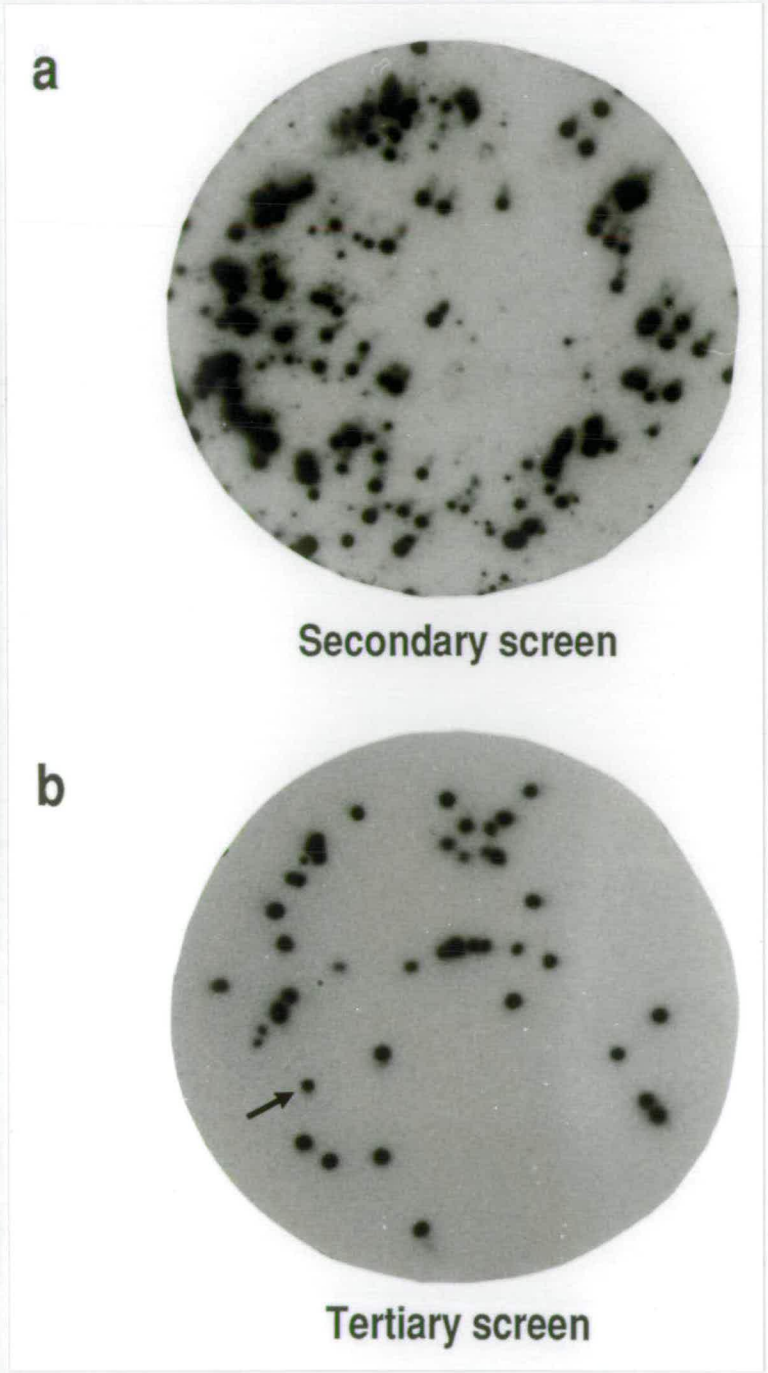


Figure 3.5 Autoradiographs showing cosmid library screen for human *Msx*-1 clone. a) secondary screen: high density of colonies with considerable proportion of positives indicating accurate picking on primary plate. b) tertiary screen showing lower density plating: majority of colonies are positive and arrow indicates the colony picked as pCosHumH7.

Figure 3.6

A)

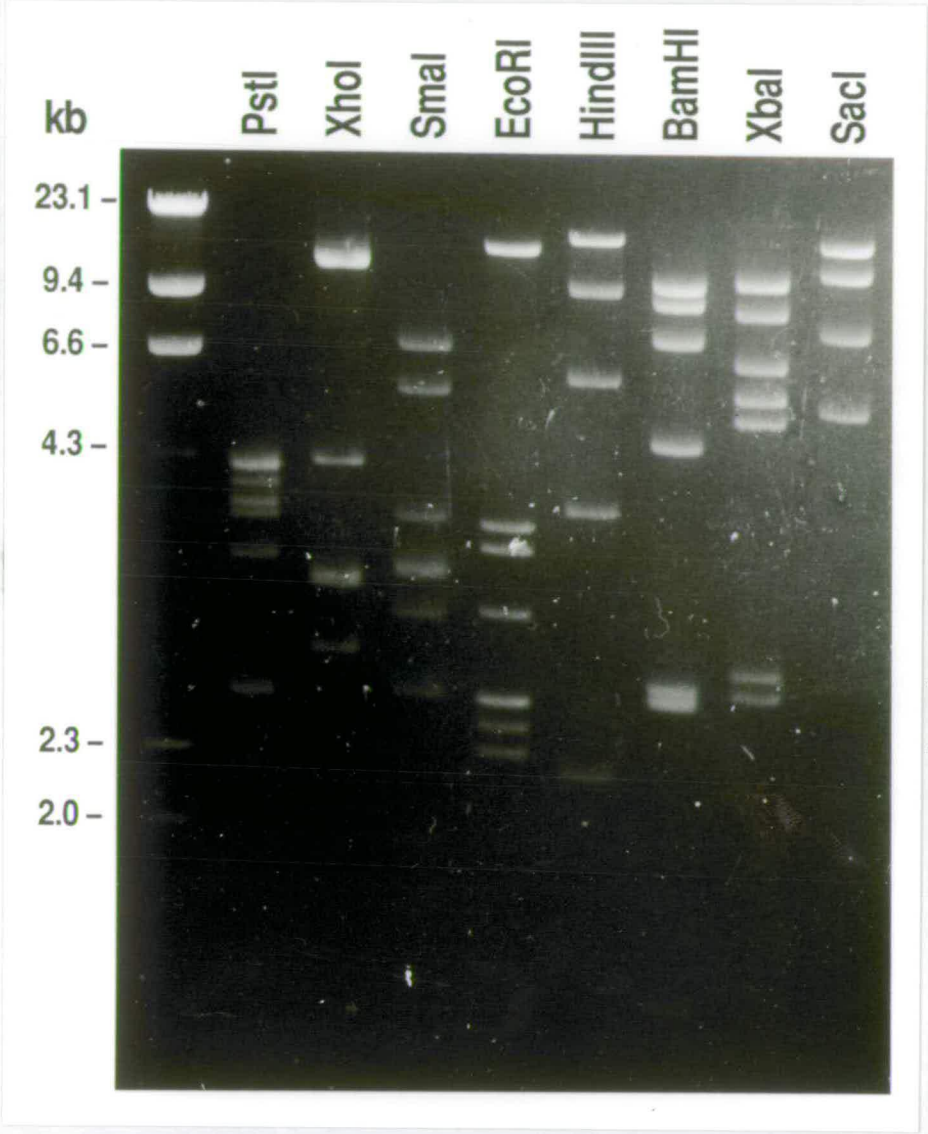


Figure 3.6 A) Ethidium bromide-stained agarose gel of single restriction digests performed on pCosHumH7. The bracket on the right highlights the three EcoRI bands at 2.3-2.5 kb (see text and figure 3.6 B)

Figure 3.6

B)

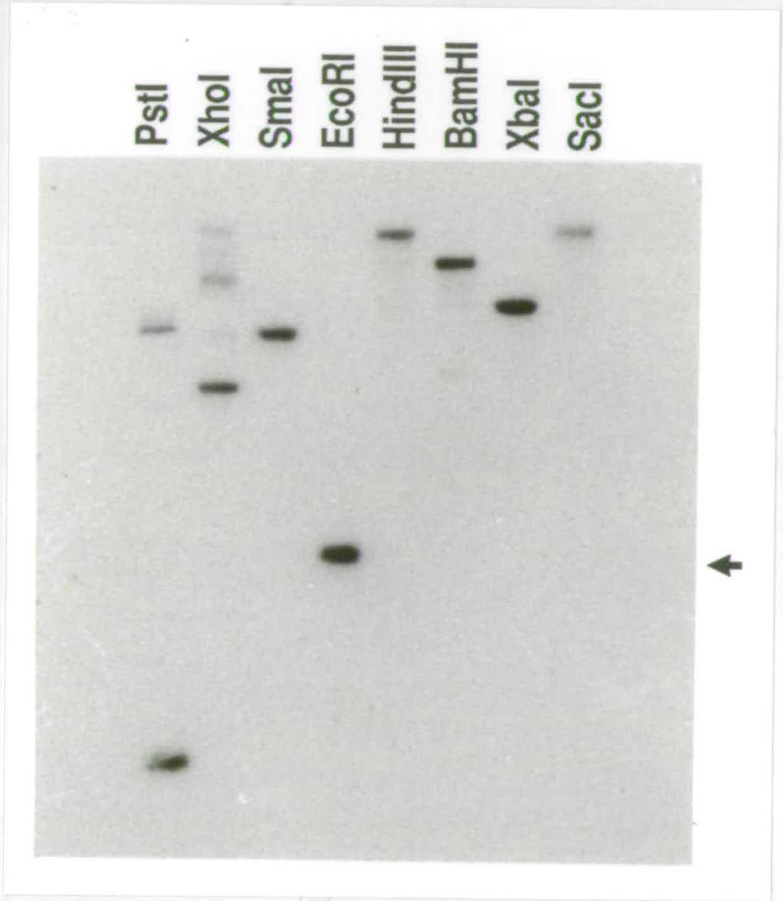


Figure 3.6 B) Autoradiograph of Southern blot from gel in figure 3.6 a). Blot was probed with SacI-NcoI fragment from 5' region of mouse *Msx*-1 cDNA. Arrow on right indicates 2.5 kb EcoRI band that was cloned to form pB1B (region bracketed in figure 3.6 a); see text)

onto Hybond and the filter hybridised to the mouse SacI-NcoI fragment (figure 3.4), used to isolate the cosmid. Hybridisation to the Southern was performed under the conditions described for the cosmid screen (section 2.5.3). Figure 3.6B shows the autoradiograph obtained from this Southern blot. Lane 4 shows that a single EcoRI fragment of approximately 2.5 kb hybridises to the probe (arrowed). To further assess the suitability of this and other fragments, a second Southern blot was performed following separation of pCosHumH7 digested with various combinations of enzymes (figure 3.7). Figure 3.7 shows a clear reduction in size of approximately 1 kb between this EcoRI fragment (lane 3) and the EcoRI/NcoI fragment hybridising in lane 7, demonstrating that at large part of the EcoRI fragment lies upstream of the NcoI and constitutes 5'-flanking DNA. I decided on the basis of this to subclone this EcoRI fragment.

The ethidium bromide stained agarose gel revealed three closely spaced bands at the mobility indicated by the Southern hybridisation signal (Figure 3.6A). It was impossible to determine which of these three bands was the target of the hybridising probe so it was decided to shotgun clone the EcoRI digested cosmid and screen the resulting clones with the mouse SacI-NcoI fragment. An excess of EcoRI digested pCosHumH7 was ligated to EcoRI digested pGEM-7Zf and the products transformed into XL1-Blue cells. Transformants were screened with the mouse SacI-NcoI fragment. One positive was picked and termed pB1B. In order to avoid sequencing the whole pB1B insert a more precisely located sub-clone was generated. The smallest band hybridising to the probe was a PstI/NcoI fragment of approximately 1.1 kb (arrow fig. 3.7B). This fragment was clearly visible on the ethidium stained gel of digested pB1B (fig 3.8, small arrow). Southern analysis confirmed that it was a sub-fragment of the pB1B insert (not shown). This fragment was cloned into PstI-NcoI digested pGEM-5Zf. The resultant plasmid was termed pE31.1. Figure 3.9 shows the relationship between the human MSX1 locus and the subclones derived.

3.2.3 Reported cloning of Human MSX1

At this stage in the work a paper was published reporting the cloning of the human MSX1 gene. Hewitt *et al.* (1991) published sequence from a genomic clone that they too isolated from a cosmid library. Comparison of the SacI-NcoI mouse *Msx-1* fragment used as probe revealed 86% homology with the human. They identified the 5' region as within the same 2.5 kb EcoRI fragment subcloned to form pB1B. An EcoRI site was present in the intron, 204 bp in from the 5' end, indicating that the 2.5 kb EcoRI insert of pB1B carried approximately 1550 bp of sequence upstream of the

Figure 3.7

A)

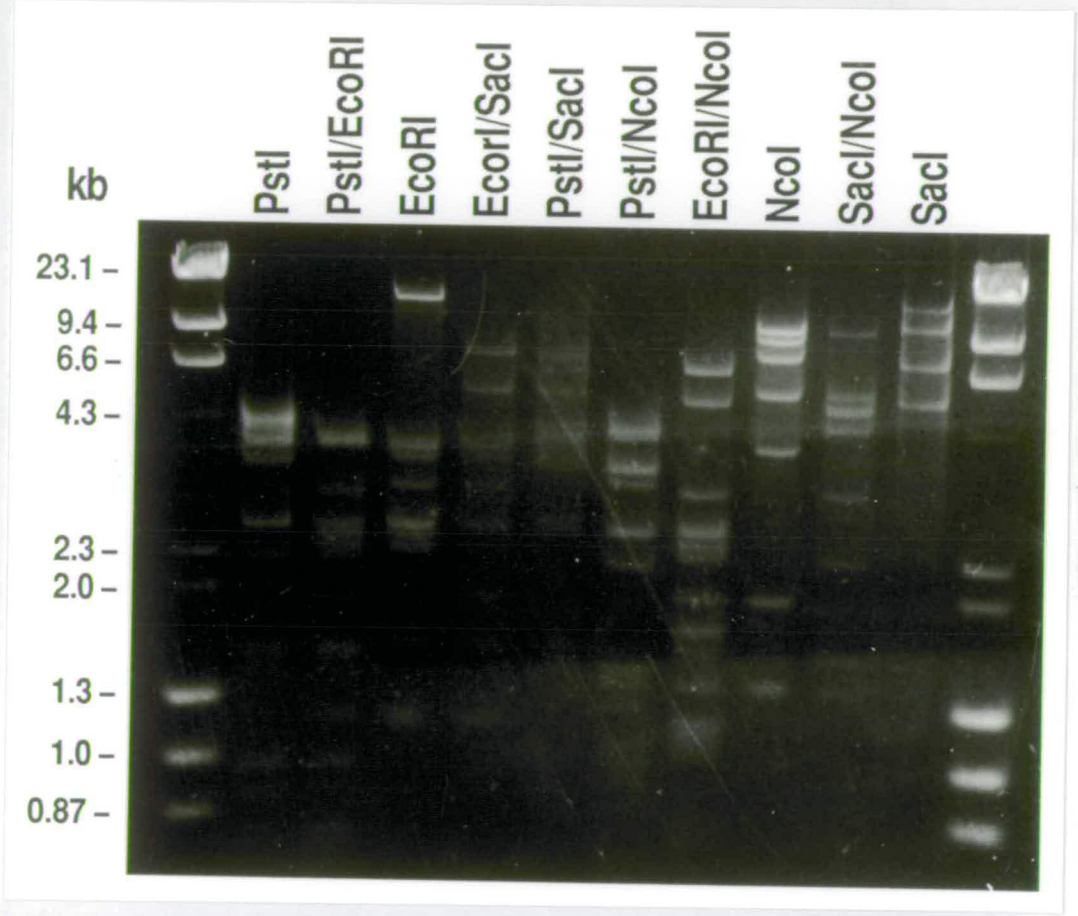


Figure 3.7 A) Ethidium bromide stained agarose gel of various single and double restriction digests on pCosHumH7. Molecular weight markers are seen in the left hand column.

Figure 3.7

B)

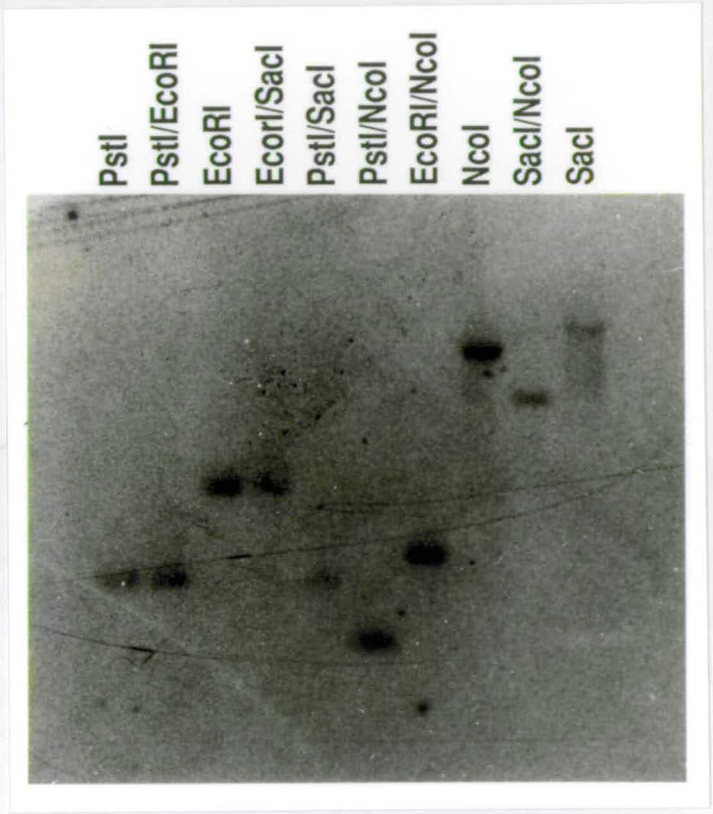


Figure 3.7 B) Autoradiograph of Southern blot from gel in figure 3.7 A) .Blot was probed with SacI-NcoI fragment from 5' region of mouse *Msx*-1 cDNA. Open arrow indicates 2.5 kb EcoRI fragment corresponding to insert of pB1B. Closed arrow indicates ~1.1 kb PstI-NcoI fragment cloned as pE31.1 (see text).

Figure 3.8

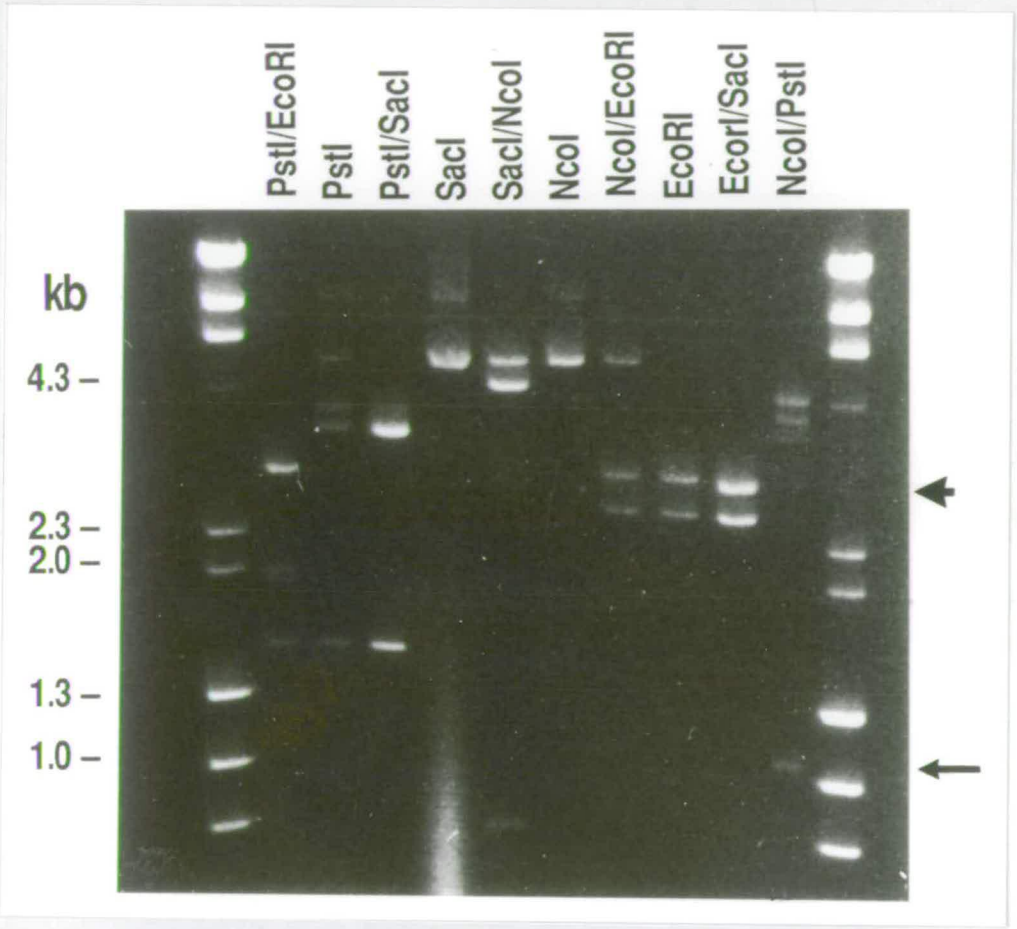


Figure 3.8 Ethidium bromide stained agarose gel of multiple digests on the cosmid subclone pB1B. The large arrowhead indicates the 2.5 kb EcoRI insert fragment, the small arrow indicates that the 1.1 kb PstI-NcoI fragment (cloned as pE31.1) is a subfragment of this insert.

Figure 3.9

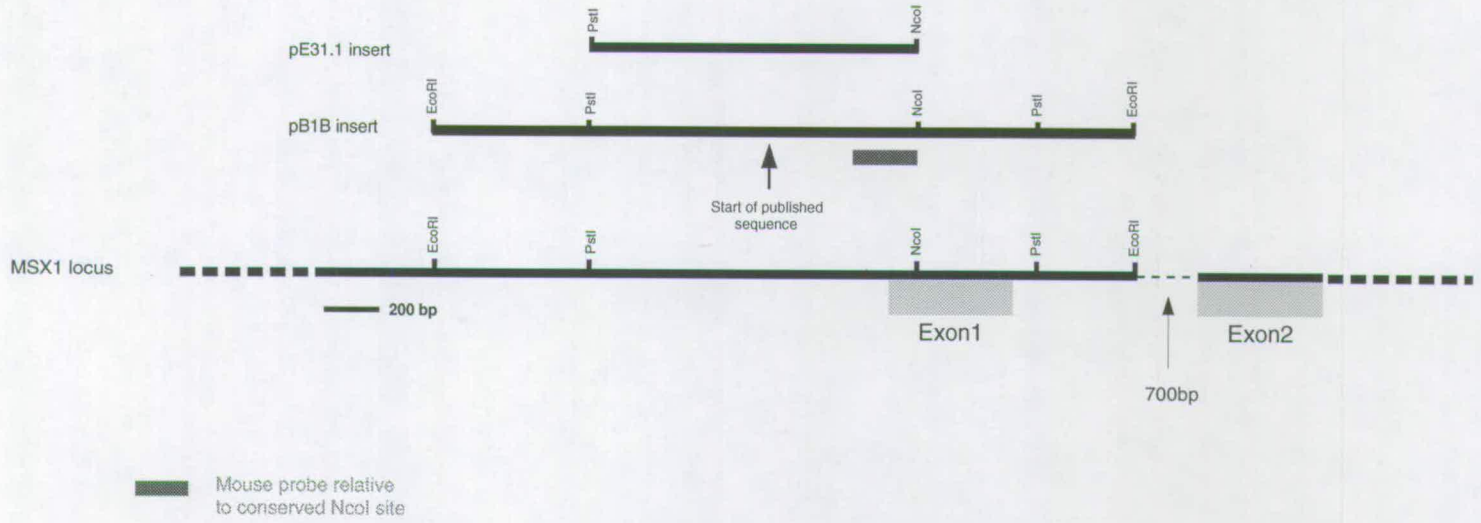


Figure 3.9 Map of the human MSX1 locus showing the relationship of the subclones derived to features of the gene structure. At the top is the insert of the smaller subclone pE31.1 and below it the insert in the pB1B subclone. The two exons (light shading) are separated by a combination of known sequence and ~700 bp of unsequenced DNA (Hewitt *et al*, 1991). The mouse probe used to isolate the human cosmid is shown (dark shading) alongside the homologous human region (86% conserved in human).

translational start codon. Hewitt *et al.* reported sequence for 420 bp upstream of the start codon leaving approximately 1100 bp of unknown sequence at the 5' end of the pB1B insert. The 5' end of the published sequence was used to design a complimentary oligonucleotide from which to sequence further upstream.

3.2.4 PCR amplification from pCosHumH7 subclones

A synthetic oligonucleotide was made complementary to bases 69-52 of the MSX1 sequence published by Hewitt *et al.* (5'-ACTCGCCCGGAGCGCTGG-3'; oligo B962). This sequence is located approximately half way into both the pB1B insert and the pE31.1 insert. Oligo B962 was used in conjunction with the pUC/M13 reverse sequencing primer (5'-TCACACAGGAAACAGCTATGAC-3') to PCR amplify a region of approximately 800 bp from pE31.1 corresponding to the full extent of the unknown sequence of this clone. An annealing temperature of 50°C was used in a protocol described (section 2.3.4). Generation of this fragment and the use of cycle sequencing facilitated linking of the published sequence and the first bases beyond it in pE31.1.

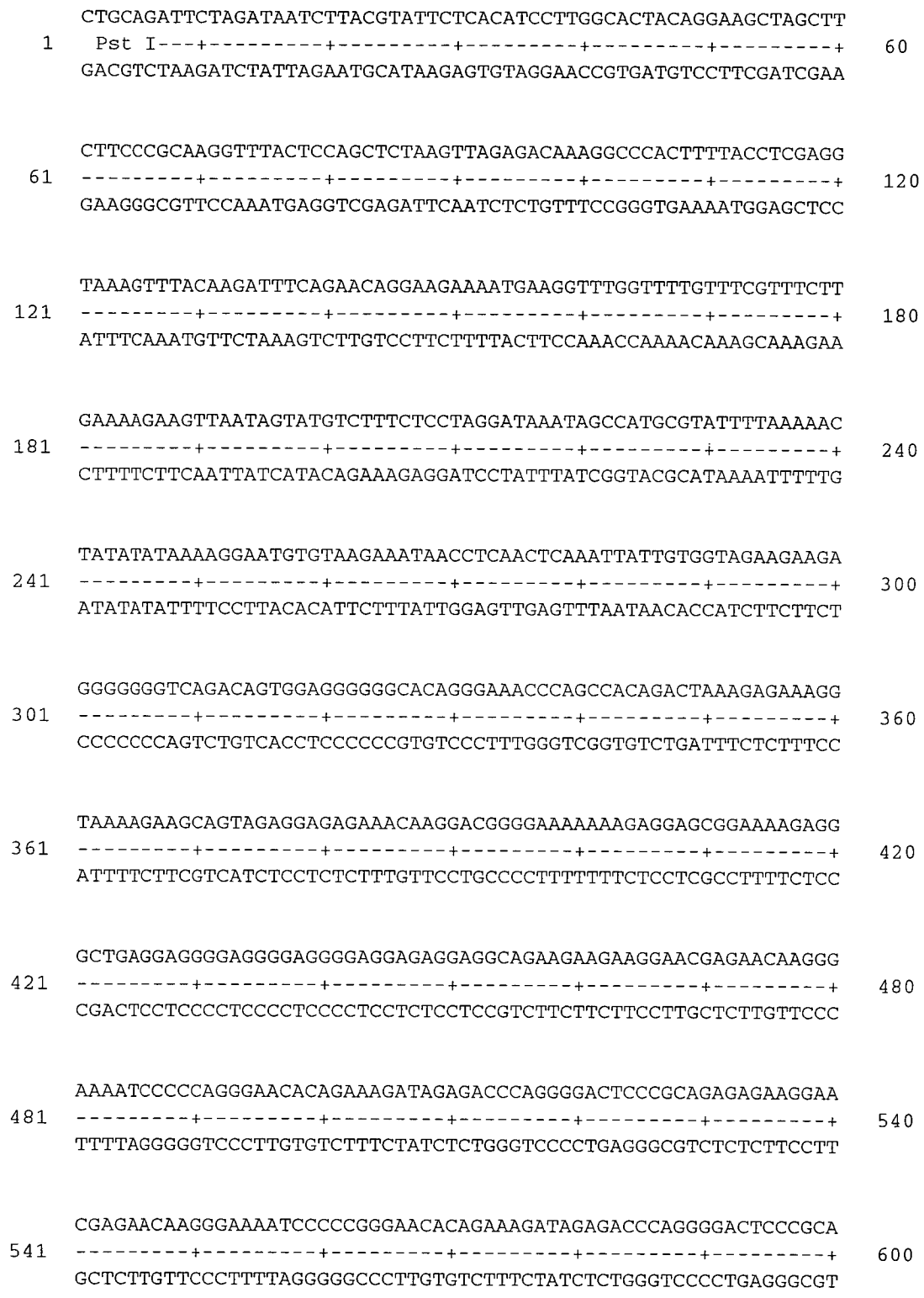
3.2.5 Sequencing of Human MSX1 subclones

The PCR amplification product was sequenced using a thermal-cycling method (GIBCO-BRL Cycle-Sequencing kit; section 2.6.2). This enabled rapid and precise generation of clear sequence data for this region. Oligos were end-labelled with ^{33}P giving strong, even bands. Sequence was initially derived from the same oligos used in amplification. Internal oligos were made to bridge the gap between the sequence from these and to confirm sequence on the opposite strand.

Sequence information from various oligonucleotides on both strands was assembled by visual comparison using the LINEUP program from the UWGCG package (Genetics Computer Group, 1991).

The sequence generated from the upstream region of the human MSX1 gene is shown in figure 3.10: in total 1195 base pairs were sequenced.

Figure 3.10


CTGCAGATTCTAGATAATCTTACGTATTCTCACATCCTTGGCACTACAGGAAGCTAGCTT
1 Pst I---+---------+---------+---------+---------+---------+---------+ 60
GACGTCTAAGATCTATTAGAATGCATAAGAGTGTAGGAACCGTGATGTCCTTCGATCGAA
CTTCCCGCAAGGTTTACTCCAGCTCTAAGTTAGAGACAAAGGCCCACTTTTACCTCGAGG
61 ---------+---------+---------+---------+---------+---------+ 120
GAAGGGCGTTCCAAATGAGGTCGAGATTCAATCTCTGTTTCCGGGTGAAAATGGAGCTCC
TAAAGTTTACAAGATTTCAGAACAGGAAGAAAATGAAGGTTTGGTTTTGTTTCGTTTCTT
121 ---------+---------+---------+---------+---------+---------+ 180
ATTTCAAATGTTCTAAAGTCTTGTCCTTCTTTTACTTCCAAACCAAAACAAAGCAAAGAA
GAAAAGAAGTTAATAGTATGTCTTTCTCCTAGGATAAATAGCCATGCGTATTTTAAAAAC
181 ---------+---------+---------+---------+---------+---------+ 240
CTTTTCTTCAATTATCATACAGAAAGAGGATCCTATTTATCGGTACGCATAAAATTTTTG
TATATATAAAAGGAATGTGTAAGAAATAACCTCAACTCAAATTATTGTGGTAGAAGAAGA
241 ---------+---------+---------+---------+---------+---------+ 300
ATATATATTTTCCTTACACATTCTTTATTGGAGTTGAGTTTAATAACACCATCTTCTTCT
GGGGGGGTCAGACAGTGGAGGGGGGCACAGGGAAACCCAGCCACAGACTAAAGAGAAAGG
301 ---------+---------+---------+---------+---------+---------+ 360
CCCCCCCAGTCTGTCACCTCCCCCCGTGTCCCTTTGGGTCGGTGTCTGATTTCTCTTTCC
TAAAAGAAGCAGTAGAGGAGAGAAACAAGGACGGGGAAAAAAAGAGGAGCGGAAAAGAGG
361 ---------+---------+---------+---------+---------+---------+ 420
ATTTTCTTCGTCATCTCCTCTCTTTGTTCCTGCCCCTTTTTTTCTCCTCGCCTTTTCTCC
GCTGAGGAGGGGAGGGGAGGGGAGGAGAGGAGGCAGAAGAAGAAGGAACGAGAACAAGGG
421 ---------+---------+---------+---------+---------+---------+ 480
CGACTCCTCCCCTCCCCTCCCCTCCTCTCCTCCGTCTTCTTCTTCCTTGCTCTTGTTCCC
AAAATCCCCCAGGGAACACAGAAAGATAGAGACCCAGGGGACTCCCGCAGAGAGAAGGAA
481 ---------+---------+---------+---------+---------+---------+ 540
TTTTAGGGGGTCCCTTGTGTCTTTCTATCTCTGGGTCCCCTGAGGGCGTCTCTCTTCCTT
CGAGAACAAGGGAAAATCCCCCGGGAACACAGAAAGATAGAGACCCAGGGGACTCCCGCA
541 ---------+---------+---------+---------+---------+---------+ 600
GCTCTTGTTCCCTTTTAGGGGGCCCTTGTGTCTTTCTATCTCTGGGTCCCCTGAGGGCGT

```
      GAGAGGGCCTCTTGGGCTCAGCGCAGAGGAAAGTTTCCCGGGCACCCCCTGTGGTCCCCT
601   ---------+---------+---------+---------+---------+---------+   660
      CTCTCCCGGAGAACCCGAGTCGCGTCTCCTTTCAAAGGGCCCGTGGGGGACACCAGGGGA

      GCACCTCCGCCGTGCCCTGCCCTGCGTGCCCCCAGGCCCAGCGCTCCGGGCGAGTCCCCA
661   ---------+---------+---------+---------+---------+---------+   720
      CGTGGAGGCGGCACGGGACGGGACGCACGGGGGTCCGGGTCGCGAGGCCCGCTCAGGGGT

      GGAGCGCGGCCCAATGGATCGGCTCCGGCCCGCCCCCTCGCGCGCTGATTGCCGCCGCCC
721   ---------+---------+---------+---------+---------+---------+   780
      CCTCGCGCCGGGTTACCTAGCCGAGGCCGGGCGGGGGAGCGCGCGACTAACGGCGGCGGG

      CCCGCTGGCCTCGCCTTATTAGCAAGTTCTCTGGGGAGGCGCGGTAGGGCCCGGAGCCGG
781   ---------+---------+---------+---------+---------+---------+   840
      GGGCGACCGGAGCGGAATAATCGTTCAAGAGACCCCTCCGCGCCATCCCGGGCCTCGGCC

      CGAGTGCTCCCGGGAAACATGCTGCCAGCGCGGCTGGCAGGCAAACGGAGGCCAGGGCCC
841   ---------+---------+---------+---------+---------+---------+   900
      GCTCACGAGGGCCCTTTGTACGACGGTCGCGCCGACCGTCCGTTTGCCTCCGGTCCCGGG

      AGTACGCCGGAGCTGGCCTGCTGGGGAGGGGCGGGCAGGCGCGCGCGGGAGGCGTGCCCG
901   ---------+---------+---------+---------+---------+---------+   960
      TCATGCGGCCTCGACCGGACGACCCCTCCCCGCCCGTCCGCGCGCGCCCTCCGCACGGGC

      CCAGGGCCCCGGGCGCTCGCAGAGGCCGGCCGCGCTCCCAGCCCGCCCGGAGCCCATGCC
961   ---------+---------+---------+---------+---------+---------+  1020
      GGTCCCGGGGCCCGCGAGCGTCTCCGGCCGGCGCGAGGGTCGGGCGGGCCTCGGGTACGG

      CGGCGGCTGGCCAGTGCTGCGGCAGAAGGGGGGGCCCGGCTCTGCATGGCCCCGGCTGCT
1021  ---------+---------+---------+---------+---------+---------+  1080
      GCCGCCGACCGGTCACGACGCCGTCTTCCCCCCCGGGCCGAGACGTACCGGGGCCGACGA

      GACATGACTTCTTTGCCACTCGGTGTCAAAGTGGAGGACTCCGCCTTCGGCAAGCCGGCG
1081  ---------+---------+---------+---------+---------+---------+  1140
      CTGTACTGAAGAAACGGTGAGCCACAGTTTCACCTCCTGAGGCGGAAGCCGTTCGGCCGC

      GGGGGAGGCGCGGGCCAGGCCCCCAGCGCCGCGGCGGCCACGGCAGCCGCCATGG
1141  ---------+---------+---------+---------+---------Nco I   1195
      CCCCCTCCGCGCCCGGTCCGGGGGTCGCGGCGCCGCCGGTGCCGTCGGCGGTACC
```

Figure 3.10 Sequence of the insert in clone pE31.1 corresponding to the 5'-flanking region of human MSX-1. The restriction sites used in the cloning of the pE31.1 insert, PstI and NcoI, are marked at the ends of the sequence (section 3.2.2). Oligo B962 is complementary to bases 698-715 (underlined; the adenine at position 715 was the 5' nucleotide).

3.3 Analysis and comparison of 5'-flanking sequences

Having determined sequence from the 5'-flanking region of the mouse and human *Msx*-1 gene I analysed it for several features. There has been considerable description of similar regions from a variety of genes enabling me to examine the *Msx*-1 genes for a number of features identified elsewhere.

3.3.1 Human and Mouse Msx-*1 genes are associated with CpG islands*

The dinucleotide CpG plays a specific role in the function of the vertebrate genome. It has been found to be unusually rare in vertebrate DNA where it is present at approximately one fifth of the expected level (Bird, 1986). Between 70% and 80% of CpG is methylated at position 5 of the cytosine ring. Deamination of 5-methylcytosine, to give thymine, is thought to be responsible for the rarity of the CpG dinucleotide. The distribution of CpG can be assessed by exploiting its presence in the recognition sites of various restriction endonucleases. Mapping the sites of cleavage by restriction enzymes such as *Mlu*I, *Not*I, *Xho*I *Sac*II and *Sma*I pinpoints regions relatively rich in CpG. This has led to the discovery that such sites often occur in clusters within the genome corresponding to the positions of genes, such that this feature is diagnostic of the location of a gene when analysing large sections of the genome. Such CpG 'islands' are proposed to be a consequence of differential methylation throughout the genome. Methylation in intergenic regions has led, gradually over evolutionary time, to a loss of CpG dinucleotides. Absence of methylation in the vicinity of genes is proposed as the reason for localised levels of expected CpG frequencies. It has been demonstrated that methylation of islands is associated with inactivation of the adjacent genes, suggesting a link between suppression of methylation and the mechanisms of gene regulation leading to maintenance of CpG. One possible relationship between the two would be a passive inhibition of upstream methylation by the binding to this region of the various *trans*-regulatory proteins associated with expression of the gene. Alternatively, methylation and its negative effects may have been selected against in promoter regions, indeed there is evidence to suggest that an active mechanism of methylation-suppression exists (Szyf *et al.*, 1990). The demonstrated absence of methylation at the island regions of several genes and the correlation of gene inactivation and methylation on the X-chromosome suggests that methylation may be part of a regulatory mechanism whereby certain, unmethylated, genes are rendered 'open' for transcription and have the *potential* for activation in response to specific regulators.

CpG islands are associated with all housekeeping genes studied so far; that is those genes performing essential functions in all cells and whose expression is constantly required (Antequera and Bird, 1993). According to the model described above, the constant presence of the transcriptional machinery at the promoter has prevented methylation and the mutability that it engenders. Such promoters often lack a TATA box and undergo transcriptional initiation at a variety of sites. It has been suggested that such multiple sites enable gene expression in the wide variety of cell types required because different combinations of transcription factors can act at different sites (Somma *et al.*, 1991). Tissue-specific genes are frequently inactive due to the specific nature of their expression requirements and would be expected to undergo methylation induced suppression of CpG. There are however examples of tissue-specific genes associated with CpG islands: the lens-specific rat γ-crystallin genes are all associated with CpG islands (denDennen *et al.*, 1989) while human α-globin and mouse MyoD are both associated with, and maintain, non-methylated CpG islands in tissues in which they are *not* expressed (Antequera and Bird, 1993). While it appears that most, if not all, CpG islands are associated with genes, the converse is not true and the relationship between the expression of a gene and its association with an island remains obscure.

Clearly, in a study of their transcriptional regulation it is of interest to examine the human and mouse *Msx*-1 loci for the presence of CpG islands. I have joined the upstream sequences, that I generated, with the published sequences (full intron sequences are not available for either gene so the sequence used is a fusion of the two exons) and produced nucleotide composition data for both species. Using the WINDOW and STATPLOT programs of the GCG package (Genetics Computer Group, 1991) I have produced plots of the CpG and GpC dinucleotide content. The WINDOW program works by passing a 'window' of specified size along the sequence in steps of a specified 'shift increment' and at each point calculating the percentage content for the given pattern. To produce these graphs I used a window of 100 bases in size and scanned at 1 base increments. The table of figures produced by WINDOW is graphically represented by STATPLOT. Figures 3.11 and 3.12 show the plots generated. In both cases the *x*-axis represents the sequence and the *y*-axis is the percentage dinucleotide content. I have provided a cartoon representation of the locus between the plots to demonstrate the position of the observed CpG clustering in relation to the structural features of the gene, also an arrow indicates the position of the transcriptional start site (Kuzuoka *et al.*, 1994). I visually divided the sequence into the region graphically demonstrated as having a raised incidence of CpG and the

Figure 3.11

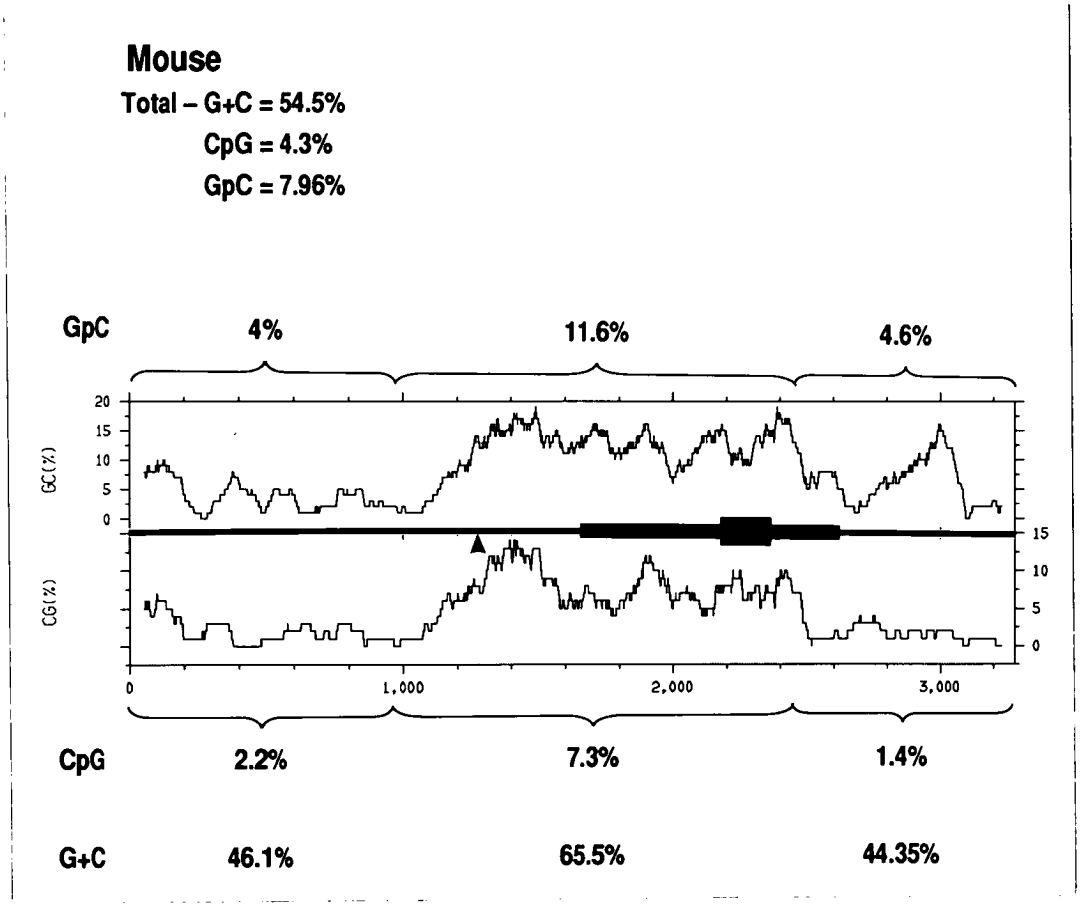


Figure 3.11 WINDOW plot of dinucleotide content across the mouse locus. The region covered is a fusion between the 5'-flank sequence reported here (figure 3.3) and the published cDNA (Hill *et al,* 1989), i.e. the intron is *not* included. Between the plots is a representation of the locus showing the ORF and homeobox in thicker boxes, respectively. Top is the G+C and dinucleotide frequencies for the whole region. Above and below are the dinucleotide frequencies for the individual regions (divided by visual criteria). Bottom are the G+C frequencies for the individual regions. The arrowhead on the central schematic marks the published position of the transcriptional start site (Kuzuoka *et al.*, 1994).

Figure 3.12

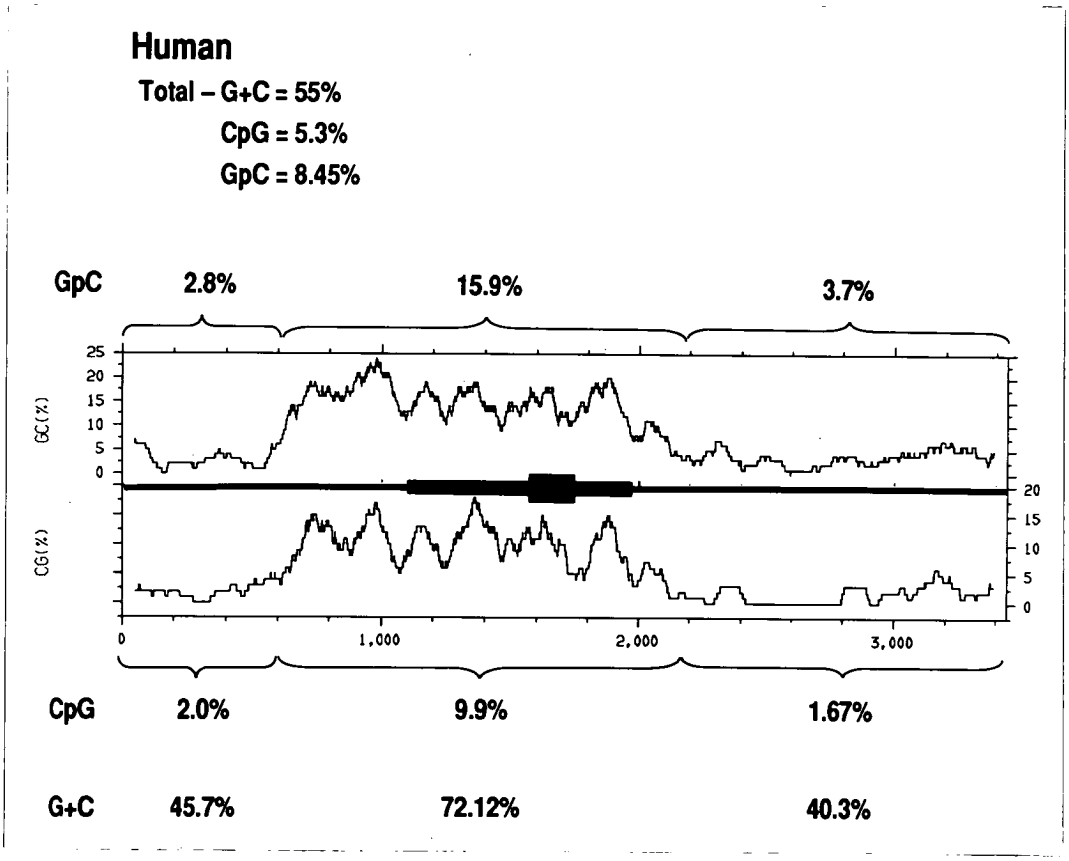


Figure 3.12 WINDOW plot of dinucleotide content across the human locus. The region covered is a fusion between the 5'-flank sequence reported here (figure 3.3) and the published sequence minus the intron sequence (Hewitt *et al,* 1991). Between the plots is a representation of the locus showing the ORF and homeobox in thicker boxes, respectively. Top is the G+C and dinucleotide frequencies for the whole region. Above and below are the dinucleotide frequencies for the individual regions (divided by visual criteria). Bottom are the G+C frequencies for the individual regions.

Table 3.1

a)

MOUSE	G	C	GpC	CpG	Total length (bp)
5' flank	217	244	40	22	1000
3' flank	165	181	46	11	780
'island'	475	508	175	109	1500
whole	857	933	261	142	3280

b)

HUMAN	G	C	GpC	CpG	Total length (bp)
5' flank	172	102	17	12	600
3' flank	264	278	50	22	1344
'island'	509	573	224	149	1500
whole	945	952	291	183	3444

Table 3.1 A&B: Composition data for the mouse and human locus. Numbers of C and G nucleotides and their dinucleotides are given along with the full length of each region (figures 3.11 & 3.12).

regions flanking this 'island' 5' and 3' (figures 3.11 & 3.12). Using the COMPOSITION program I determined the number of each nucleotide and dinucleotide in the island and flanking regions (Tables 3.1A & B). From these data I calculated the percentage G+C content. The DNA flanking the 'island' has on average 43% and 45.25% G+C content for the human and mouse genes respectively. This fits well with the previously determined 40% G+C content of bulk DNA (Bird, 1986). Within the 'island' regions this value rises to 65.5% in mouse and 72.1% in human. Again, this corresponds well with the 65% described for CpG islands (Bird, 1986; Antequera and Bird, 1993a).

Bird has described the island region as a sequence with a CpG content approximately 5 fold greater than that of bulk genomic DNA and with an approximately equal content of GpC and CpG dinucleotides, as would be expected in DNA free of suppression, (Bird, 1986). The percentage CpG content for the island regions was higher than that of the flanking regions by a factor of 5.5 (9.9÷1.8) in human and 4.06 (7.3÷1.8) in mouse, fitting the criteria well. The 'island' regions of the two *Msx*-1 cognates have a CpG:GpC ratio of 0.63 (7.3÷11.6) in mouse and 0.62 (9.9 ÷15.9) in human. These values depart somewhat from the 1.0 predicted for a CpG island, though they are closer to this value than the flanking regions, 0.42 (1.8÷4.3) and 0.56 (1.83÷3.25) for mouse and human respectively. This may reflect a state of decay for the CpG islands associated with mouse and human *Msx*-1.

The two regions ('island' and flank) have very different nucleotide content, with features reminiscent of a CpG island. This possibility is further examined below.

Is apparent clustering of CpG statistically significant?

To assess whether the apparent heterogeneity of CpG distribution revealed in figures 3.11 and 3.12 is genuine I have constructed contingency tables and performed a χ^2 analysis. The null hypothesis is that CpG is equally distributed along the length of the locus. Testing the significance of the uneven distribution between the island and non-island regions will elucidate the actuality of a CpG island. Tables 3.2 and 3.3 test whether there is significant variation between the proportion of CpG in the island and in the flanking regions. Expected values represent the number of CpG's that would be found in each region, as a fraction of the known total (table 3.1), if the dinucleotide were homogeneously distributed. The method used for calculation of the expected (legend to tables 3.2 & 3.3) does *not* take into account the variable G+C value for the alternative regions, it addresses only the question of whether there is CpG clustering. The probability of the result is determined by comparison to a χ^2-distribution table. Both mouse and human show highly significant ($p<0.01$) deviation from the null hypothesis showing that CpG's *are* clustered and validating the visual division of the

Contigency tables

Table 3.2

Mouse - CpG

a) OBSERVED

	CpG	Non-CpG	Totals
Island	109	1390	1499
Non-island	33	1747	1780
Totals	142	3137	3279

b) EXPECTED

	CpG	Non-CpG	Totals
Island	64.9	1434.1	1499
Non-island	77.1	1702.9	1780
Totals	142	3137	3279

$$\chi^2 = \frac{(109-64.9)^2}{64.9} + \frac{(1390-1434.1)^2}{1434.1} + \frac{(33-77.1)^2}{77.1} + \frac{(1747-1702.9)^2}{1702.9} = \mathbf{57.688}$$

$$\mathbf{p << 0.001}$$

Table 3.3

Human - CpG

a) OBSERVED

	CpG	Non-CpG	Totals
Island	149	1350	1499
Non-island	34	1910	1944
Totals	183	3260	3443

b) EXPECTED

	CpG	Non-CpG	Totals
Island	79.7	1419.3	1499
Non-island	103.3	1840.7	1944
Totals	183	3260	3443

$$\chi^2 = \frac{(149-79.7)^2}{79.7} + \frac{(1350-1419.3)^2}{1419.3} + \frac{(34-103.3)^2}{103.3} + \frac{(1910-1840.7)^2}{1840.7} = \mathbf{112.74}$$

$$\mathbf{p << 0.001}$$

Tables 3.2 & 3.3: Contingency tables used to determine significance of variation between Island CpG and Non-Island CpG frequencies. Expected values are based on the null hypothesis that the observed number of CpG dinucleotides are homogeneously distributed throughout the locus; e.g. expected CpG (island) = (total CpG/total length) x island length. The χ^2 value, equal to $\Sigma(O-E)^2/E$, is calculated beneath each pair of tables along with the probability (p) obtained from χ^2-distribution tables.

locus (into 'island' and flank) on the basis of the plots in figures 3.11 and 3.12. This crude test most likely reveals the unequal distribution of G and C nucleotides as similar tests showed clustering of GpC dinucleotides.

Do any of the regions show CpG suppression?

Is there any variation between 'island' and flank that is specific to CpG dinucleotides, possibly reflecting a discriminatory mechanism? Examination of tables 3.4 and 3.5 shows that in *all* cases the observed level of CpG dinucleotide is *lower* than the expected level. The status of GpC dinucleotide is more variable in this regard. I have used the binomial theorem to assess the significance of variation from the expected in the numbers of CpG and GpC in each region. The expected value in this case is calculated using the G and C composition data (table 3.1) for the respective regions. Therefore in this case there *is* correction for the effect of the general variation in G+C content.

In mouse *Msx*-1 I have shown that there is highly significant reduction of CpG in both island and flanking DNA suggesting a CpG suppression in all regions (Table 3.4; equations 1 & 2). This suppression is, however, much stronger in the flanking region than in the island (binomial fraction of 6.74 compared to 4.35). In contrast, there is no significant variation from the expected in the level of GpC dinucleotides in either the 'island' sequence or the sequence flanking it (Table 3.4; equations 3 & 4).

In human, CpG frequencies, as in the mouse, show a highly significant reduction for both island and flanking regions (Table 3.5; equations 5 & 6) though the deviation is again considerably greater in the flanking regions (binomial fraction of 5.58 compared to 3.47). Unexpectedly, the incidence of GpC within both island and flanking DNA is significantly[1] different from the expected frequency ($0.05 > p > 0.01$; Table 3.5; equations 7 & 8) however the discrepancy in the island is due to an excess of GpC and in the flanking DNA significant deviation, due to reduced GpC, is only found at the 5' flank (not shown).

There are no regions, in either the mouse or human loci, where CpG suppression is completely absent but the level of this suppression is markedly reduced in the 'island' region. It should be noted that the 'island' region as I have defined it extends beyond the region traditionally viewed as the promoter (section 1.5.1). One interpretation of these observations is that in the germ-line the 5' region of the gene and its immediate 5' flank have been protected to some degree, though not completely,

---

[1]A probability of ≤ 0.05 is considered statistically significant while ≤ 0.01 is considered highly significant.

Table 3.4

Mouse *Msx*-1

	Length (bp)	G + C %	Obs. CpG % (No.)	Exp. CpG % (No.)	Obs. GpC % (No.)	Exp. GpC % (No.)
Whole	3280	54.6 (857+933)	4.3 (142)	7.4 (243)	7.9 (261)	7.4 (243)
Island	1500	65.6 (476+508)	7.3 (109)	10.7 (161)	11.7 (175)	10.7 (161)
5' flank	1000	46.1 (217+244)	2.2 (22)	5.3 (53)	4.0 (40)	5.3 (53)
3' flank	781	44.3 (165+181)	1.4 (11)	4.9 (38)	5.9 (46)	4.9 (38)

CpG :-Island - $\left|\frac{(109-161)}{\sqrt{(1499\times 0.107\times 0.893)}}\right| = 4.35$ $p < 0.00003$ (1)

Flank - $\left|\frac{(33-91)}{\sqrt{(1780\times 0.051\times 0.949)}}\right| = 6.74$ $p < 0.00003$ (2)

GpC :-Island - $\left|\frac{(175-161)}{\sqrt{(1499\times 0.107\times 0.893)}}\right| = 1.17$ $p = 0.121$ (3)

Flank - $\left|\frac{(86-91)}{\sqrt{(1780\times 0.051\times 0.949)}}\right| = 0.538$ $p = 0.2964$ (4)

Table 3.5

Human *Msx*-1

	Length (bp)	G + C %	Obs. CpG % (No.)	Exp. CpG % (No.)	Obs. GpC % (No.)	Exp. GpC % (No.)
Whole	3444	55.1 (945+952)	5.3 (183)	7.6 (262)	8.4 (291)	7.6 (262)
Island	1500	72.1 (509+573)	9.9 (149)	12.9 (194)	14.9 (224)	12.9 (194)
5' flank	600	45.7 (172+102)	2.0 (12)	4.9 (29)	2.8 (17)	4.9 (29)
3' flank	1345	40.3 (264+278)	1.6 (22)	4.1 (55)	3.7 (50)	4.1 (55)

CpG :-Island - $\left|\frac{(149-194)}{\sqrt{(1499\times 0.129\times 0.871)}}\right| = 3.47$ $p = 0.00029$ (5)

Flank - $\left|\frac{(33-84)}{\sqrt{(1944\times 0.045\times 0.955)}}\right| = 5.58$ $p < 0.00003$ (6)

GpC :-Island - $\left|\frac{(224-194)}{\sqrt{(1499\times 0.129\times 0.871)}}\right| = 2.31$ $p = 0.01044$ (7)

Flank - $\left|\frac{(67-84)}{\sqrt{(1944\times 0.045\times 0.955)}}\right| = 1.86$ $p = 0.0314$ (8)

Tables 3.4 & 3.5: Measuring significance of variation in dinucleotide content using the binomial theorem: $(O-E)/\sqrt{npq}$ where O is observed value, E expected value, n=sample size (length) and p=probability with p+q=1: p=(E/n)/100. G+C and dinucleotide content are shown as percentages of the sequence region. Figures in brackets are actual numbers. The expected CpG and GpC content for a given region was calculated by multiplying the C fraction and the G fraction. A probability of 0.00003 represents the limit of the tables used.

from the molecular machinery effecting CpG loss. Such a reduction in promoter accessibility might be due to nucleosomes or molecular apparatus concerned with their positioning, or might result from near constant occupancy of this region by proteins directly regulating transcription. A second interpretation is that in a common vertebrate ancestor the *Msx*-1 gene 5' region was acted upon by the forces that created CpG islands (whatever they were/are) and that in the period since there has been a gradual and partial loss of the island, possibly at slightly different rates in the two lineages once they diverged, resulting in the present situation. The relevance, if any, of the methylation status in the germ-line (directly causative to maintenance of the island) to the generation of the complex and highly regulated embryonic expression patterns remains obscure and will require further insight into the origin and function of the CpG island as a general phenomenon.

3.3.2 Search for known transcription-factor binding-sites

For many transcription factors the target DNA sequence to which they bind has been identified. David Ghosh has generated a database comprising published consensus binding-sites and actual binding sites from individual genes; the Transcription Factor Database or TFD (Ghosh, 1990; 1992; 1993). This database is maintained at the UK HGMP-RC where it can be searched using the SIGNALSCAN program (Prestridge, 1991; 1993). Database consensuses are written using the IUAPC ambiguity codes (Nom. Cttee. Int. Union. Biochem., 1985) and SIGNALSCAN matches database entries with the query sequence according to the ambiguous and non-ambiguous positions. Using this tool I have searched the mouse and human *Msx*-1 5'-flanking sequences for matches to previously characterised transcription factor binding-sites. A sequence match to a specific consensus site does not necessarily imply that the transcription factor in question *does* bind to this sequence *in vivo*, merely that it *may be capable of binding*. A further qualification is that consensus sites are in many cases defined by *in vitro* binding assays and their applicability *in vivo* is often unclear. Determination of *potential* binding sites enables identification of candidates for regulatory genes acting upon *Msx*-1, which in some cases may seem particularly appropriate due to an overlap in temporal and spatial expression of the transcription factor and the *Msx*-1 gene. Functional conservation of *cis*-regulatory elements between species as divergent as *Drosophila* and mouse has been demonstrated (Awgulewitsch and Jacobs, 1992). The inclusion in TFD of *cis* elements from a wide variety of organisms may enable identification of protein binding sites for the products of non-mouse genes. This is particularly important when examining the flanking DNA from

genes, such as *Msx*-1, that were cloned by virtue of their homology to *Drosophila* genes. The presence of binding sites for *Drosophila* transcription factors may indicate structural conservation and the binding of homologous mouse genes to these sites.

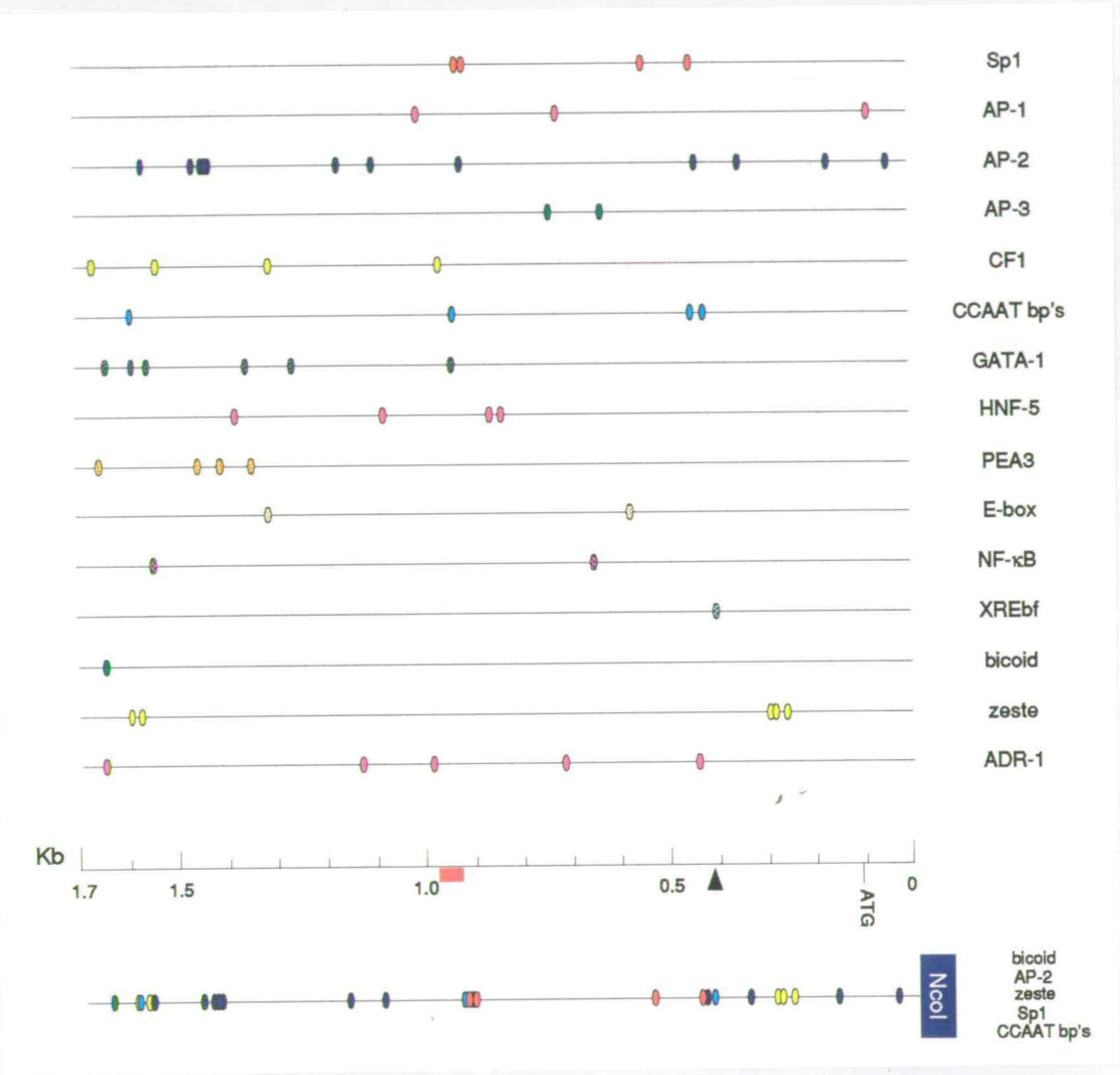
I have compared the 5'-flanking sequences of *Msx*-1 and MSX1 to all eukaryotic entries in the TFD. Both sequences provided matches with a large number and wide variety of different sites, reflecting the lack of functional criteria for matching. Sequences were found corresponding to binding sites for both general, ubiquitous transcription factors such as Sp1 and CTF and for specific, regionally expressed factors such as AP-2 (section 1.5.2). Due to the rather small size of some consensus sequences and the non-discriminating nature of those consisting of several ambiguous positions, such sites are found in large numbers in any sequence of this length. An example is the HSTF consensus binding site (Williams and Morimoto, 1990). This has the sequence NGAAN and clearly fortuitous matches will occur very frequently (once every 64bp in random DNA). For this reason I have omitted the matches found to consensus sequences of less than 5 bases (or equivalent). The single exception to this rule is the E-box (see below).

The large amount of data generated by a TFD search is unsuitable for presentation here, consequently I have been selective with the information presented. Many of the matches were omitted on the grounds of insufficient consensus length. Several were grouped under the heading of CCAAT binding proteins. These all bind the same sequence but represent a varied set of factors of both the general and specific class (see section 1.5.1). The sites remaining include those for a variety of factors identified in human, mouse, rat, *Drosophila* and even yeast. Figures 3.13 and 3.14 show graphical representations of the mouse and human *Msx*-1 5'-flanking sequences with the positions of a variety of putative binding sites, determined by TFD comparison, indicated by small coloured blobs. The relation of these positions to the start of translation is indicated on the scale at the bottom.

Of the factors that provided a binding site match within the sequences, I have selected those known as common proximal promoter factors and those which have or may have developmental significance based upon their expression pattern or the genes that they are known to regulate. Tables 3.6 and 3.7 summarises the factors included in figures 3.12 and 3.13 giving their recognition sequence, structural details, organism of origin and the frequency at which a match would be found in random DNA.

AP-1 is a transcriptional activation activity comprising the products of the *fos* and *jun* oncogenes (reviewed by Curran and Franza, 1988). It was first identified as an activator of the SV40 enhancer and the human metallothionein IIA gene. AP-1 has

Figure 3.13



Figures 3.13 Diagrammatic representation of the 5' *Msx*-1 sequence from mouse. Coloured blobs represent positions at which a match was found to the consensus binding-site of a specific transcription factor. The scale below marks the distance from the NcoI site conserved between the two species, the start codon and the position of the conserved sequence (red rectangle). An arrowhead denotes the transcriptional start site (Kuzuoka *et al.*, 1994). At the bottom is a composite of the potentially more interesting sites.

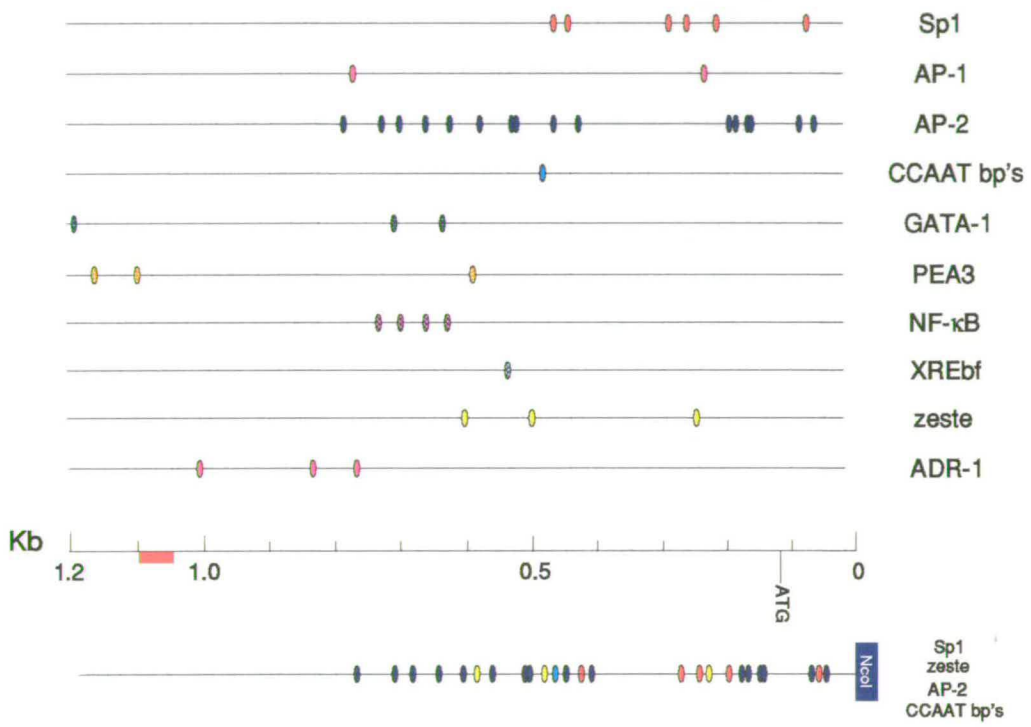
Table 3.6

Mouse

Factor	Concensus Recognition sequence(s)	Sequence(s) found	Structure	Species (where 1st identified)	Frequency of random occurrence (every *x* bp)
Sp1	GGGCGG	GGGCGG	zinc-finger	human	4096
AP-1	TGANTMA TGACTTCT GAGAGGA	TGATTAA TGACTTCT GAGAGGA	leucine-zipper	human	2048 65536 16384
AP-2	GSSWGSCC	GGCAGCCC GCCTCCC	helix-span-helix	mouse	4096
AP-3	TGTGGWWW	TGTGGAAA TGTGGAAA	?		2048
CF1	ANATGG	AAATGG ATATGG	?	human	1024
CCAATbp's	CCAAT	CCAAT	leucine-zipper		1024
GATA-1	GATAAG CCAATCT MYWATCWY TATCTT	GATAAG CCAATCT CCAATCAC CTAATCTC ACAATTCA TATCTT	zinc-finger	human	4096 16384 2048 4096
HNF-5	TRTTTGY	TGTTTGC TATTTGT	?	rat	4096
PEA3	AGGAAR	AGGAAA	ETS domain		2048
E-box	CANNTG CACCTG (MyoD)	CAAATG CACCTG	leucine-zipper	multiple	256 4096
NF-κB	GGGRHTYYHC	GGGGCTTTTC GGGACTTTTC	*Rel*-homology domain	human	14563
XREbf	CACGCW	CACGCT	zinc-finger	mouse	2048
bicoid	TCTAATCTC	TCTAATCTC	homeodomain	*Drosophila*	262144
zeste	YGAGYC CGAGCG	CGAGCC TGAGCC CGAGCG	?	*Drosophila*	1024 4096
ADR-1	GGAGA	GGAGA	zinc-finger	*S.cerevisiae*	1024

Table 3.6 Column 2 shows the consensus sequences derived from TFD. Column 3 shows the actual sequences matching the consensus that was found in the 5' flanking region of the mouse *Msx*-1 gene. Column 4 describes the structural class of the DNA-binding motif for the factor, where known. Column 5 states the organism in which the factor was first identified. Column 6 gives base frequency at which the concensus site would appear in random DNA (eg every 4096 bases).

Figure 3.14



Figures 3.14 Diagrammatic representation of the 5' *Msx*-1 sequence from human. Coloured blobs represent positions at which a match was found to the consensus binding-site of a specific transcription factor. The scale below marks the distance from the NcoI site conserved between the two species, the start codon and the position of the conserved sequence (red rectangle). An arrowhead denotes the relative position of the mouse transcriptional start site (Kuzuoka *et al.*, 1994). At the bottom is a composite of the potentially more interesting sites.

Table 3.7

Human

Factor	Concensus Recognition sequence(s)	Sequence(s) found	Structure	Species (where 1st identified)	Frequency of random occurrence (every *x* bp)
Sp1	GGGCGG	GGGCGG	zinc-finger	human	4096
AP-1	TGACTTCT	TGACTTCT	leucine-zipper	human	65536
	GAGAGGA	GAGAGGA			16384
AP-2	GSSWGSCC	GGCAGCCC			4096
		GCCTCCC	helix-span-helix	mouse	
	CCCCAGGC	CCCCAGGC			65536
	CCSCRGGC	CCCCGGGC			16384
CCAATbp's	CCAAT	CCAAT	leucine-zipper		1024
GATA-1	MYWATCWY	CCAATCAC			2048
		CTAATCTC			
		ACAATTCA	zinc-finger	human	
	TATCTT	TATCTT			4096
	WGATAR	AGATAG			1024
		AGATAA			
PEA3	AGGAAR	AGGAAA	ETS domain		2048
NF-κB	GGGRHTYYHC	GGGGCTTTTC			14563
		GGGACTTTTC	*Rel*-homology	human	
	GGGRNTYYC	GGGACTCCC	domain		8192
	GGGATTTTCC	GGGATTTTCC			1048576
XREbf	CACGCW	CACGCA	zinc-finger	rat	2048
zeste	YGAGYC	CGAGTC		*Drosophila*	1024
		TGAGCC	?		
	CGAGCG	GCAGCG			4096
ADR-1	GGAGA	GGAGA	zinc-finger	*S.cerevisiae*	1024

Table 3.7 Column 2 shows the consensus sequences derived from TFD. Column 3 shows the actual sequences matching the consensus that was found in the 5' flanking region of the human MSX1 gene. Column 4 describes the structural class of the DNA-binding motif for the factor, where known. Column 5 states the organism in which the factor was first identified. Column 6 gives base frequency at which the concensus site would appear in random DNA (eg every 4096 bases).

Figure 3.15

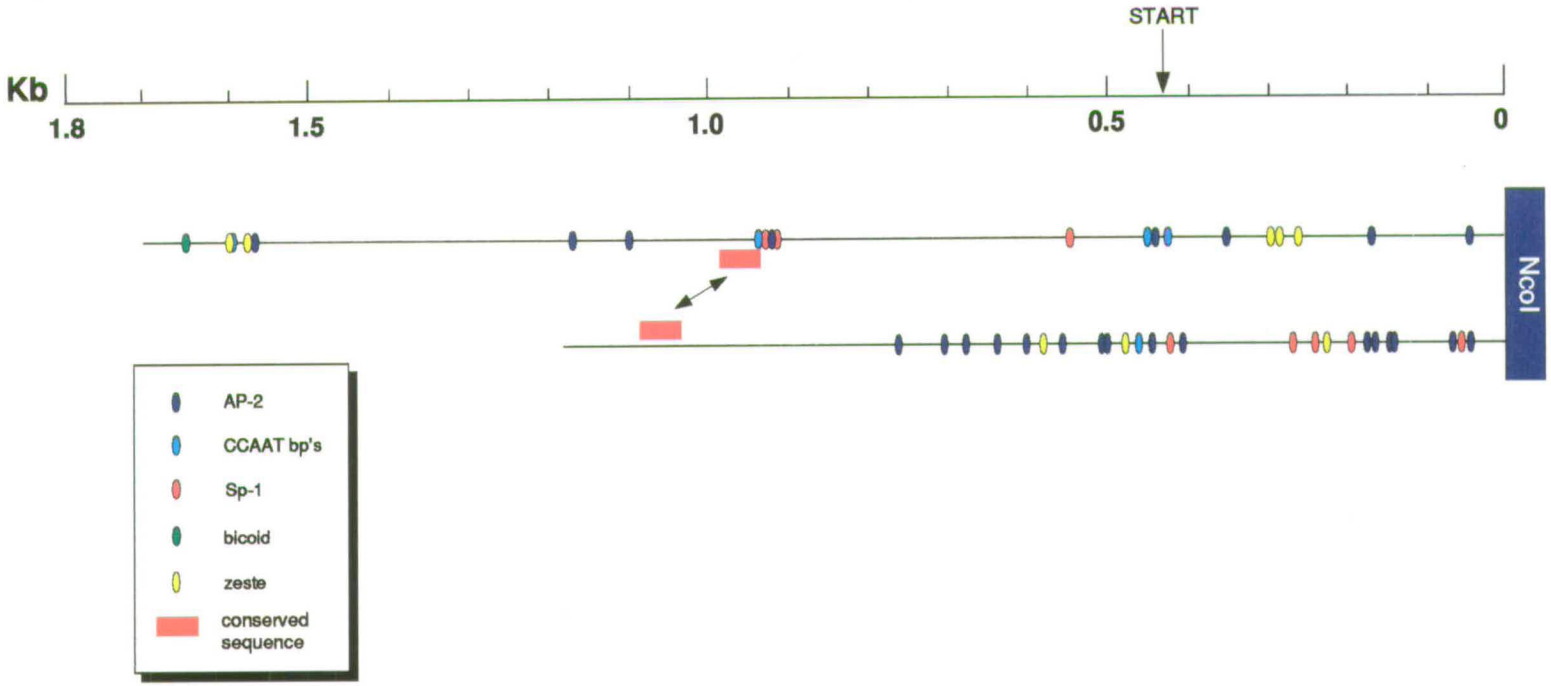


Figure 3.15 Compilation diagram providing a comparitive lineup of the two 5' flanking regions and the position of recognition sequences for a selected group of DNA-binding proteins. The arrow indicates the position of the mouse transcriptional start site (Kuzuoka *et al.*, 1994). The position of the conserved sequence is seen to be greater from the start site in human (by ~100bp) presumably caused by loss of sequence in the mouse or gain in human in the interval between the conserved sequence and the start site.

been shown to transduce the activation of these and other genes by phorbol esters (such as TPA), agents that mimic the action of diacylglycerol in the activation of protein kinase C (Angel *et al.*, 1987). This transduction pathway activates AP-1 by means of the phosphorylation-dependent inactivation of an AP-1 inhibitor, IP-1 (Auwerx and Sassone-Corsi, 1992).

The CCAAT binding protein group consists of CBP, CBF, CCAAT-bf, CRF, CTF and CDF (Cohen *et al.*, 1986; Graves *et al.*, 1986; Kingston *et al.*, 1987; Goding *et al.*, 1987; Green *et al.*, 1987; Barberis *et al.*, 1987). The consensus binding-sequences included in TFD vary subtly according to the bases flanking the core CCAAT motif and variants were found upstream of different genes. The references given above are those quoted by the database for each individual CCAAT binding activity: how many different genes encode these activities is not clear though it is likely that some represent the same gene product.

GATA-1 (also known as NF-E1, GF-1 & Ery-1) is a member of a family of transcription factors with related zinc-finger DNA-binding domains. Analysis of the preferred recognition sequences for several GATA family members has shown GATA-1 to have considerable variation in its target site specificity (Merika and Orkin, 1993). This is reflected in the variety and the degenerate nature of the consensus sequences present in TFD: tables 3.6 and 3.7 show the various matches found to these. Given the broad spectrum of consensus it is perhaps not surprising that matches are found in sequences of this length however they cannot be dismissed as purely fortuitous on this basis. GATA-1 has been studied in the context of its essential role in erythroid differentiation. Targeted disruption in ES cells has revealed a requirement for the gene in the erythroid lineages of chimaeric mice derived using these cells (Pevny *et al.*, 1991). This fits closely with the previously determined expression patterns and regulatory influences of this gene (Wall *et al.*, 1988). Studies on the expression of GATA family genes in *Xenopus* have revealed that GATA-1 is active in the early gastrula, possibly involved in commitment of mesoderm toward haemopoietic tissue (Zon *et al.*, 1991). Together these observations provide no support for the view that GATA-1 regulates expression of *Msx*-1 though this cannot be formally discounted.

The E-box is a binding site for a number of proteins, involved in developmental and differentiation pathways, which are characterised by the helix-loop-helix DNA-binding/dimerisation domain (Murre *et al.*, 1989a & 1989). These proteins are capable of forming heterologous complexes that determine their mode of action. One such protein is the myogenic determinant MyoD which binds the sequence shown in table 3.6. An E-box found in the proximal portion of the mouse *Msx*-1 promoter is not conserved in human and its significance is therefore doubtful.

Potential binding sites for *Drosophila* proteins bicoid and zeste are of interest for the reasons of functional homology described earlier. In mouse there is a single bicoid-binding site found near the distal end of the sequence determined. As shown in table 3.6 this sequence is a perfect match to the non-ambiguous site quoted by TFD. This consensus match is 9 bp in length and is therefore found fortuitously at a low frequency in random DNA (once every 262 kb), unfortunately its position at the distal end of the mouse promoter sequence puts it beyond the equivalent human sequence thereby preventing assessment of the level of conservation (further discussed in section 3.5). Several *zeste* sites are found in both mouse and human though close examination of the sequence reveals no conservation between the two (figure 3.15 shows no positional homology). Presence of these sites is likely to be a result of their high GC content and that of the 5'-flanking region, though maintenance by selection pressure cannot be ruled out.

Several AP-2 consensus binding-sites are found in mouse and human *Msx*-1 5-flanks. A small number are conserved between the two species but many matches are found to be due to the degenerate nature of the consensus sites, matching runs of cytosine for example. AP-2 is an attractive candidate as a regulator of *Msx*-1 for several reasons discussed in detail below (section 3.5).

3.3.3 Search for homology between Mouse and Human Msx-*1*

The strategy for identifying novel *cis*-regulatory regions acting upon *Msx*-1 was based upon a search for non-coding sequences that have been conserved, due to functional constraints, during the evolutionary interval separating mouse and human (~80 million years). Having obtained sequence of the proximal 5' flanking region of the gene for both species a comparison was carried out by computer.

The COMPARE program of the GCG sequence analysis package (Genetics Computer Group, 1991) is a useful tool in the sensitive comparison of nucleotide and peptide sequences. A 'window/stringency' comparison using this program enables the production of a graphical display showing similarity as a series of points plotted as Cartesian co-ordinates in units of nucleotides of the compared sequences. Briefly, the user specifies a 'window size' and a 'stringency'. The vertical sequence slides along the horizontal sequence comparing all possible combinations. At each combination the program slides a window along the pair of sequences and calculates the value of the matches for all the nucleotide pairs within the window (based upon a Symbol Comparison Table for all possible symbol combinations). If the sum of the values within the window is greater than or equal to the stringency, a match is recorded and a

Figure 3.16

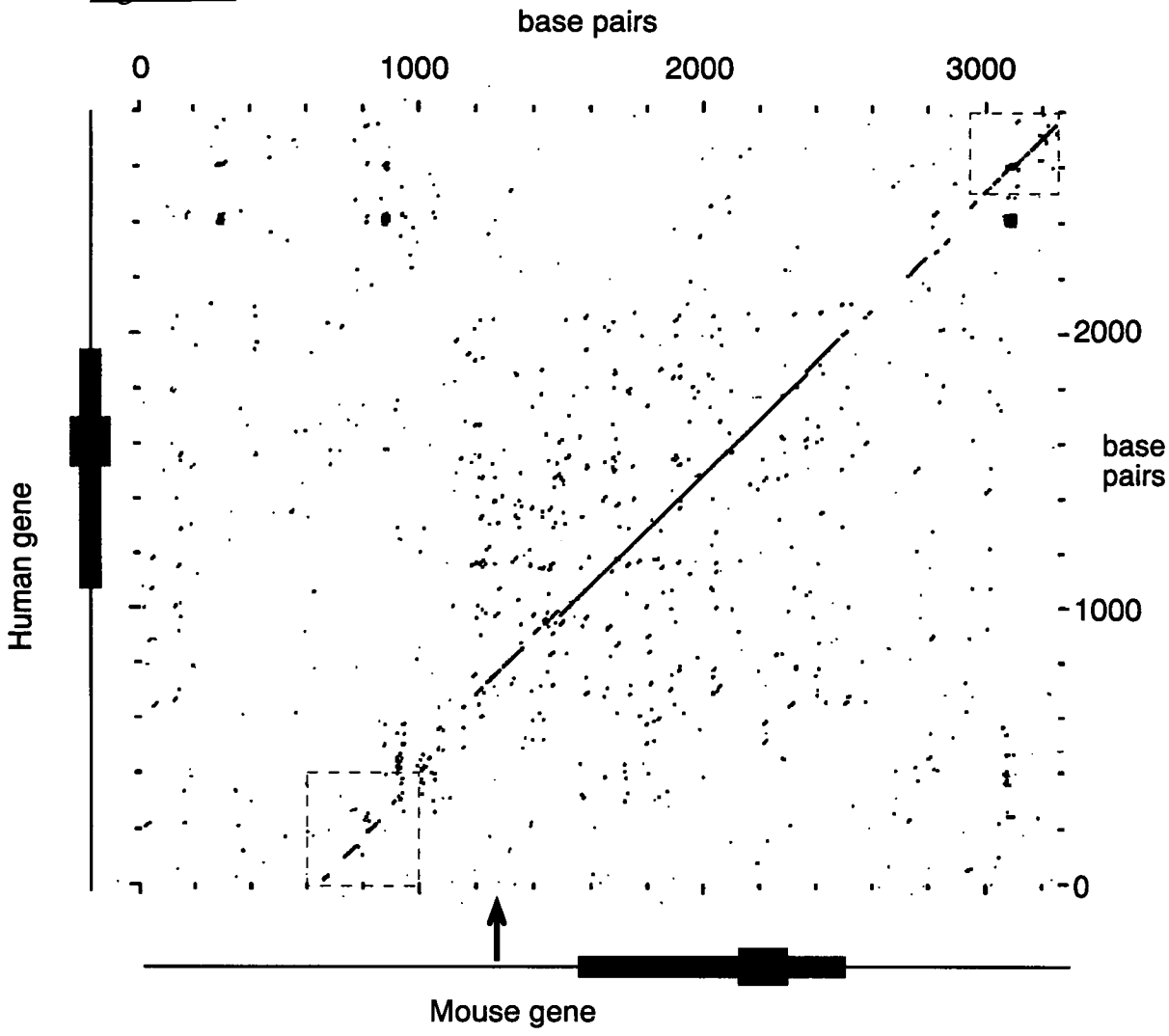


Figure 3.16 DOTPLOT comparison, generated by COMPARE, between promoter and coding sequences for mouse and human. Along the axes are schematic representations of the two loci indicating the position of the coding regions (thicker line) and the homeoboxes (even thicker line). The transcriptional start site of the mouse gene is marked by an arrow (Kuzuoka *et al.*, 1994). The dotted boxes highlight regions of homology outside the coding region (see text).

point is registered on the graph. COMPARE performs the calculations involved in this comparison and the graph is produced from these figures by the DOTPLOT program.

The default values for window size and stringency recommended for nucleotide comparison are 21 and 14 respectively. These values were used in the analysis and revealed the portions of the sequences likely to bear an interesting relationship to one another.

Figure 3.16 shows the DOTPLOT output for a default window/stringency comparison (alternative values were tested and shown to provide no additional information). The comparison was between the mouse and human *Msx*-1/MSX1 genes, a fusion of both previously published coding sequence (Hill *et al.*, 1989; Hewitt *et al.*, 1991) and 5'-flanking sequences described here (sections 3.1.5 & 3.2.6). It shows a long, largely continuous diagonal line running bottom left to top right. The central, unbroken line represents homology between the coding region, including the homeobox, and the sequences immediately 5' and 3' it. The diagrammatic representation of the loci next to each axis illustrates how the comparison relates to the different regions of the genes. The thickened line represents the coding region while the thicker section within it is the homeobox. We see that the extended solid diagonal closely corresponds to the open reading frames and that outside of them the homology is considerably reduced. Additional short regions of homology can be seen outside the open reading frame. These are highlighted and will be dealt with individually.

The comparison dotplot in figure 3.16 shows that conservation between the coding regions is considerable. Analysis with the GAP program shows there to be 88.75% identity (not shown; also Hewitt *et al.*, 1991). The dotplot also shows that homology extends beyond the initiation codon. The major line continues to a point close to the start of the published *Msx*-1 cDNA (Hill *et al.*, 1989). The recently published mouse cap site position is a C nucleotide at position 1283 (Kuzuoka *et al.*, 1994; figure 3.3). This position is marked on the dotplot (arrow, figure 3.16) and is surrounded by a region of high homology, distinct from the ORF.

Conservation in the 3' untranslated region

Homology at the 3' end ceases abruptly at the stop codon, however there is a region of homology at the extreme 3' end of the message (upper box, figure 3.16). Closer inspection of this region reveals approximately 90bp which is perfectly conserved between human and mouse (figure 3.17). Remarkably, when the chicken *Msx*-1 sequence (Robert *et al.*, 1991; Genbank accession number X61922) was examined for homology in this region a perfect match was found with the identity of

Figure 3.17

GAP of:
Mouse *Msx*-1 5' flank+cDNA from: 3000 to: 3280
to:
Human MSX1 5' flank+coding region from: 2500 to: 2800

Gap Weight: 5.000
Length Weight: 0.300

Quality: 186.0 Length: 320
Ratio: 0.662 Gaps: 3
Percent Similarity: 76.98 Percent Identity: 76.98

```
3000   .....GCAGCAAAGCATTGCTCTGAGGGGGCAGGGCGCATGCTGCTGCTT   3049
              | |||| ||||        ||||||||||    |  | | ||  |
2500   .....ACAACAAAACATTTGCTCTGGGGGGCAGGGAAAACACAGATGTGT   2540

3050   CACCAAGGTAGGTTAAAGAGACTTTCCCAGGACCAGAAAAAAAGAAGTAA   3099
         | |||||||||| ||| ||| |  |    |||         || |  |
2541   TGCAAAGGTAGGTTGAAGGGACCTCTCTCTTACC.....AGTACCAG      2582

3100   AAAAAAAAAAAAAAAAAAAAAAAAAAAACAAATCTGTTCTATTAACAGT   3149
       ||| | ||    ||||  |||||||||||| || ||| |    ||||||||
2583   AAACACAATTGTAAAATTAAAAAAAAAAAAAACTCTTTCTATTTAACAGT   2632

3150   ACATTTTCGTGGCTCTCAAGCATCCCTTTTGAAGGGACTGGTGTGTACTA   3199
       ||||| | |||||||| || ||| ||| | |||||     ||||||||||
2633   ACATTGT.GTGGCTCTGAAACAT.CCTCTGGAAGGATTATGTGTGTACTA   2680

3200   TGTAATATACTGTATATTTGAAATTTTATTATCATTTATATTATAGCTAT   3249
       ||||||||||||||||||||||||||||||||||||||||||||||||||
2681   TGTAATATACTGTATATTTGAAATTTTATTATCATTTATATTATAGCTAT   2730

3250   ATTTGTTAAATAAATTAATTTTAAGCTACA.....mouse.....       3280
       ||||||||||||||||||||||||||||
2731   ATTTGTTAAATAAATTAATTTTAAGCTACA.....human.....       2780
```

Figure 3.17 GAP comparison of the 3' untranslated ends of human and mouse *Msx*-1, showing the high level of conservation in this region. There is 77% identity over the whole sequence pictured here. At the extreme 3' end there are 90 bp of continuous sequence unchanged between the two species. Numbers relate to sequence position within the files generated by fusion of the 5'-flanking sequences and the open-reading frames. Details of the comparison are shown above. The upper sequence is mouse, the lower human.

Figure 3.18

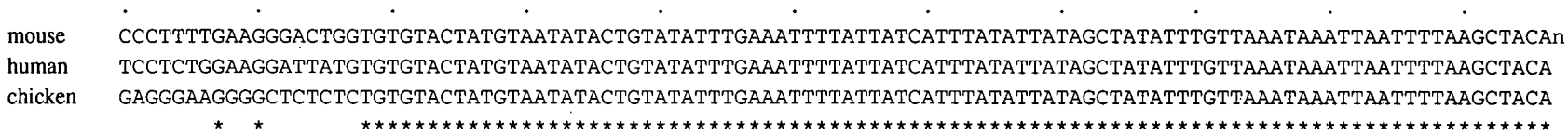


Figure 3.18 LINEUP comparison of the 3' ends of the mouse, human and chicken *Msx*-1 transcripts. The right-hand end of the sequence shown for mouse and chicken represents the known poly-adenylation site; the human gene was sequenced from a genomic clone but it is probable that the same position is also the polyA site.

the human and mouse sequences (figure 3.18). The identical sequences extend over the same length of DNA with only one extra base pair matching between human and mouse. A stretch of 88bp in a non-coding portion of the transcript (as far as we know, though the absence of variation at the 1st and 3rd positions of the 'codons' argues against conservation at the protein level) has been perfectly conserved since the divergence of the lineages leading to birds and mammals, over 300 million years ago (Benton, 1990). The extremely high level of conservation in this sequence surely reflects constraint by natural selection on a function that it performs. One such possible function is that the 3' untranslated region (UTR) has an independent, *trans* regulatory capacity. A *trans* regulatory role has been assigned to the 3' UTR of several myogenic genes where expression of this region of the transcript was associated with exit from the cell cycle and myogenic differentiation. (Rastinejad and Blau, 1993). Expression of *Msx*-1 has been associated with the opposite effect of blocking terminal myogenic differentiation (Song *et al.*, 1992; section 1.4.2), however the construct transfected in these experiments did not extend to the conserved 3' UTR. A common mechanism may be acting through the 3' UTRs of these genes: the expression of *Msx*-1 in several regions of active cell proliferation (e.g., the progress zone of the limb, facial primordia) and the link that has been established between exit from the proliferative phase and determination of cell fate, particularly in the limb (Summerbell *et al.*, 1973), make this an attractive proposal. Transfection experiments using a full length transcript of *Msx*-1 will be required to examine this possibility. No sequence data is yet available for the 3' UTRs of the myogenic genes concerned so comparison with the *Msx*-1 sequence was not possible, though comparison of this sequence with the GenEMBL database revealed no significant matches.

Conservation in the 5'-untranslated region

Figure 3.16 reveals homology extending beyond the start codon toward the transcription start site. Examination of the conservation in this region has revealed an interesting feature also found in other homeobox gene sequences. The 5'-untranslated region contains a small open reading frame also found in the human gene and in the chicken gene. Comparison of these reading frames shows that the carboxy-terminal portion of the protein is conserved while the remainder shows little homology across species. Possibly the most significant feature that is maintained is the length of the ORF which is 23 amino acids long in all three species. Figure 3.19 shows a lineup of the nucleotide and peptide sequences of the *Msx*-1 5' ORFs from these species, and a fourth short ORF found upstream of the *fushi tarazu* gene of *Drosophila*. The *Drosophila* ORF is of approximately the same size (22 amino acids) though shows no

Figure 3.19

A)

```
HUMAN     ATGCCCGGCGGCTGGCCAGTGCTGCGGCAGAAGGGGGGGCCCGGCTCTGCATGGCCCCGGCTGCTGACATGA
MOUSE     ATGATCCAGGGCTGTCTCGAGCTGCGGCTGGAGGGGGGGTCCGGCTCTGCATGGCCCCGGCTGCTGCTATGA
CHICK     ATGAGCCGCGGCCGCCGCTCCCGCGAGACGGCGACACGGGGCCGCCCTGCATGGCCCCGGCTGCGGAGATGA
          ***  *   *** * *     *       *  *    **  * ** ****************** *  ****

ftz       ATGCAGGATCTGCCGCAGGACCAGCTCATTCGCAAACTCACCAGCGTTGCGTGCACATCGCAGAGTTAG
```

B)

```
MOUSE     MIQGCLELRLEGGSGSAWPRLLL*
HUMAN     MPGGWPVLRQKGGPGSAWPRLLT*
CHICK     MSRGRRSRETATRGRPAWPRLRT*
          *  *            *****

ftz       MQDLPQDQLIRKLTSVACTSQS*
```

Figure 3.19 Lineup showing short open reading frames found in the 5' untranslated region of *Msx*-1 in mouse, human and chick. Asterisks represent positions of identity between the three at both nucleotide and peptide level. Also shown is the short open reading frame in the 5' UTR of *Drosophila ftz*. The mouse, human and chicken 5'-ORFs overlap the first 7 codons of the major reading frame. The DNA sequences shown begin at position 1016, figure 3.10 for human and position 1534, figure 3.3 for mouse

homology to the *Msx*-1 5' ORFs. In mouse, human and chicken this small ORF overlaps the first 7 codons of the reading frame encoding the homeobox protein. It is possible that in this position it has a regulatory influence upon translation of the major protein. Upstream open reading frames (uORFs) are a feature of the translational regulation of several *Saccharomyces cerevisiae* genes involved in amino acid biosynthesis, including the transcription factor GCN4 (Hinnebusch and Liebman, 1991). The introduction of a mutation of the *Msx*-1 uORF start codon would prove an interesting test of its function.

Conservation in the 5'-flanking sequence

The continuous line seen in figure 3.16, representing the highly homologous coding regions of mouse and human *Msx*-1, begins to break down the further 5' one looks. A region of approximately 150bp (from 1200-1350 in the mouse gene) shows homology at a position beyond the main block associated with the open-reading frames. This was initially viewed as a potential location for the transcriptional start site though the absence of a clear TATA box in either gene prevented any firmer prediction in the absence of experimental data. Recently published work (Kuzuoka *et al.*, 1994) has put the transcriptional start site at position 1283, in the centre of this conserved region. Gap analysis of the mouse sequence around the start and the human region highlighted as homologous from the dotplot reveals that the start site is flanked on either side by substantial homology, including a conserved CCAAT site immediately upstream (figure 3.20).

A region of conservation in the 5'-flank (lower box, figure 3.16) is shown enlarged in figure 3.21. This 400bp x 400bp comparison emphasises a short unbroken diagonal plot (boxed) revealing a region of homology near the end of the known human sequence. Gap analysis of this region is seen in figure 3.22.

The bracketed region in figure 3.22 comprises 44 bp of which 38 are identical between mouse and human (86%). Of the six differences between them, 3 are due to transversions and 3 to transitions. No sequences of similarly high homology are found within the surrounding regions (figure 3.16). Comparison of the 5'-flanking sequences shows a homology of 51% across the entire region. Clearly the sequence highlighted in figure 3.22 has an uncharacteristically high level of conservation. A sequencing-gel autoradiograph of this sequence from the mouse is shown in figure 3.23. In mouse this conserved sequence starts 808 bp 5' from the start codon, in human the homologous region is 929 bp 5' from the start codon. This difference of ~100 bp is likely to be due to expansion, in human, of an interval between the coding region and the homologous

Figure 3.20

GAP of:
mouse *Msx*-1 from: 1100 to: 1500
to:
human *Msx*-1 from: 500 to: 900

Gap Weight: 5.000
Length Weight: 0.300

Quality: 162.0 Length: 494
Ratio: 0.404 Gaps: 6
Percent Similarity: 63.312 Percent Identity: 63.312

```
                .         .         .         .         .
1100.............MOUSE.............TGTGGGGGTCCAGCCGG 1116
                                                |||  |  |||
 550 GGGAAAATCCCCCGGGAACACAGAAAGATAGAGACCCAGGGGACTCCCGC 599
               .         .         .         .         .
1117 ACCGATGCCCACCT.GACTTAGCTAGGCGGAAAAGCTCCCCAGGTACTCC 1165
     |  || | || | | | || |||   | |||||   ||||  |   | ||
 600 AGAGAGGGCCTCTTGGGCTCAGCGCAGAGGAAAGTTTCCCGGGCACCCCC 649
               .         .         .         .         .
1166 GGCTCTGTCGCCTGTGCGGGTCAGGCCCTTCCCCGAGCG.CCCCAGGCCG 1214
      |   |  |          | |  ||||| ||| | | | |||||||||
 650 TGTGGTCCCCTGCACCTCCGCCGTGCCCTGCCCTGCGTGCCCCCAGGCCC 699
               .         .         .         .         .
1215 AGCGCGCCTCGGG.....GCACGAGCACAGCCCAATGGTTCTCTCCGGAC 1259
     ||||| ||  | |       || |||| | |||||||||| ||    | | |
 700 AGCGCTCCGGGCGAGTCCCCAGGAGCGCGGCCCAATGGATCGGCTCCGGC 749
               .         .  ↓      .         .         .
1260 CCGCCCCCTCGCGCTCTGATTGGC.CGCTGCCACGCTGGCCTTGCCTTAT 1308
     |||||||||||||| |||||||  | |    || |||||||| |||||||
 750 CCGCCCCCTCGCGCGCTGATTGCCGCCGCCCCCCGCTGGCCTCGCCTTAT 799
               .         .         .         .         .
1309 TAACAAGTTCTCAGGGGAGCGGCGG.CGGACCCGGAGCCGGCGAGTGCGC 1357
     || |||||||||  ||||||  ||||  || ||||||||||||||||| |
 800 TAGCAAGTTCTCTGGGGAGGCGCGGTAGGGCCCGGAGCCGGCGAGTGCTC 849
               .         .         .         .         .
1358 CTGGG.AACTCGGCCTGAGCGGCGCAGGGATCCAGGCCCCGCTCGCTCGA 1406
     | ||| |||  |     ||||  || || |  ||  |   |  |
 850 CCGGGAAACATGCTGCCAGCGCGGCTGGCAGGCAAACGGAGGCCAGGGCC 899
               .         .         .         .         .
1407 GTTGGCCTTCTGGGGAAGCCGCAGGAGGCTCGCGCGCGAGAGCCGGCCGG 1456

 900 C...................HUMAN..................... 900
```

Figure 3.20 A GAP comparison of the region surrounding the transcription start site. The C residue at the start site is arrowed at position 1283 (Kuzuoka *et al.*, 1994). Upstream is a conserved CCAAT sequence, inderlined at position 1240.

Figure 3.21

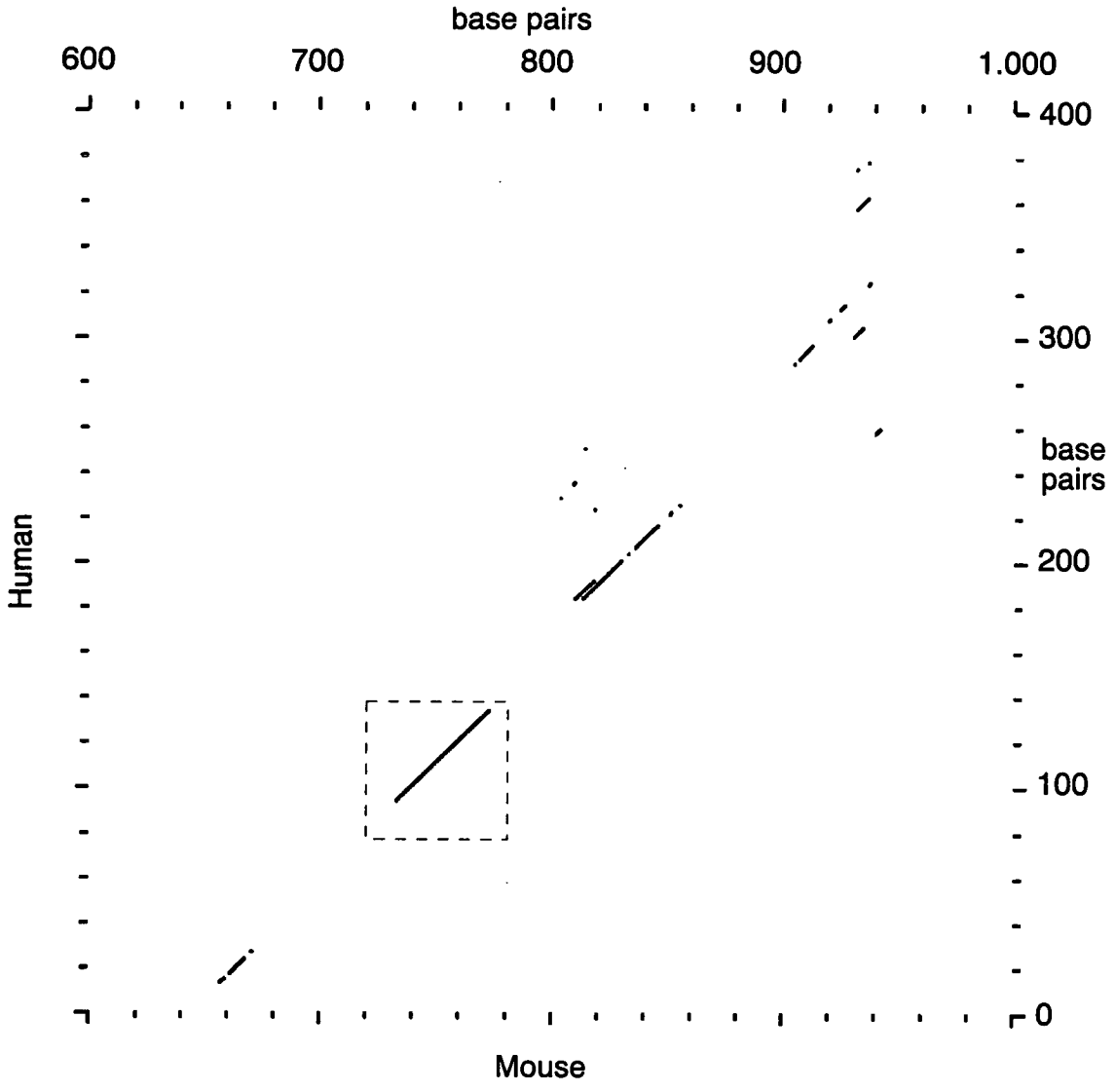


Figure 3.21 Expansion of the distal homologous region in the 5' flank of *Msx*-1, as highlighted in the lower dotted box of figure 3.16. Sequence of the region of homology highlighted here is shown in Figure 21.

Figure 3.22

GAP of:
Mouse *Msx*-1 5' flanking region: from: 700 to: 800
to:
Human *Msx*-1 5' flanking region: from: 60 to: 160

Gap Weight: 5.000
Ratio: 0.554
Percent Identity: 56.000

Quality: 56.0 Length: 102
Gaps: 0
Length Weight: 0.300

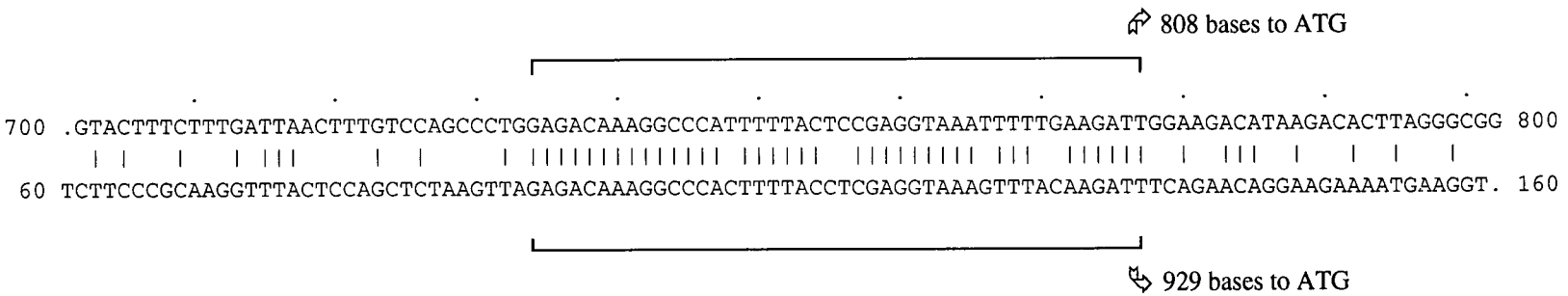


Figure 3.22 A GAP alignment of the sequences from mouse and human *Msx*-1 5' flanking region shown to be similar by the COMPARE analysis (figures 3.16 & 3.21). The upper sequence is that of mouse, the lower is human. The brackets highlight the region of homology. Details of the alignment are given above.

Figure 3.23

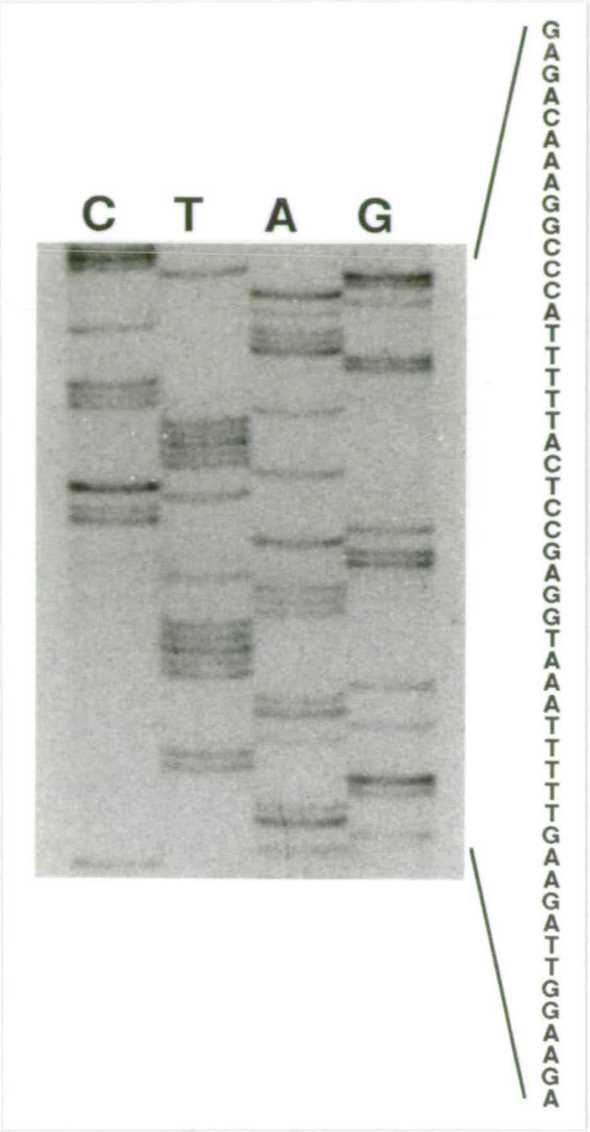


Figure 3.23 Autoradiograph showing the products of nucleotide sequencing from the conserved 5'-flanking region of the mouse *Msx*-1 gene. At the right, going 5' to 3' in descending, is the nucleotide sequence itself.

sequence. Sequence between bases 300-500 (section 3.2.6) is a potential site for such expansion due to the long stretches of polyA and polyG.

3.3.4 Sequence homology is found with the Hox*D cluster*

The conserved sequence identified upstream of *Msx*-1 was compared with the Genbank/EMBL database and the Eukaryotic Promoter Database (Bucher, 1993) to determine whether a similar sequence was present in any previously identified gene or gene promoter. In addition to this search of computer databases, I compared the *Msx*-1 conserved sequence with promoter sequences being published at the time that would not yet be included in the databases. This latter source was the sole origin of significant homology. One such sequence was published as part of an evolutionary comparison (similar to that presented here) between mouse and human *Hox*d-9 and *Hox*d-10 (Renucci *et al.*, 1992). Renucci and colleagues sequenced the region of the *Hox*D cluster between these two genes from both mouse and human and compared the sequences from the two species in an attempt to identify regions conserved due to functional constraints. I compared the conserved *Msx*-1 sequence (described in section 3.3.3) with this entire intergenic region (3657 bp), using a combination of the COMPARE and GAP programs, and discovered that there was homology between the *Msx*-1 sequence and a portion of the *Hox*D sequence which had itself been conserved between mouse and human. Bases 5123-5167 of figure 2 from the paper by Renucci and colleagues have similarity to the *Msx*-1 conserved sequence (opposite orientation): this sequence is seen in figure 3.24B. Figure 3.24 is a lineup of the various similar sequences illustrating the level of homology and the positions conserved. Figure 3.24A shows that of the region conserved between mouse and human *Msx*-1, the 5' half is also highly homologous to the sequence in the *Hox*d-9/d-10 intergenic region. This region is itself conserved between mouse and human (figure 3.24A; Renucci *et al.*, 1992).

Figure 3.24

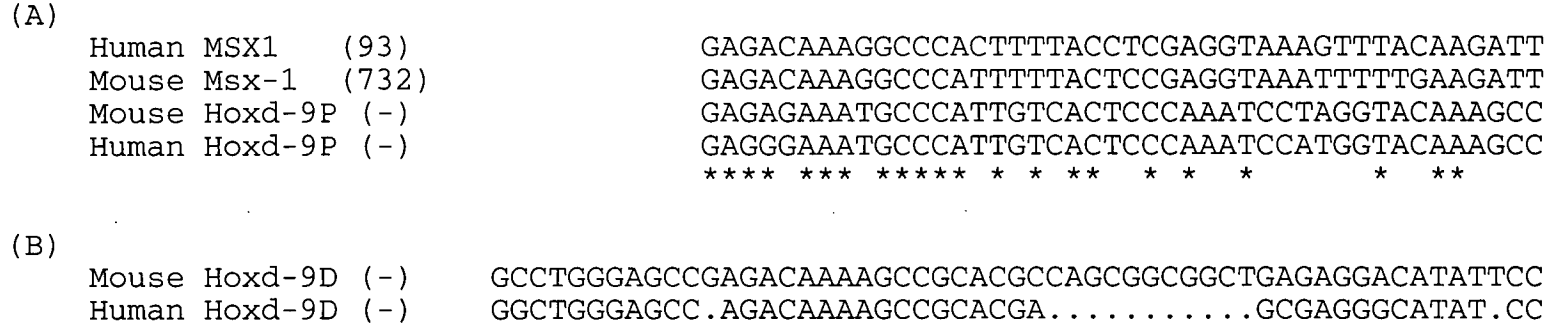


Figure 3.24 Side-by-side comparison of the homologous sequences found upstream of mouse and human *Msx*-1 and *Hox*d-9. A) shows a line up of the two sequences from both species with the invariant positions marked below with asterisks. B) shows an additional similar sequence from *Hox*d-9 and its human counterpart. This sequence also comes from region B of Renucci *et al* (1992). The numbers in brackets refer to the positions of the sequence in figure 3.3 for mouse *Msx*-1 and figure 3.10 for human MSX1. A minus sign (-) indicates that this sequence is found in the opposite orientation with respect to transcription. Other sequences shown are found in the same orientation as the transcript. The human sequence in B) is shown including gaps for optimal alignment.

3.4 Comparison of *Msx* gene coding sequences

The evolutionary relationship between *msh*-like genes has clear relevance to the project described and in this regard I attempted to clarify this relationship by comparison of the gene sequences. Several computer programs are available to enable multiple alignment of sequences and to provide data from which one can determine sister groups and construct unrooted cladograms. From such comparisons it is possible to infer an evolutionary history of the *Msx* genes.

Invertebrates (*Drosophila*; Gehring, 1987) and primitive chordates (Holland, 1991 and personal communication) have a single *Msh*-like gene, while vertebrates characteristically have three such genes (section 1.4). It appears that there was a series of duplication steps, along the lineage leading to *vertebrata*, resulting in the modern situation. Comparison of *msh*-like genes from a variety of organisms may elucidate both the sequence of duplication events that has led to this amplification, and the relative position of each event in such a phylogeny. Holland has proposed that the vertebrate-specific amplification may have had a causal role in the evolution of the sub-phylum as the genes are expressed during the formation of several of the features that distinguish vertebrates from other chordates. A combination of sequence-determined phylogenies and *in situ* study of expression patterns will enable a better resolution of the relationship between these genes and the evolution of the vertebrates.

The *msh*-like genes represent an ideal opportunity for a truly informative comparison across phyletic boundaries due to the wide variety of organisms from which genes of this family have been cloned. Unfortunately sequences from some of the more 'primitive' organisms (those capable of clarifying the fundamental structure at the base of a phylogenetic tree) are limited to reports of homeobox sequence. Cnidarian, Ascidian, *Drosophila* and Honeybee sequences available at present are limited to the homeobox region. Given the highly conserved nature of the homeobox, its comparison between different species does not reveal an instructive degree of divergence. Trees that were constructed using homeobox sequence alone demonstrated this in the clearly incorrect nature of some of the clades predicted (e.g. chicken and human more closely related than mouse and human). Reluctantly it was decided to confine the comparison to genes for which full coding sequence was available. In practice this restricted comparison to vertebrate *msh*-like genes. While little new information regarding the evolutionary timing of duplication events in the *msh*-like gene family will be forthcoming from such a comparison, it is a useful exercise in assessing the relative divergence of mouse and human *Msx*-1 and other orthologous genes. The comparison performed also clarifies the paralogous or orthologous status

of various *msh*-like genes. Full coding sequence is available for two mouse, two human, two Xenopus, two chicken, five zebrafish and one quail gene. Sequences were compared using the CLUSTAL program of the GCG package (Genetics Computer Group, 1991). This program uses the Neighbour Joining (NJ) method of Saitou and Nei (1987) to produce cladograms. This is a distance method which calculates similarities by the method of Wilbur and Lipman (1983). Figure 3.25 shows a cladogram generated using data in the '.nj' output file of CLUSTAL. The tree was rooted by inclusion of an outgroup known to be unrelated (human IGFII promoter was used here: the same root position was obtained using alternative outgroups). The tree divides into two major branches, one for the *Msx*-1 lineage and one for *Msx*-2. The *Msx*-2 branch seems to conform to accepted views of phylogeny based on paleontological data. The most closely related species are quail and chicken with human and mouse also closely linked to one another and comprising a clade along with the birds. The bird and mammal clade branches earlier from the amphibia, represented by *Xenopus*. Together the tetrapods diverge from the bony fish (exemplified by zebrafish) at a yet earlier stage, though still post-dating the divergence of the *msh* orthologues. The *Msx*-1 branch conforms in all but one respect. Again mouse and human are closely related, branching from chicken. The clade comprising birds and mammals, however, branches from zebrafish which clade branches earlier from *Xenopus* causing an apparent reversal of the relationship between fish and amphibia. The relationship of the various zebrafish genes is a point of considerable interest. There are at least 5 *msh*-like genes in the zebrafish genome and relating them to the more familiar vertebrate genes may provide information with regard to the timing and mechanism of their amplification. MsxA and MsxD form a trichotomy with the *Msx*-1 branch suggesting an additional duplication in this lineage soon after the initial duplication of the ancestral *msh*-like gene, which may have been coincidental with (or even causal to) the vertebrate radiation. MsxB and MsxC appear to have derived from a secondary duplication event in the *Msx*-2 lineage, though possibly later than that producing MsxA & D. The position of MsxE within the *Msx*-1 clade is more problematic and probably results from uncharacteristic rates of substitution. The extra genes present in the zebrafish genome may be due to an early genomic duplication (tetraploidy), or partial genomic duplication event. Such a hypothesis is supported by the presence of atypical numbers of members in other gene families in this organism (Ekker *et al.*, 1992a).

This analysis confirms that the primary *msh* duplication event within the chordates occurred prior to the vertebrate radiation, and clarifies the paralogous relationship between Zebrafish *Msx*E and the various *Msx*-1 genes, and Zebrafish

Figure 3.25

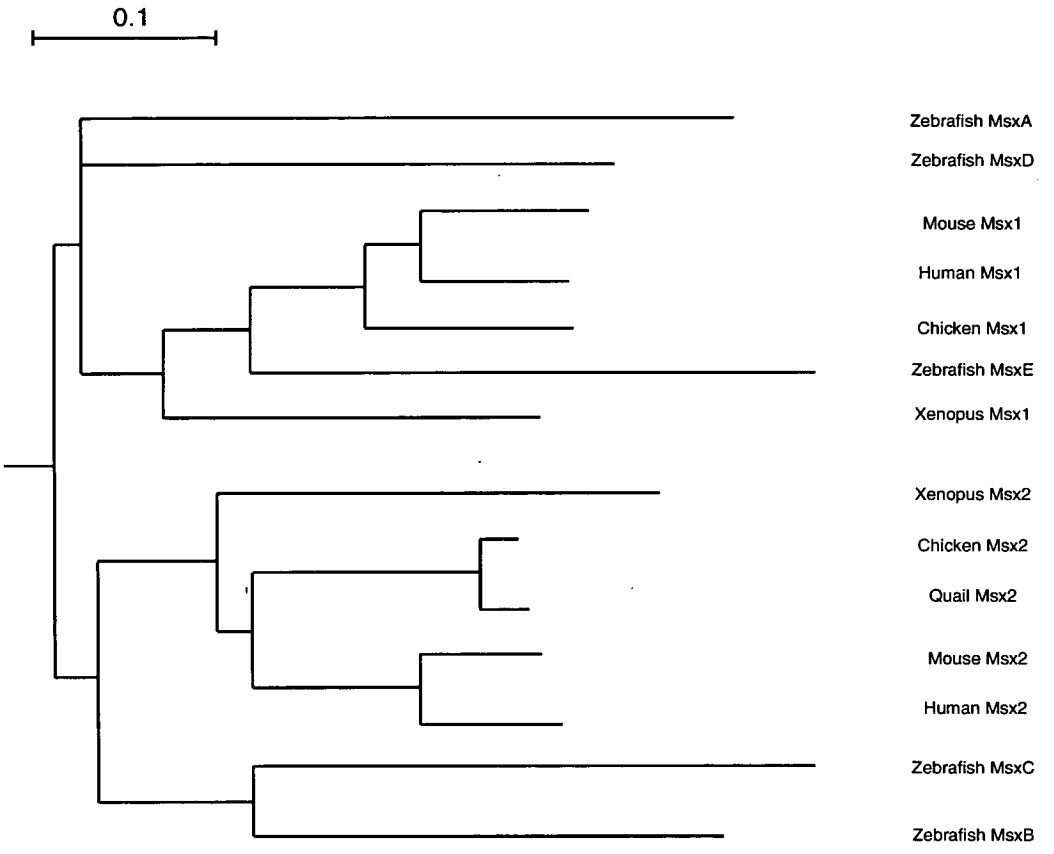


Figure 3.25 Cladogram displaying the phylogenetic relationships between the vertebrate *msh*-like homeoboxes. Calculated using full-length nucleotide coding sequences by a nearest-neighbour method (see text). The tree is rooted by inclusion of a known outgroup.

*Msx*C and *Msx*D and the *Msx*-2 genes. In the absence of full-length *Msx*-3 coding sequence the story is limited to the two genes described, though phylogenies based on second exon sequence alone (not shown) suggest that the *Msx*-3 branch may have split first from the branch leading to a later divergence of the *Msx*-1 and *Msx*-2 genes However this data is not conclusive due to the problems of comparing homeoboxes (encoded in the second exon), where a single base substitution may result in a shift between clades, or even major branches.

3.5 Discussion

I have described analysis of 5'-flanking sequences of the human and mouse Msx-1 genes. Examination of the nucleotide content, a search for specific sequence motifs and comparison of the two sequences has pointed to several features characteristic of transcriptionally active regions and to sequences with a probable role in transcriptional regulation.

Analysis of the mouse and human *Msx*-1 loci, including the 5'-flanking sequences, provides evidence for relief of CpG suppression in the 5' region of the gene. The existence of CpG islands is dependent upon their non-methylation in the germ-line. While the methylation state of the island in somatic cells may be significant to expression or repression of the associated gene it has no bearing upon the next generation. The absence of methylation at an island in germ-line cells should ensure maintenance of CpG content due to escape from the suppressing mechanism described, though the mechanistic link between promoter activity and suppression of methylation remains unclear. Based upon the data presented, an assessment of the state of the *Msx*-1 promoter in the germ-line might be expected to reveal reduced methylation. *Msx*-1 is not constitutively active but apparently has a promoter that escapes methylation to some degree. As mentioned there are examples of genes with restricted expression patterns that are associated with CpG islands. Larsen *et al.* (1992) have calculated that of among all genes, in human, associated with CpG islands approximately half are restrictively expressed genes, representing about 40% of the total number of genes expressed in this way. As alluded to previously (section 3.3.1), *Msx*-1 may have a promoter of the type at which there is assembly of the transcriptional machinery but pausing of elongation (Lee *et al.*, 1992). This situation may enable rapid induction of transcription in response to signals such as are responsible for the epithelial-mesenchymal interactions with which *Msx*-1 is implicated. Such a situation might mimic that found in constitutively active genes and result in the observed island effect. This possibility is discussed below in regard to potential GAGA elements in the conserved region. Alternatively, the CpG state of the promoter (showing raised levels but not to the degree expected in the absence of methylation) may reflect the presence of a 'fully intact' CpG island at some time during the history of the gene. This 'island' may now be in decay and what we see is a snapshot of this process. Such a model has been proposed by Antequera and Bird (1993b) who suggest that over time an increasing number of genes are losing their CpG islands, "succumbing to methylation" in their words, and that genes associated with such loss are exclusively those showing

restricted expression patterns. Interestingly, Antequera and Bird suggest that this process of island loss has occurred more rapidly in mouse than in human. If the *Msx*-1-associated islands in mouse and human are in a process of gradual erasure they appear not to conform to this general species difference.

The presence, in the *Msx*-1 promoter region, of consensus binding sites for a variety of transcription factors points directly to possible regulators of the gene. The properties of several of these factors are discussed below with reference to features that make some better candidates than others.

Sp1 is a common site found in the proximal regions of a large number of promoters. It is generally considered a basal promoter element though sites are also found in enhancer regions providing transcriptional specificity. Sp1 acts synergistically and a number of sites are often found in proximity to one another (Courey *et al.*, 1989; Anderson and Freytag, 1991). Figures 3.13 and 3.14 show several consensus Sp1 sites in mouse and human *Msx*-1 promoters. Comparison of the position of Sp1 binding sites does not pinpoint any conservation (figure 3.15). This may reflect the general maintenance of consensus sites (GGGCGG) within GC-rich regions such as these, or that the sites are fortuitous and non-functional. The proximal sequence of *Msx*-1 shows no obvious consensus TATA-box (Kuzuoka *et al.* (1994) have recently highlighted a TATA sequence at 1104 bp upstream of the CAP site as determined by them. I have not considered this sequence given the uncharacteristically large distance involved). There is evidence that binding of Sp1 assists in positioning the initiation complex in the absence of a TATA-box (Pugh and Tjian, 1991). The numerous Sp1 sites in the mouse and human *Msx*-1 promoters could participate in such a mechanism.

CCAAT binding factors are common proximal promoter regulators (section 1.5.1). It is of no great surprise that both mouse and human *Msx*-1 promoters contain this sequence motif. Examination of the two sequences shows that the single CCAAT sequence in the human promoter is conserved with the second of the three CCAAT sequences found in the mouse (figure 3.15). This sequence may represent the binding site for a ubiquitous proximal promoter transcription factor.

AP-2 is a transcriptional activator protein initially identified as a factor binding to the promoter of the human metallothionein II_A gene and to SV40 enhancers (Imagawa *et al.*, 1987). The gene encoding AP-2 has been cloned revealing a novel class of regulator but bearing a common structural theme, a DNA-binding/dimerisation domain, in this case termed a 'helix-span-helix' motif (Williams *et al.*, 1988; Williams and Tjian, 1991b). This gene encodes both the originally identified activator protein and, by means of alternative splicing, a second protein apparently lacking DNA-binding

activity but acting as a repressor of the activator function (Buettner *et al.*, 1993). It is of particular interest that AP-2 binding sites should be found upstream of *Msx*-1 (both mouse and human) as the two genes are expressed in a largely overlapping pattern during mouse embryogenesis (Mitchell *et al.*, 1991; section 1.4). AP-2 is first expressed in the anterior portion of the embryo at 8.5 dpc. As development progresses expression is seen in the folds of the neural tube that will contribute the migratory neural crest cells. Sites such as the primordia of the developing face, the facial and acoustic ganglia primordia and the dorsal root ganglia of the spinal cord are all derived from the neural crest and show marked expression of AP-2. The early limb-bud expresses AP-2 in a diffuse pattern throughout, though as it develops further expression becomes confined to the distal mesenchyme, corresponding to the progress zone. Significantly, these sites of AP-2 expression in the developing craniofacial structures, limb and nervous system are all targets for retinoid-induced embryopathy. The RA-dependent induction of transient AP-2 expression in human teratocarcinoma cells had previously been characterised (Lüscher *et al.*, 1989). Together these observations lend support to the idea that AP-2 may be involved in a regulatory cascade activated by retinoids, possibly effecting the morphogenetic responses that they induce. Expression of *Msx*-1 in neural crest cells and lineages to which they contribute has been described (section 1.4). Expression patterns of *Msx*-1 in the facial primordia and developing limbs are particularly striking in their similarity to those of AP-2. The coincidental expression, in several domains, of *Msx*-1 and various retinoic acid receptors (RARs; Lyons *et al.*, 1992), and the demonstration that RA can modify *Msx* gene expression in the limb (Yokouchi *et al.*, 1991) provides a further potential link between AP-2 and *Msx*-1, as RA-regulated genes. There are clearly large numbers of cells within these regions of the embryo that will be expressing both *Msx*-1 and AP-2 simultaneously. In this context the presence of AP-2 consensus binding sites upstream of mouse and human *Msx*-1 may reflect a regulatory interaction between the two, possibly in response to RA-induced signals. It will be of interest in this regard to discover whether the AP-2 promoter contains *Msx*-1 binding sites: cross regulatory interactions and feedback loops are common features in *Drosophila*. These loops often include the action of extracellular signalling molecules such as growth factors. AP-2 responds to protein kinase C and adenylate cyclase signal transduction pathways (Imagawa *et al.*, 1987) making it a candidate as a molecule causing changes in gene expression patterns in response to extracellular stimuli.

The bicoid-binding sequence in the *Msx*-1 5'-flanking region represents not the consensus site for wild type bicoid protein (TCTAATCCC) but a mutant variant generated as part of a detailed study of homeodomain-DNA interaction (mutant 5 -

Hanes and Brent, 1991). This mutant gave a high level of binding activity in a yeast co-transfection assay but with the added qualifier that the plasmid used contained a large number of such sites, possibly artificially inflating the level of transactivation observed. The mouse genome is known to encode at least one homeoprotein with a bicoid-like homeodomain. The *goosecoid* gene encodes a protein with the same helix 2 sequence in the helix-turn-helix motif (Blum *et al.*, 1992). *goosecoid* has a lysine residue at position 9 of the recognition-helix as does bicoid giving it a similar binding specificity (section 1.1.6). *goosecoid* is active during early gastrulation (6.4-6.8 dpc) and in mid-gestation (Blum *et al.*, 1992; Gaunt *et al.*, 1993) with the latter phase including expression in the branchial arches, the developing ear and the limb-buds, all sites of *Msx*-1 expression (section 1.4). *Msx*-1 expression is first seen around 7 dpc making it quite possible that either *goosecoid* protein persists and positively regulates *Msx*-1 or that its down regulation releases *Msx*-1 from *goosecoid*-mediated repression. Later, expression of *Msx*-1 is restricted to the distal limb whereas *goosecoid* is found more proximally (Gaunt *et al.*, 1993), suggestive of a possible negative interaction. While the sequence found was initially identified as a binding site for the bicoid protein it should be noted that all homeoproteins have a similar sequence specificity, i.e. the requirement for a TAAT core (section 1.1.6). As the suffix to the TAAT core has been changed in this mutant site it is quite possible that an alternative homeodomain interacts with this sequence, though none has yet been identified that has a preferred specificity to TAATCT (Treisman *et al.*, 1992a).

zeste is a *Drosophila* gene with a rather enigmatic status. It is known as a transcription factor acting upon a number of genes, including *Ubx,* upstream of which are specific binding sites for the protein. *zeste* is ubiquitously expressed throughout the *Drosophila* embryo and its status in the provision of region-specific transcriptional stimulation has been examined in this regard. A process known as transvection, in which regulatory elements act in *trans* on the second gene copy following homologous chromosome pairing, is known to involve *zeste*. Whether *zeste* acts as a site-specific transcription factor, upregulates by transvection or is a basal transcription activator is unclear. Null mutants in the *zeste* gene show no apparent phenotype and express *Ubx* in a normal manner. Nevertheless *zeste* has been shown to activate *Ubx* through the adjacent binding sites (Laney and Biggin, 1992). While transvection relies upon expression of *zeste* the absence of this process has no apparent ill effects on the fly (Goldberg *et al.*, 1989). It appears that *zeste* has a redundant action as a transcription factor (at least in the case of *Ubx* regulation) where the activity of other ubiquitous factors is capable of compensating for its loss. Dissection of the *Ubx* promoter (Laney and Biggin, 1992) and use of artificial transgenic constructs suggests that multiple

ubiquitous factors (*zeste*, GAGA and NTF-1) act redundantly, and also co-operate with specifically expressed factors to provide regional restriction to downstream genes. Selective distribution of binding sites for these factors may augment the region-specific regulation by spatially localised factors to increase the complexity of expression patterns that can be generated with a limited number of proteins.

Neither mouse nor human 5' flanking sequence contained *Msx*-1 binding sites Catron *et al.*, 1993), providing no evidence for a transcriptional autoregulatory loop operating to control *Msx*-1 expression.

The comparative sequencing approach has been adopted in a variety of systems, to study either the evolution of complex genomic regions, such as multigene clusters, or to exploit presumed evolutionary stability as a means of identifying functional domains within genes and their regulatory sequences. The most extensive comparative sequencing has been performed in the globin gene clusters and the gene complexes of the immune system, particularly the T-cell receptor locus (Shehee *et al.*, 1989; Hardison and Miller, 1993; Koop and Hood, 1994). These examples are of considerable value as the *cis*-regulation of these genes has been functionally characterised, enabling assessment of the level of conservation of *known* enhancers and promoter elements. The largest contiguous DNA comparison performed comprises approximately 100 kb of mouse and human sequence from the T-cell receptor locus (Koop and Hood, 1994). Unlike previous comparisons of the γ-crystallin and β-globin loci between human and rodent genomes (denDennen *et al.*, 1989; Shehee *et al.*, 1989) the T-cell receptor locus was found to be highly conserved throughout coding, non-coding *and* intergenic regions. The level of conservation in this complex genomic region suggests that maintaining organisation of the coding segments has had more selective advantage than maintenance of their specific sequences, as might be expected from our knowledge of T-cell receptor function and the mechanism by which it generates its extraordinary diversity. With regard to *cis*-regulatory signals, the previously defined 3' Cα and 5' Cδ enhancers were both highly conserved (95% and 80% respectively over limited core regions). The authors of this study conclude that "comparative sequence conservation may be a very powerful approach to identifying candidate regulatory regions", and have themselves detected potential *cis*-elements on the basis of their observations. Nuclear factors specific to B and T cells have been shown to bind at least one such sequence (Hood *et al.*, 1992). Work on the γ-globin gene has been performed using a similar strategy to that adopted here (Gumucio *et al.*, 1991). Nuclear factors were shown to bind to several conserved sequences found at least a kilobase upstream of the CAP site. Comparison of β-Globin genes, in this case

from human and rabbit, demonstrates the co-localisation of homologous sequence and regions with regulatory function (Hardison and Miller, 1993). The comparative study of most relevance to this work is one performed on a region of the *Hox*D homeobox-gene complex comparing mouse against human (Renucci *et al.*, 1992; section 1.2). This is the only published study of this kind on a homeobox gene. The region of the *Hox*D cluster examined extended from the 5' leader of the *Hox*d-10 (*Hox*-4.5) gene, through the two exons and intron of that gene, the intergenic region leading to *Hox*d-9 (*Hox*-4.4) and on through to the *Hox*d-9 transcription termination site. Outwith the coding regions, which are extremely highly conserved, there are four regions where similarity exceeds 75% over a region of 100 bases or more: the *Hox*d-9 intron (the *Hox*d-10 intron was not sequenced from human) and three discreet regions of varying size in the intergenic region. Detailed examination of the three intergenic regions shows that they comprise several blocks of sequence showing near identity between mouse and human. Transgenic analysis suggested that one of the conserved regions (region C) may have a role in delineating the anterior-posterior expression boundary. In the absence of this region, expression still appeared to be directed to the correct tissue types suggesting that one of the other regions provides tissue-specificity, such as expression in mesodermal derivatives, though this hypothesis was not tested. It was suggested that region B (the region containing the *Msx*-1 5'-flanking homology) generated this specificity as region A comprises a series of functional homeodomain-binding sites more likely involved in cross regulatory interactions between homeobox genes (Zappavigna *et al.*, 1991). *In vitro* protein binding studies were performed using these homologous regions and these will be discussed in chapter 4.

The presence of a markedly conserved sequence in the 3'-UTR of *Msx*-1 may be significant in indicating a regulatory function for this region. This conservation has previously been noted as part of a detailed study examining the incidence of highly conserved regions (HCRs: defined as regions of ≥100nt with ≥70% similarity between cognate genes of species diverged by at least 300 million years) in non-coding sections of entries to the Genbank database (Duret *et al.*, 1993). Duret and colleagues showed that such features are not uncommon, occurring in approximately 30% of genes examined, and with a bias for genes encoding DNA-binding or cytoskeletal proteins (about 10-fold more frequent than in genes encoding hormones, hormone receptors and enzymes). HCRs were predominantly found in the 3'-UTR, rather than the 5'-UTR or intron(s), and in the transcribed portion of the gene indicating a requirement for their presence in the mature message, likely reflecting a post-transcriptional regulatory role. Such regions of conservation may function in mRNA localisation, regulation of translation, control of transcript stability or yet undiscovered levels of post-

transcriptional regulation. The presence of this type of sequence in the *Msx*-1 gene is not directly informative but it may hint at important levels of control and suggests a direction for future studies on distribution and regulation of *Msx*-1 protein synthesis.

The best candidate for potential *cis*-regulatory function is the region highly conserved between mouse and human (between positions -550 and -507 from the transcription start site in mouse), and similar to the *Hox*D sequence. This sequence has been conserved at 38 of 44 positions (86 %) over a period of approximately 80 million years since the mammalian radiation. The conservation of this sequence and its similar position with relation to the coding region in both mouse and human (section 3.3.3) leads to the conclusion that this represents an island of conservation in a region of the gene that has diverged under non-selected drift.

Interestingly the two conserved sequences, from *Msx*-1 and *Hox*d-9, both extend 3' for 43 bases from the GAGA sequence. The *Hox*d-9 sequence is part of a block of homology that extends further 5' while the GAGA represents the 5' limit of the *Msx*-1 homology. Comparison reveals that although the two sequences are very similar in the 5' half (for ~21 bases) they diverge as we move 3'. Such a 'bipartite' structure may reflect the presence of more than one protein-binding site in the whole 43 bases and that only the 5' most site is conserved between the two genes. Alternatively the 3' half may not have such a strict level of similarity due to a requirement rather for a general base content than a specific pattern of nucleotides. Such "complex and degenerate functional requirements" have been proposed as a possible explanation for the apparent lack of conservation of certain previously characterised regulatory elements in the large scale study on the T-cell receptor discussed above (Koop and Hood, 1994). The 3' portion of the *Msx*-1 conserved sequence has two polyT stretches that are largely conserved in human *Msx*-1 but are absent in the *Hox*D sequence. Recent work on the function of *Drosophila* transcription factor *dorsal* has shown that this single protein is capable of positively *or* negatively regulating target genes dependent upon the alternative presence of two possible binding sites (Kirov *et al.*, 1993; Jiang *et al.*, 1993). The dorsal-dependent enhancer element consists of a single protein binding site specific for the *dorsal* protein. A dorsal-dependent silencer however comprises a similar *dorsal*-binding site adjacent to an AT-rich sequence that has been proposed to bind a co-repressor responsible for modification of the *dorsal* activation properties. It is possible that the stretches of T in the 3' half of the *Msx*-1 conserved sequence reflect a similar mechanism, with potential significance for the absence of the T's in the *Hox*D sequence. The interaction of a tissue-specific regulator (dorsal), general factors (co-repressors) and selective use of alternative DNA elements provides added levels of complexity without using additional

genes. Nothing is known regarding the mechanisms whereby the two proteins interact or cause suppression. In a similar way the silencing activity of the yeast mating-type-specific protein α2 is dependent upon an interaction with a general factor MCM1 (section 1.1.6). The correct recognition of a *cis*-element in this case is dependent upon the α2-MCM1 interaction.

The sequence homology to the upstream region of *Hox*d-9 is of particular interest in relation to gene expression in the developing limb. A considerable body of work has described the patterns, morphological consequences and regulating influences of *Hox*D gene expression in the vertebrate limb (Dollé *et al.*, 1989; Izpisúa-Belmonte *et al.*, 1991; Morgan *et al.*, 1992; Dollé *et al.*, 1993b; Niswander *et al.*, 1993b). The 5 *Abd*-B cognate genes at the 5' end of the *Hox*D complex (figure 1.1) provide early patterning information in the limb bud (from here on '5' *Hox*D genes' refers to these 5 genes). Of these genes, *Hox*d-9 is the most 3' and has the broadest expression pattern (section 1.3.4). *Hox*d-9 is expressed throughout the early limb bud while its 5' neighbours are increasingly proximo-distally restricted (Dollé *et al.*, 1989). All other 5' *Hox*D genes (*Hox*-d10-13) are expressed within the *Hox*d-9 domain. *Msx*-1 is expressed in the limb bud at a similar stage of development (10-10.5 dpc) in a distally restricted domain with a posterior bias (Hill *et al.*, 1989). Large regions of the early limb bud mesenchyme therefore express both genes. While the homology to the *Msx*-1 conserved sequence is upstream of *Hox*d-9 it may exert an influence on genes throughout the *Hox*D cluster. The long range function of enhancer sequences is one mechanism proposed for maintenance of cluster integrity, supported by results from transgenic studies that suggest sharing of *cis* elements by more than one gene (Sham *et al.*, 1992; Eid *et al.*, 1993). If the *Msx*-1/*Hox*D conserved sequence (section 3.3.4) acts as a functional *cis* element it may regulate all the 5' *Hox*D genes. The position of this sequence upstream of *Hox*d-9 may be significant in this respect as this is the first of the 5' *Hox*D genes expressed, according to the temporal colinearity rule. Models predicting an 'opening' of the complex to account for colinearity might be expected to require an 'open' status for any global regulatory sequences, thereby necessitating their position early in the sequence. By predicting that a common *cis* element regulates *Msx*-1 and 5' *Hox*D genes we might expect both to respond to similar signals. Work on limb patterning has defined a role for the Apical Ectodermal Ridge (AER) in specification of positional identity (Saunders, 1948; Summerbell, 1974). Combining surgical manipulations with gene expression studies has demonstrated a relationship between the AER and expression of *Msx*-1 and *Hox*D genes in the underlying mesenchyme (Robert *et al.* 1991; Coelho *et al.*, 1993a; Izpisúa-Belmonte *et al.*, 1992). In brief, expression of both *Msx*-1 and 5' *Hox*D genes is dependent upon an intact AER. It is

possible then that *Msx*-1 and 5' *Hox*D genes respond to the same inductive/maintenance signal emanating from the AER. Recent work has gone some way to defining the molecular nature of AER function with the demonstration that fibroblast growth factor 4 (FGF-4) can largely substitute for the AER (Niswander *et al.*, 1993). A factor binding the *Msx*-1/*Hox*D homologous sequence might respond to the signal transduction pathway activated by FGF-4 and thereby induce or stabilise expression of those genes in limb mesenchymal cells.

The *Msx*-1 conserved sequence does not include any lengthy or complex consensus binding sites as determined by comparison to TFD. As described the 5' end of this sequence, the region most widely conserved, includes a GAGA motif. Sequences consisting of alternating purines have been shown to bind proteins from a variety of sources (Gilmour *et al.*, 1989; Kennedy and Rutter, 1992). In *Drosophila* a protein has been characterised, and recently cloned, that binds to the GAGAG sequence found in the proximal promoter regions of several genes, among them *Ultrabithorax*, *fushi tarazu*, *even-skipped*, *Krüppel*, *hsp70* and *hsp26* (Soeller *et al.*, 1993). This so-called GAGA factor is implicated in transcriptional activation by antirepression of the non-specific negative effects upon transcription of histone H1 (Croston *et al.*, 1991). GAGA protein has also been shown to enhance activation and nuclease sensitivity of the *hsp26* promoter, when fused to a reporter gene (Lu *et al.*, 1993). It has been proposed that GAGA factor has a role in the maintenance of promoter accessibility and absence of nucleosome suppression of transcription. In support of this hypothesis, recent work has demonstrated that GAGA factor is involved in an energy-dependent disruption of nucleosome order at the *Drosophila* hsp70 promoter (Tsukiyama *et al.*, 1994). In this system GAGA factor is essential to the correct assembly of the transcription initiation complex at the promoter by means of realigning adjacent nucleosomes. It has been suggested that GAGA is one of an alternative class of transcription regulating molecules whose role is to prepare promoters by rearranging the chromatin structure that surrounds them (van Holde, 1994). A study by Lee *et al.* (1992) demonstrated a role for GAGA in assembly of a paused transcription complex at the *hsp70* promoter of *Drosophila*. The polymerase becomes engaged and then pauses in elongation after approximately 25 nucleotides thereby setting up a system capable of extremely rapid response to transcriptional inducing signals, in this case those associated with heat-shock response. Both *Msx*-1 and the *Hox*D genes respond to the inductive influence of the AER (and in the case of *Msx*-1 several other epithelial components), it is possible that such induction could occur by release of a stalled transcription of this type.

Chapter 4

In vitro binding studies

4 Introduction

Given the complexity of the mammalian genome and the cellular environment that it occupies, characterisation of individual molecular interactions in a cellular system is, in the majority of cases, impractical. To surmount the problems such interactions are first studied *in vitro*, in a controlled, low complexity system. While findings from such experimental work should not be casually extrapolated to the *in vivo* situation, many fundamental properties of the molecules examined can be accurately assessed by this approach. Characterisation of the protein-DNA interactions that effect transcriptional regulation upon a gene is a typical example of such a situation. The wide spectrum of nuclear proteins and the large, complex nature of the genome serve to obscure attempts to identify interactions occurring between a sequence element, adjacent to a particular gene, and an individual protein or small number of proteins. By using an *in vitro* approach the complexity of at least one of the two components (protein and DNA) can be greatly reduced.

This chapter describes the use of two key *in vitro* techniques for studying sequence-specific DNA-protein interactions, the gel retardation assay and South-western blotting, in an attempt to assess the significance of the conserved sequence identified by comparing the mouse and human *Msx*-1 5'-flanking region (section 3.3.3). Crude nuclear extracts were used in the presence of short, defined DNA sequences enabling assessment of any protein-binding by these sequences in the broadest possible context. HeLa cell extract is the best characterised cell-free transcription system, containing components of both basal (e.g. Usuda *et al.*, 1991) and regulated (e.g. Vasseur-Cognet & Lane, 1993; Virbasius *et al.*, 1993) RNA polymerase II transcription machinery. The HeLa cell line is a human epitheloid carcinoma (cervical adenocarcinoma) derived line (Gey *et al.*, 1952). It has been used as a source in the purification of several well-known transcription factors such as Sp1, CAAT-binding proteins and Oct-1 (e.g. Kadonaga and Tjian, 1986; Fletcher *et al.*, 1987). However, these factors are all ubiquitous in their expression patterns and in the case of the Oct family (Schaffner, 1989), additional factors have been found that show cell-type-specific expression and are not expressed in HeLa cells (namely Oct-2, a lymphoid-specific factor). For these reasons HeLa extract was used along with extracts from cell lines with alternative lineages (B16 and H3M). The B16 cell line (more accurately termed B16C3 but from here on referred to as B16) is derived from a spontaneous melanoma in C57BL/6 mouse (Bennett, 1983). As a melanoma cell line, B16 represents a member of the neural crest lineage. *Msx*-1 is expressed in cells and tissues derived from the neural crest (section 1.4) and consequently it was felt that programs

of gene expression activated in this lineage may be represented by factors present in B16 cells. Novel transcription factors have previously been isolated from this cell line (Tagawa *et al.*, 1990). Also this cell line was being successfully used in transcriptional studies by another group in the host institute (P. Budd & I. Jackson, pers. comm.). H3M derives from embryonic neural tube, including cells of the neural crest and for similar reasons was considered a good candidate as a cell line expressing regulators of *Msx*-1. In an attempt to relate the findings of these *in vitro* experiments to the situation in the embryo nuclear extracts were also prepared from dissected portions of 11 dpc mouse embryos.

4.1 Gel-retardation assays

The Gel-Retardation assay (a.k.a. Electrophoretic mobility-shift assay - EMSA, or Bandshift assay) is a convenient and flexible way to study DNA-protein interactions (see Lane *et al.*, 1992 for review). The assay exploits the observation that migration of a length of DNA through a non-denaturing polyacrylamide gel is retarded when the DNA is present in the form of a DNA-protein complex (Fried and Crothers, 1981; Garner and Revzin, 1981). Protein is added to the DNA probe, which is radiolabelled, and the mixture is applied to a native polyacrylamide gel. Uncomplexed DNA migrates through the gel faster than DNA bound by protein thereby resolving the two populations of DNA which are detected by autoradiography. The ability to resolve complexed and uncomplexed DNA is dependent upon a variety of factors, briefly discussed below.

The electrophoretic properties of DNA are well studied; equation a) describes, at a first approximation, the migration of DNA through a polyacrylamide gel (Lumpkin and Zimm, 1982).

a) $$v = h^2 \times Q \times E / (L^2 \times f)$$

where -
- v = velocity of migration
- h = end-end distance of DNA molecule
- Q = effective charge
- E = electric field
- L = contour length
- f = frictional coefficient

We can see that several factors influence DNA electrophoresis through polyacrylamide. Movement of DNA through the gel matrix has been assumed to occur in a worm-like fashion. A conformational change in the molecule, such as a bend, slows down this movement; the degree of retardation is dependent upon the position of the bend along the molecule and the degree of bend, hence the significance of the spatial distance between the two ends of the DNA (coefficient *h*; Marini *et al.*, 1982).

Binding of a protein to the DNA increases the mass of the complex thereby decreasing its relative mobility (observed mobility÷expected mobility). Many DNA-binding proteins also produce a conformational change in the DNA adding to the retardation effect. It has been shown that resolution of complexed DNA is dependent not upon the absolute mass of the protein/DNA complex but upon the ratio of the masses of the two components of the complex. This is demonstrated by increasing the mass of the protein, which reduces the relative mobility of the complex, or increasing the mass of the DNA which has the opposite effect (Lane, 1992). However, particularly when using crude extracts, potentially containing a variety of proteins capable of binding the sequence under scrutiny, the relative mobility of various complexes cannot be used as an indication of the size of the proteins bound due to additional factors such as protein charge, conformation and the possibility of multiple proteins interacting with the DNA that may also affect movement through the gel (Fried, 1989).

In order that *specific* interactions are analysed using this assay it is essential to suppress the large amount of non-specific DNA-binding activity present in protein extracts. This is of particular importance when using crude cellular extracts as a protein source. Non-specific binding of protein to DNA can be suppressed in two ways. Firstly, the presence of salt in the binding buffer disrupts non-specific electrostatic interactions; the higher the salt concentration, the more stringent the conditions. Secondly, non-specific 'competitor' DNA is included in the binding mixture. The use of fragmented *E.coli* or salmon sperm DNA as competitor has now largely been superseded by non-complex, synthetic copolymers such as poly d(I-C). Such copolymers have the advantage that they include no fortuitous recognition sites that might titrate the proteins binding the labelled probe.

Gel concentration is important as this determines pore size within the matrix. Polyacrylamide gels have a sufficiently small pore size to resolve complexes and are favoured. Agarose gels have considerably larger pores and are neither able to discriminate on a conformational basis nor to resolve complexes unless the ratio of complex mass to free DNA mass is very large. DNA protein complexes appear more stable during electrophoresis than free solution kinetics suggest. This has been put

down to the 'caging' effect caused by the structure of the gel which prevents diffusion, maintaining a high local concentration of protein and DNA within the small 'cell' of the matrix (Fried and Crothers, 1981). It was shown that the effective salt concentration drops sharply as the complex enters the gel; this could also explain the gel-dependent stabilisation of the interactions.

In many cases the gel-retardation assay is used to study the binding of purified or cloned proteins. In the case of the latter, it can be used in conjunction with various truncated proteins as an assay to localise DNA-binding function within the peptide (Ramsay *et al.*, 1991; Ha *et al.*, 1993). Alternatively this assay can be used in conjunction with crude nuclear extracts (Strauss and Varshavsky, 1984). The use of crude extracts broadens the scope of the assay, enabling its use to determine additional data. Extracts can be produced from cell lines of different status, for example embryonal carcinoma cells and differentiated cells (Flamant *et al.*, 1987) or from cells in a synchronised culture at different stages of the cell cycle to look for stage-specific binding activity (Walsh *et al.*, 1992). Protein-DNA interactions involving higher order complexes composed of more than a single protein can assemble in a crude extract, including both oligomeric binding-proteins and situations where a stretch of DNA has binding sites for more than one protein (Suzuki and Suzuki, 1988).

The gel-retardation assay was employed in this study as the most convenient method for determining whether regions of *Msx*-1 5'-flanking DNA, shown to have homology between mouse and human (section 3.3.3), bound proteins in a sequence-specific manner. The size of the region of homology (44 bp) lends itself to the use of synthetic oligonucleotides as probes in the assay. The use of oligos enables a plentiful supply of DNA of constant quality. Oligos were purified by polyacrylamide-gel electrophoresis followed by elution to eliminate incomplete synthesis products (short oligos) and substances remaining from the synthesis that might affect enzymatic manipulations, such as radiolabelling or ligation (section 2.8.1). The oligos used were synthesised in complementary pairs and following purification these were annealed. Table 4.1 shows the sequence of the oligos used. Complementary oligos were synthesised representing both the mouse and human regions of *Msx*-1 5'-flanking homology, the mouse *Hox*d-9 sequence and an Octamer site used as a control.

Table 4.1

Code	Gene/Species	Sequence
D13	*Msx*-1/mouse (top)	5'-ATTAACTTTGTCCAGCCCTGGAGACAAAGGCCCATTTTTACTCCGAGGTAATTTTTGAAGATT-3'
D12	*Msx*-1/mouse (bottom)	5'-AATCTTCAAAAATTACCTCGGAGTAAAAATGGGCCTTTGTCTCCAGGGCTGGACAAAGTTAAT-3'
C611	*Msx*-1/human (top)	5'-TTTACTCCAGCTCTAAGTTAGAGACAAAGGCCCACTTTTACCTCGAGGTAAAGTTTACAAGATTT-3'
C679	*Msx*-1/human (bottom)	5'-AAATCTTGTAAACTTTACCTCGAGGTAAAAGTGGGCCTTTGTCTCTAACTTAGAGCTGGAGTAAA-3'
D14	*Hox*d-9/mouse (top)	5'-GGCAGCCTTGAATCTGAGAGAAATGCCCATTGTCACTCCCAAATCC-3'
C617	*Hox*d-9/mouse (bottom)	5'-GGATTTGGGAGTGACAATGGGCATTTCTCTCAGATTCAAGGCTGCC-3'
B349	Octamer (top)	5'-GTGAGCGAGAGGAATTTGCATTTCCACCGACCTTCCGC-3'
B351	Octamer (bottom)	5'-TGTAGCGGAAGGTCGGTGGAAATGCAAATTCCTCTCGC-3'

Table 4.1 Sequences of the six oligos used as gel retardation probes and in the construction of concatenated South-western probes. They form complementary pairs, with the 'top' strands of the first six corresponding to the homologous sequences shown in the lineup of figure 3.24.

4.1.1 Conserved sequences bind similar proteins

To determine whether the conserved sequence identified (section 3.3.3) binds nuclear proteins in a sequence-specific manner the gel-retardation assay was employed using protein extracts prepared by hypotonic lysis of cells and high salt extraction of the nuclei (Andrews and Faller, 1991; section 2.7.2). Binding conditions used were as follows - 0.25 mM DTT, 2 mM $MgCl_2$, 2.5% Glycerol, 4 mM spermidine, 100 mM KCl, 100 mM NaCl, 10 mM N-[2-Hydroxyethyl]piperazine-N'-[2-ethanesulphonic acid] (HEPES) pH 7.9 plus poly d(I-C) to a final concentration of 150 ng/µl (3µg per binding reaction). This buffer was used in the initial experiments and later titration of the salt content (not shown) showed that the concentrations used were close to the upper level at which binding was observed, i.e. providing the maximum suppression of non-specific interactions (see above).

Figure 4.1 shows a retardation assay performed using both mouse and human probes (mouse probe = oligos D12/D13; human probe = oligos C611/C679; table 4.1 for oligo sequence; section 2.8.1 for probe construction). Binding assays with these probes were performed in the presence of HeLa (Promega, HeLaScribe™) and B16 nuclear extracts (section 2.7.2). Lanes 2 and 3 show binding of the mouse and human probe, respectively, to the HeLa cell extract . Lanes 4 and 5 show binding of the mouse and human probe (respectively) to the B16 cell extract. Clearly the two probes have very similar protein-binding properties. Differences are, however, apparent in the binding patterns generated by the two protein sources used. The B16 extract gives two clearly retarded bands of differing mobility. DNA-protein complexes of a similar mobility are formed with this extract by both the mouse and human *Msx*-1 probes. The HeLa extract also shows an identical pattern of complexes with both mouse and human *Msx*-1 probes. Two major complexes are seen, one of identical size to the smaller B16 complex, the second of lower mobility than the second B16 complex. A faint signal is seen at a similar position to that of the upper B16 band suggesting a low level this complex in HeLa cells. It is possible that the upper HeLa band represents a higher order complex, specific to HeLa cells, but comprising in part similar components to the upper B16 band.

Figure 4.2 shows gel-retardation of Octamer oligos B349/B351 by the HeLa extract (Table 4.1; the octamer core sequence is ATTTGCAT). This sequence was shown to generate a single retarded band that was of different (lower) mobility to any of those in Figure 4.1 (relative mobility of the octamer band is 0.075 compared with 0.164 for the upper band in figure 4.1, lanes 2 & 3). In addition the octamer did not

Figure 4.1

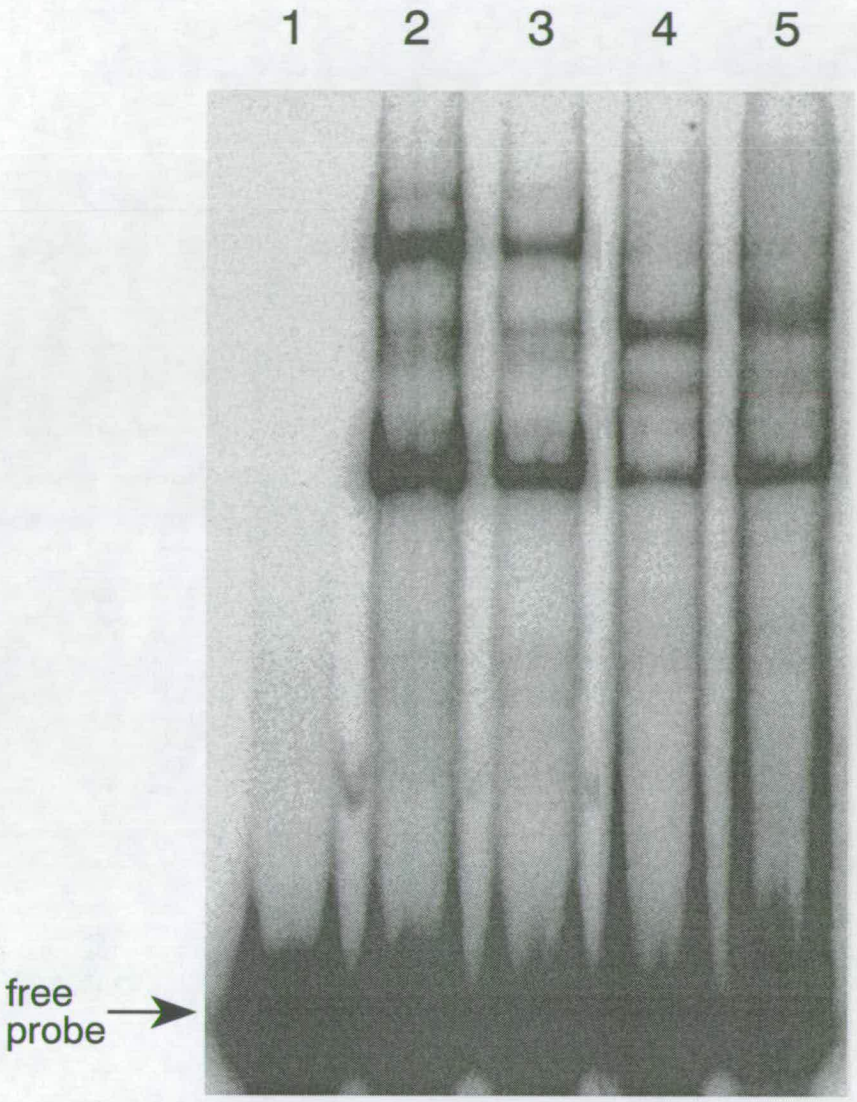


Figure 4.1 Gel-retardation assay showing identical activity by both mouse and human conserved sequences. Lane 1 shows the mouse probe in the absence of protein. Lanes 2 and 3 show HeLa protein (6.8 µg/lane) binding mouse and human probes. Lanes 4 and 5 show B16 (4.2 µg/lane) extract binding mouse and human probes. Unbound probe is arrowed at the bottom of the gel; bands above this represent probe retarded by complex formation with protein. In addition to the protein each lane contains 1X binding buffer (section 4.1.1), 2.5% glycerol, ~10^5 cpm of probe and 3 µg poly d(I-C) in a total of 20 µl.

Figure 4.2

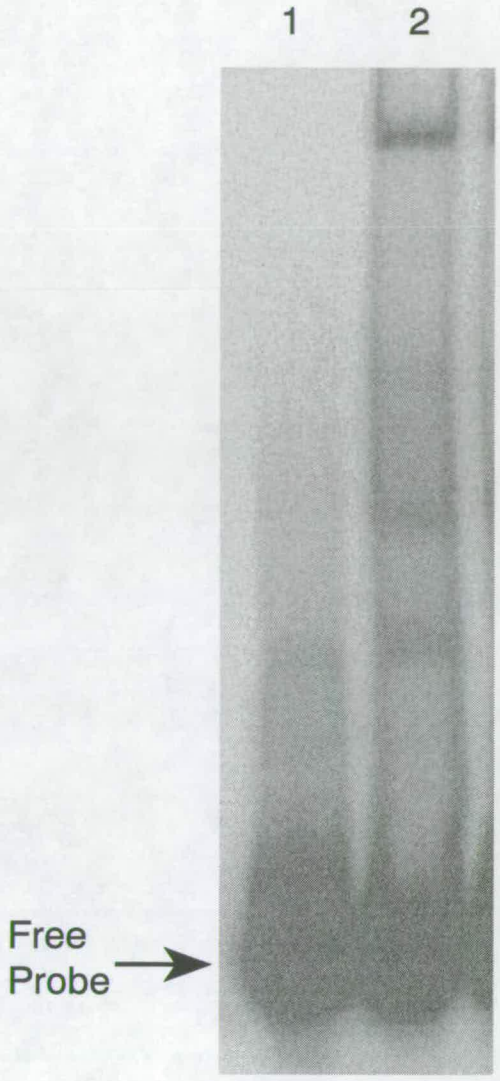


Figure 4.2 Gel-retardation showing HeLa cell extract bound to Octamer probe (oligos B349/B351). Lane 1 contains no protein, Lane 2 contains 6.8µg of HeLa protein. Binding reactions were in the buffer described using ~10^5 cpm of probe.

compete with the *Msx*-1 sequence for complex formation (not shown). From this I conclude that the complexes seen in figure 4.1 are sequence-specific.

4.1.2 Conserved sequences compete for binding of mouse proteins

Binding assays performed for analysis by gel-retardation can be assessed for their specificity by inclusion into the reaction mixture of additional, unlabelled oligonucleotides potentially capable of competing with the labelled, probe oligo for protein binding. If the unlabelled 'competitor' binds the same proteins as the labelled probe oligo then the retarded protein-DNA complex will be titrated out and disappear as an autoradiograph signal. Oligos used as competitor can be identical to the probe, confirming the interaction observed, or can have a different sequence in order to investigate the relationship between proteins binding diverse sequences.

Figure 4.3 shows competition of the complexes formed between the mouse *Msx*-1 conserved element and the B16 extract. Unlabelled, double-stranded oligonucleotides corresponding to this mouse sequence and to the mouse *Hox*D sequence are shown to compete these complexes. Oligonucleotides used as competitors were gel purified prior to annealing (section 2.8.1) as for the probe oligos. It was found that purified oligos made more potent competitors than non-purified. Competition similar to that seen in figure 4.3 was achieved with unpurified oligos only when used in at least 200-fold molar excess. There is possibly a lower effective concentration of oligonucleotide in the unpurified sample as a result of a high proportion of partial synthesis products thereby reducing the actual amount of full length double-stranded oligo generated upon annealing.

Lane 1 shows complex formation between proteins in the B16 extract and the mouse *Msx*-1 conserved sequence. This is the same as a similar binding reaction shown in figure 4.1 (lane 4), though the bands are resolved with slightly less clarity here. Lane 3 shows an identical binding reaction to which was added a 50-fold molar excess of unlabelled purified double-stranded oligonucleotide identical to that used as probe (D12/D13). There is a clear reduction in the level of the DNA-protein complexes, which is interpreted as competition by the unlabelled oligos for the proteins bound in the retarded complex bands. Similar competition is also seen with a 10-fold molar excess of unlabelled competitor (lane 4), suggesting that the concentration of the protein (or proteins) binding the radioactive probe in this complex is strictly limiting. Lane 2 shows the resolution of a binding reaction, identical to that in lane 1, in which unlabelled purified double-stranded DNA corresponding to the *Hox*D sequence (oligos D14/C617) was added as a competitor to 50-fold molar excess. Again there is

Figure 4.3

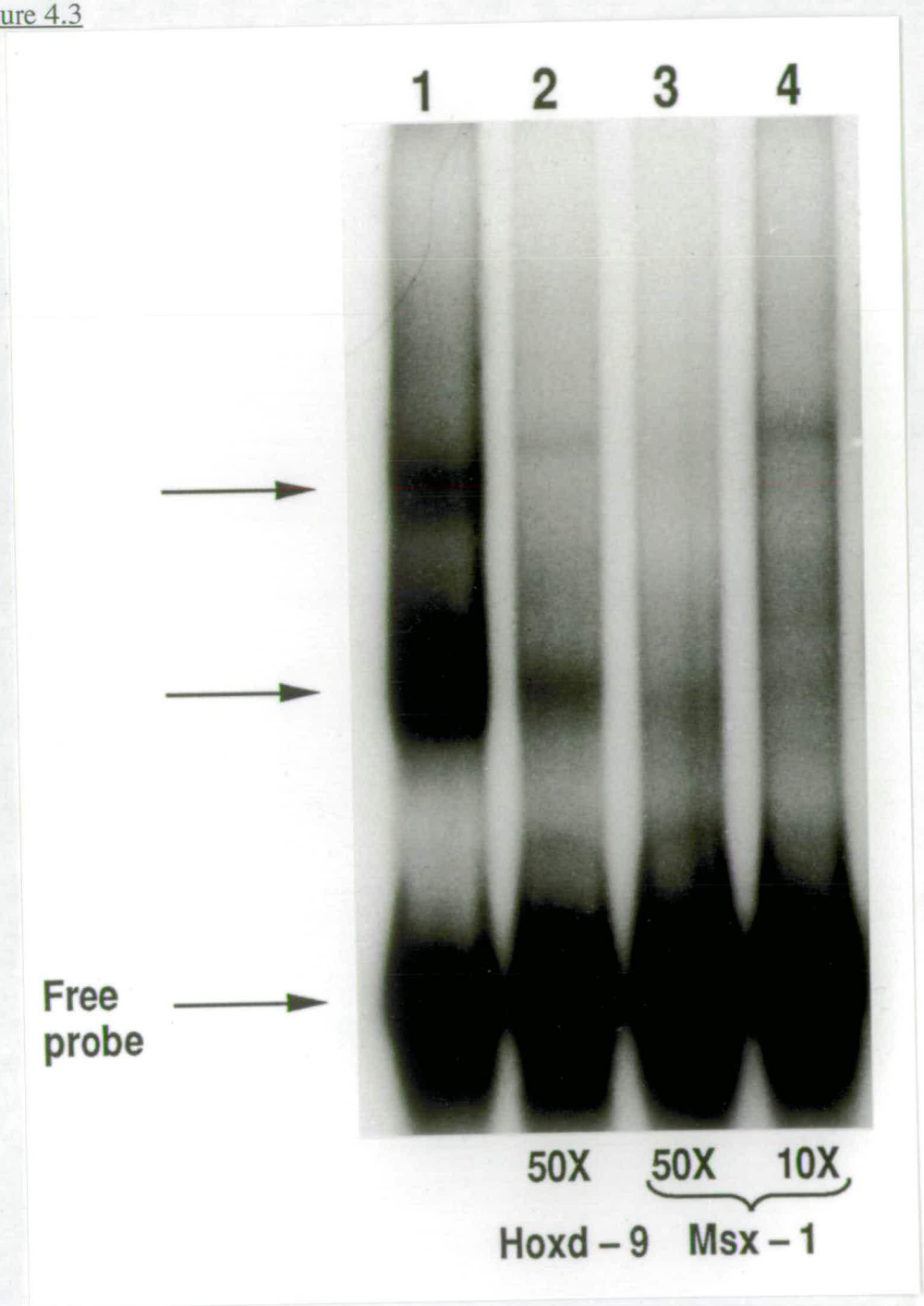


Figure 4.3 Gel-retardation assay showing use of various unlabelled competitor oligos. Each lane contains 1X binding buffer (section 4.1.1), 2.5% glycerol, 3µg poly d(I-C), radiolabelled probe, the described molar ratio of unlabelled competitor and 4.2µg of B16 extract, in a total volume of 20µl. Lane 1 shows an uncompeted binding assay between B16 extract and the mouse *Msx*-1 conserved sequence (D12/D13). Lane 2 shows an identical binding reaction competed by inclusion of 50-fold unlabelled *Hox*D sequence (D14/C617). Lane 3 shows an identical binding reaction with inclusion of 50-fold unlabelled probe (mouse *Msx*-1 sequence: D12/D13). Lane 4 shows an identical binding reaction with inclusion of 10-fold unlabelled probe (mouse *Msx*-1 sequence: D12/D13).

competition of the complexes formed (a reduction in their intensity) but to a lower level than in lane 3, when the probe itself was used unlabelled. Both the mouse *Msx*-1 sequence and the *Hox*D sequence are capable of competing with the mouse *Msx*-1 probe for the binding of proteins in specific DNA-protein complexes. Self competition by the probe sequence itself is expected. Competition by the HoxD sequence demonstrates that the two related sequences form complexes with the same protein components of the nuclear extract.

4.2 South-Western blotting

South-Western blotting is a method developed to detect a specific DNA-binding protein within a heterogeneous protein mix, such as a crude nuclear extract (Bowen *et al.*, 1980; Miskimins *et al.*, 1985). As its name implies it is a hybrid technique using principles from Southern and Western blotting enabling the characterisation of DNA-protein interactions on a solid membrane support.

SDS denaturation followed by polyacrylamide-gel electrophoresis enables separation of individual proteins according to size (Laemmli, 1970; section 2.9.2). These proteins are transferred onto a nitrocellulose membrane by migration toward the anode of an electric current, passed across a sandwich of the gel and the membrane (Towbin, 1979). South-western blotting involves the incubation, under optimal binding conditions, of the protein-laden filter with double-stranded DNA corresponding to a putative binding-site and radiolabelled to high specific-activity. As for the gel-retardation assay, reagents are used to block non-specific interactions: in this case these are non-fat dried milk and poly-d(I-C) DNA (section 2.8.6). Conditions for DNA interaction with filter-bound protein have been optimised by two key modifications of the original protocol: repeated cycles of *in situ* denaturation (by guanidine hydrogen chloride) and renaturation of the proteins, and the use of concatenated DNA sites as probes (Vinson *et al.*, 1988). Denaturation/renaturation is thought to facilitate the adoption of a correct conformation by the DNA-binding domain, as the denatured protein refolds, and was shown to increase the efficiency of filter-bound protein-DNA interactions. The use of concatenated copies of the DNA site decreases the dissociation rate constant thereby enhancing detection of membrane bound proteins. Such concatenated probes can also be labelled to very high specific-activity by nick-translation, improving detection. After washing the filter has removed uncomplexed probe, proteins that specifically bind the labelled probe DNA will show up as bands on an autoradiograph, corresponding to the protein bands on the SDS-PAGE gel.

The technique has been widely used in the study of DNA-binding proteins, in particular transcription factors (Matsuo *et al.*, 1991; Benyajati *et al.*, 1992; Dikstein *et al.*, 1992; West *et al.*, 1992; Lenormand *et al.*, 1993). It provides quantitative information on the size of proteins bound by a particular sequence and it can provide qualitative information regarding the presence of such proteins in extracts from varying sources.

South-western blotting was used in this study to provide information on the nature of the protein (or proteins) binding the conserved DNA sequence identified, to compare the proteins bound by related sequences and to examine the distribution of binding proteins by using extracts from different regions of the embryo.

Extracts shown to give DNA-binding activity by the gel-retardation assay, along with additional cellular and embryonic extracts, were examined by South-western blotting and DNA-binding proteins within them were identified as single bands with a measurable molecular weight. Various reagents designed to suppress non-specific hybridisation were tested, including gelatin, dried milk and bovine serum albumin. Dried milk gave the best results and it was used in the generation of all blots presented here (section 2.8.6).

4.2.1 Mouse and Human sequences bind identically-sized proteins

South-western blots were performed to assess whether the related sequences identified upstream of mouse and human *Msx*-1 genes bind proteins with similar properties, as gel-retardation assays had suggested (section 4.1.1). Figure 4.4 shows a composite of two identical filters; one was probed with concatenated complementary oligonucleotides D12 and D13 (section 2.9.1), representing the mouse sequence, and the other with probe derived from oligonucleotides C611 and C679, representing the human sequence (see table 4.1 for oligonucleotide sequences). Details of the filter processing and hybridisation conditions are given in section 2.9.4 and these apply for all blots shown.

All lanes show a hybridisation signal to a large polypeptide of molecular mass (M_r) approximately 127kD (large arrow figure 4.4; see section 4.2.2 for determination of molecular weight). In addition, all lanes show a much less distinct signal at approximately 31kD (small arrow figure 4.4). In the case of the B16 cell extract this second signal resolves into two indistinct bands; the HeLa extract does not demonstrate two clear bands though the weakness of the signal does not preclude such a possibility. Variation between the two protein sources is seen in the presence of an

Figure 4.4

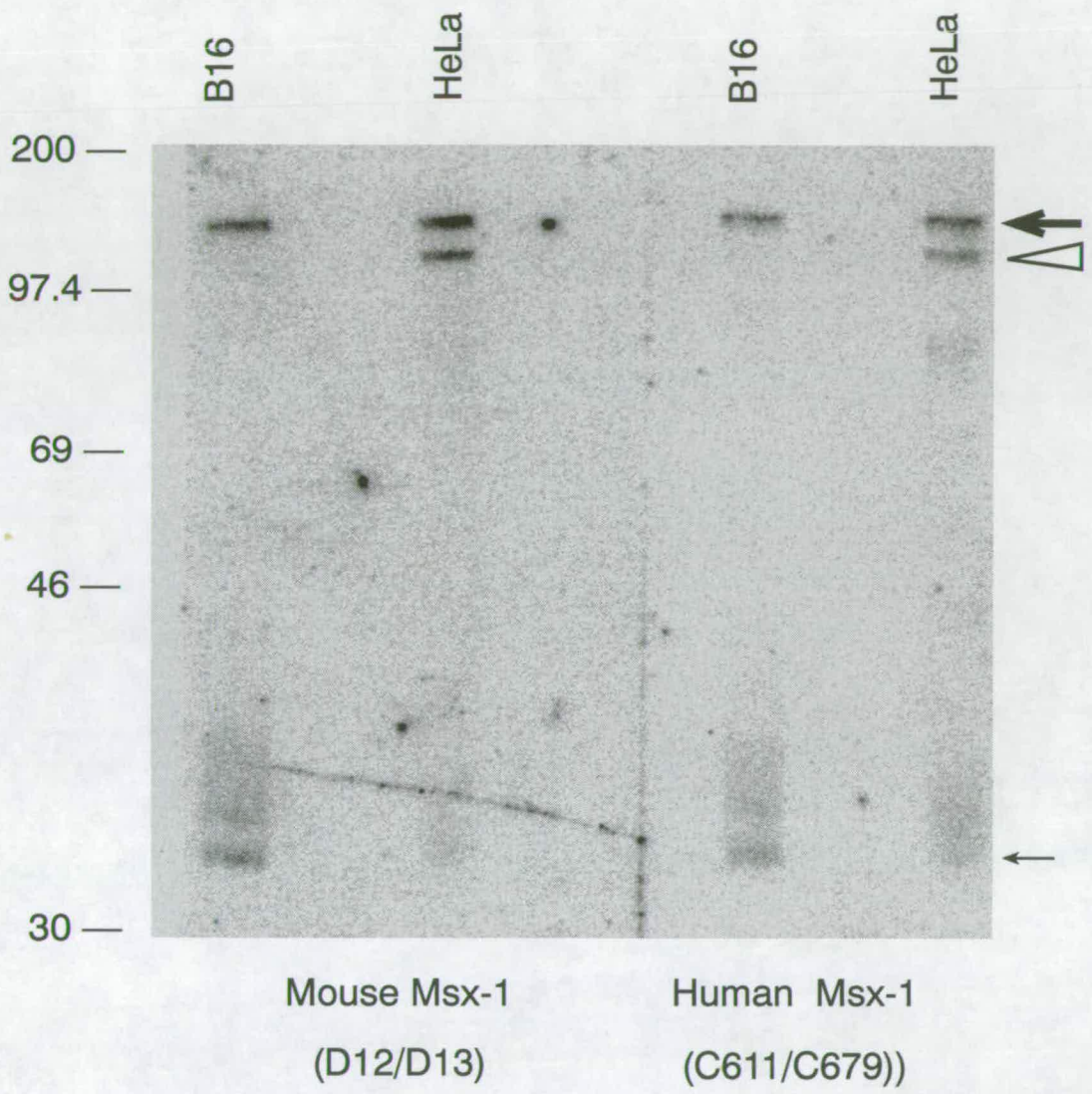


Figure 4.4 South-western blot showing similar proteins bound by both mouse and human *Msx*-1 sequences. Mouse and Human sequences are shown binding to proteins from B16 melanoma cell line (mouse) and HeLa cell line (human). Lanes 1 & 3 loaded with 8µg B16 extract (section 2.8.3), lanes 2 & 4 loaded with 10µg HeLa extract (Promega). kiloDalton sizes on the left are derived from rainbow molecular weight markers (Amersham) run on the SDS-PAGE gel.

additional signal, at approximately 108kD (section 4.2.2), in the HeLa cell extract. A coomassie-blue stained SDS-PAGE gel, loaded with identical samples, did not show major proteins at the position of the peptides detected by hybridisation so these bands are unlikely to represent non-specific binding to highly abundant proteins. Background hybridisation is low, suggesting adequate competition of non-specific hybridisations by the poly-d(I-C) competitor. The large molecular mass signals are clear, strong and even in their intensity supporting the view that they represent a specific interaction between probe and protein.

4.2.2 Similar DNA-binding proteins are found in several cell types

Figure 4.5 shows the result of a South-western blot performed using protein extracts from several cell types probed with the mouse conserved sequence (oligos D12/D13). HeLa cell and B16 cell extracts are present along with protein from H3M cell line, ES (embryonic stem) cells and F9 mouse teratocarcinoma cells. Details of the gel loading are presented in the figure legend. The mouse H3M cell line is an immortalised cell line derived from an explant of neural tube from a 9dpc mouse embryo (a kind gift of S. Clay); Embryonic stem cells are cultured cells derived from the early embryo that maintain their pluripotent state (Martin, 1981); F9 cells are a line of mouse teratocarcinoma cells - an undifferentiated cell type cultured from malignant embryonic tumours (Bernstine *et al.*, 1973).

HeLa and B16 extracts show results comparable to those presented in section 4.2.1; in addition the H3M, ES and F9 cells all show binding activity similar to the B16 cells. HeLa shows the 127kD and 108kD signals shown in figure 4.4 though no signal of 31kD is detectable. B16 shows a signal at 127kD but nothing at 108kD: there is an indistinct signal at 31kD but rather like the HeLa cell extract of figure 4.4 it has not resolved into clear bands. H3M, ES and F9 all show the 127kD signal along with the indistinct signal at 31kD, identical to the B16 extract.

Using this particular blot the molecular mass (M_r) of the DNA-binding proteins was ascertained by constructing a standard curve based upon the rainbow molecular-weight markers used (Amersham). Figure 4.6 shows this curve, relating it to the bands on the blot and producing size-estimates (used in this chapter) for the DNA-binding polypeptides detected.

The conserved *Msx*-1-flanking DNA binds a protein (of ~127kD) common to a variety of several cell types; cells derived from the early embryo, the embryonic neural tube, an epitheloid carcinoma (adenocarcinoma) and a melanoma cell line all express this protein.

Figure 4.5

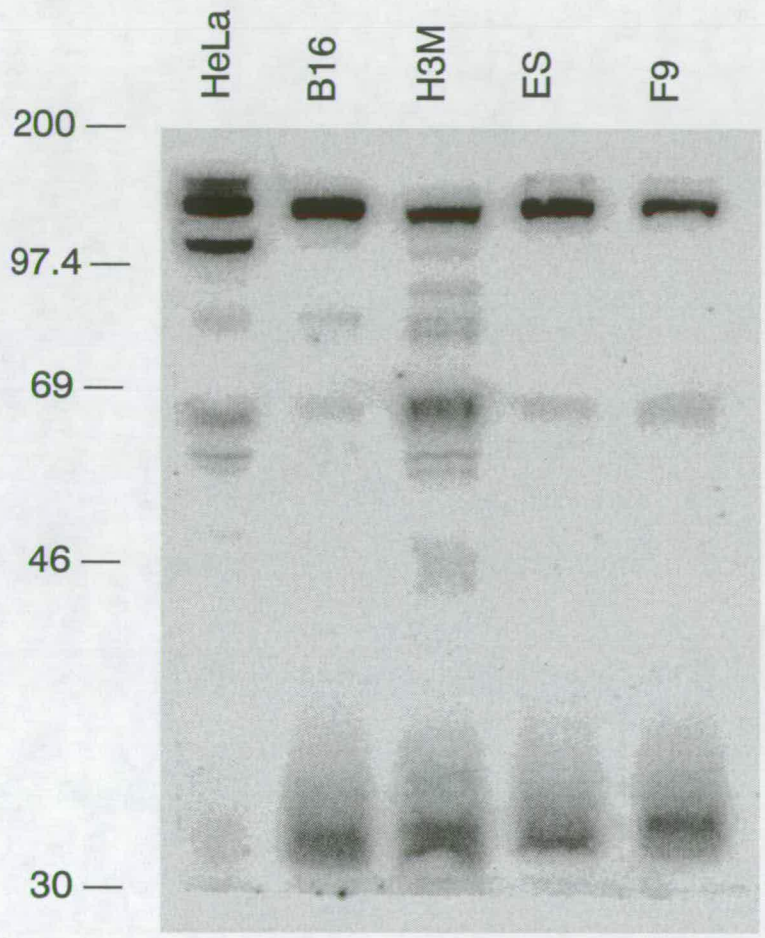


Figure 4.5 South-western blot using proteins from a variety of cell lines. These were probed with oligos D12/D13 corresponding to the conserved sequence in the mouse Msx-1 5'-flanking region. Lane 1 - 34µg HeLa cell extract; Lane 2 - 14µg B16 cell extract; Lane - 3 15µg H3M cell extract; Lane 4 - 15µg Es cell extract; Lane 5 - 15µg F9 cell extract. kiloDalton sizes on the left are derived from rainbow molecular weight markers (Amersham) run on the SDS-PAGE gel.

Figure 4.6

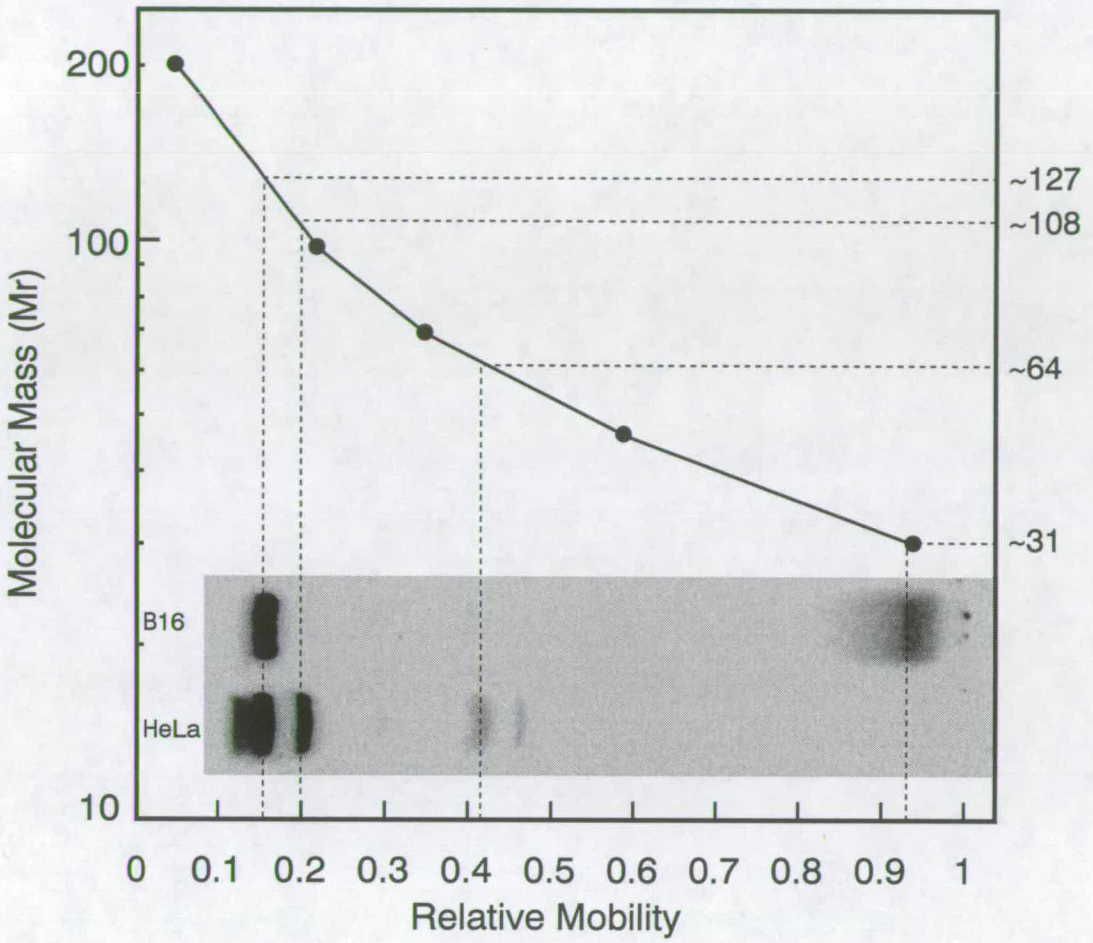


Figure 4.6 Assignment of molecular weight to the peptides detected by South-western blotting is shown by comparison to a standard curve constructed from plotting the relative mobility (distance migrated/distance of electrophoretic front) against the known sizes (logarithmic scale) of the markers used (Amersham Rainbow markers). A section of the blot shown in figure 4.5 is aligned accordingly showing intersections of the relative mobilities of the bands with the standard curve and the size estimates they provide, in kD.

4.2.3 *A sequence in the HoxD cluster binds identically sized proteins*

The *Hox*D sequence described as similar to the conserved *Msx*-1 sequence (section 3.3.4) was tested for protein-binding activity by South-western blotting. Figure 4.7 shows a composite blot made from three separate filters, identical in their protein content and processing conditions but hybridised to different probes. On each filter are lanes carrying HeLa protein and B16 protein (HeLa left, B16 right). The left panel (A) was probed with the human *Msx*-1 probe, the centre panel (B) was probed with the mouse *Hox*d-9 probe and the right panel (C) was probed with the mouse *Msx*-1 probe (see legend for details). There is a clear and marked similarity among all three panels. As before the mouse and human *Msx*-1 probes gave the same pattern of a band at 127kD and a diffuse signal at 31kD, plus the HeLa-specific band at 108kD. Significantly the centre panel, probed with the *Hox*d-9 flanking sequence, shows a similar pattern, strongly suggesting that all three sequence motifs are interacting with the same protein components of the extracts tested.

4.2.4 *The 127kD protein is expressed throughout the embryo*

Given our detailed knowledge of the spatial and temporal pattern of *Msx*-1 transcription (section 1.4), proteins with a causal role in the formation of this pattern, including transcription factors, may themselves show such restrictions.

The DNA-binding proteins identified are putative transcriptional regulators of *Msx*-1 and examining their distribution within the embryo may provide insight into any role they have in determining *Msx*-1 transcription patterns. Not having a molecular probe for the protein (or proteins) itself, the South-western blotting technique was used as a means whereby regions of the embryo could be assayed for the presence of the DNA-binding protein activity. Dissection of 11 dpc mouse embryos was performed, based on a knowledge of *Msx*-1 expression, and protein extracts produced for use in the South-western assay. Figure 4.8 shows a photograph of an 11 dpc mouse embryo across which several lines have been drawn indicating the mode of dissection. Both the forelimb and the hindlimb were severed at the point where they join the flank (figure 4.8 lines A & B); the hindbrain and majority of the mid and forebrain were removed (and discarded) by a single cut (line C) leaving the area of the outgrowing branchial arches which were removed from the trunk (line D); the remaining region anterior to the first limb bud was removed (line E) and used as a region of low expression (termed the '-ve' region in figures 4.9 & 4.10) - this represents a region of the embryo showing much lower (though not zero)

Figure 4.7

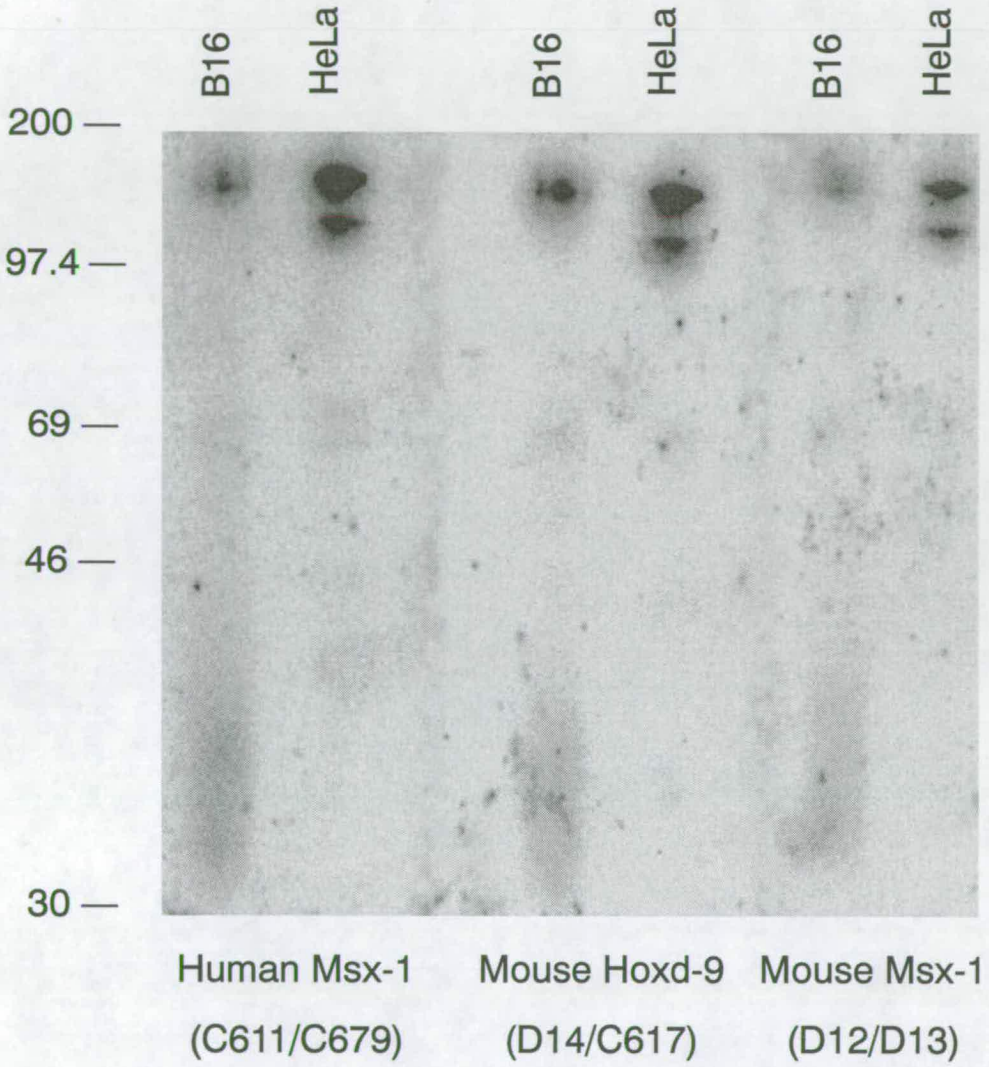


Figure 4.7 South-western blot showing identical protein-binding activity by the three related sequences, from mouse and human *Msx*-1 and from the *Hox*D cluster. Three South-western blots are shown, A, B and C. All have two protein samples loaded; in lane 1 the B16 cell extract and in lane 2 the HeLa cell extract. Blot A was probed with the human *Msx*-1 sequence (oligos C611/C679). Blot B was probed with the mouse *Hox*d-9 sequence (oligos D14/C617). Blot C was probed with the mouse *Msx*-1 sequence (oligos D12/D13).

Figure 4.8

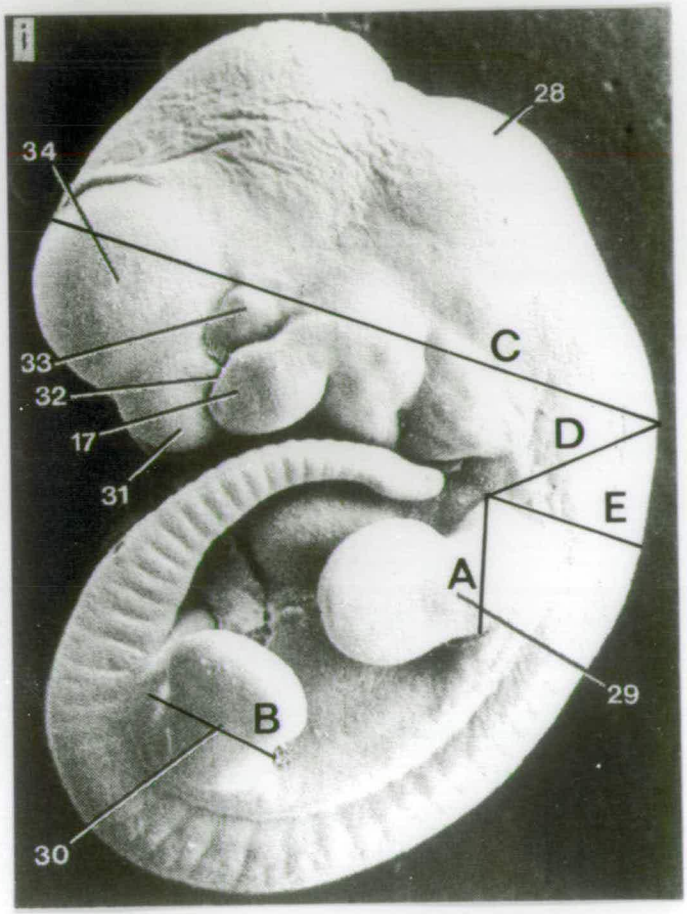


Figure 4.8 11 day mouse embryo showing mode of dissection for generation of embryonic extracts. Lines marked by letters indicate cuts made (see section 4.2.4). Ignore numbers.

expression than the others used; the remaining thorax and tail made up the 'trunk' extract.

The proportion of *Msx*-1 expressing cells in each of these regions varies greatly, and with it of course the concentration in a given extract of any factors specific to *Msx*-1 expressing cells. The arch area has a very high level of expression at this stage, with *Msx*-1 strongly transcribed in the primordia of the developing face (section 1.4.1). Both the fore and hind limbs express *Msx*-1 strongly in their distal mesenchyme: the relative temporal delay of hindlimb development means that it will have a higher proportion of *Msx*-1 expressing cells than the forelimb at this stage due to increased distal restriction of expression as the limbs grow out (section 1.4.2). Expression in the trunk is limited to the midline of the neural tube and the tail-bud/genital-ridge region (Lyons *et al.*, 1992). The trunk section as a whole will therefore contain a low proportion of *Msx*-1 expressing cells. The so called '-ve' area is a portion of the embryo in which expression is limited to the narrow stripe of cells at the midline of the neural tube and also contains very few expressing cells.

Figure 4.9 shows South-western hybridisation of the mouse probe (concatenated oligos D12 and D13) to the embryonic extracts described. Lanes 1 and 2 carry B16 and HeLa extracts, respectively, and give similar results to those described above (section 4.2.1). The remaining lanes show a binding pattern very similar to the B16 pattern, with the 127kD protein and the doublet (resolved in this example) at 31kD. This demonstrates that the cell lines tested and the embryo are expressing the same DNA-binding protein. Furthermore, there do not appear to be any gross variations in the distribution of the protein throughout the embryo. Faint signals are also visible on the blot at 64kD, as in the HeLa and H3M extracts in figure 4.5. The detection of these bands most likely reflects the better quality of hybridisation and background suppression in these two blots but it is impossible to say whether they represent low levels of additional polypeptides or degradation products, or high levels of non-specific proteins binding a detectable amount of the probe.

Figure 4.10 shows a blot of an identical gel hybridised to a probe generated from oligos B349/B351 (Table 4.1) representing an octamer element. A single band of approximately 70kD is hybridised in all cases (this band may be slightly smaller in the branchial 'arch' region of the embryo), with the exception of the HeLa cell extract which contains an additional, higher molecular weight binding-protein of approximately 80kD. The quality of signal is poor, possibly due to sub-optimal conditions for this interaction (the salt concentration may be above the optimal level for binding thereby destabilising the interaction), however it shows that the signals

Figure 4.9

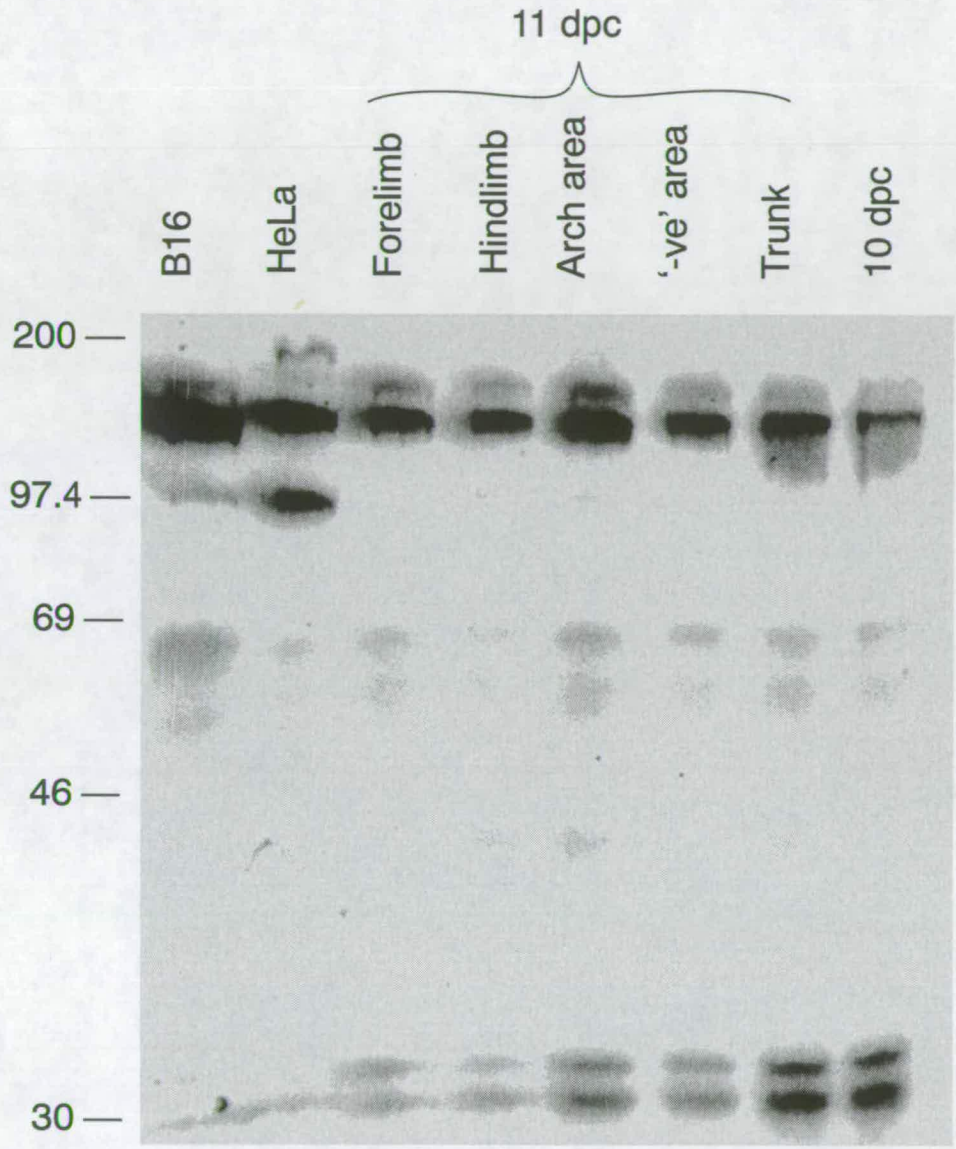


Figure 4.9 Southwestern blot in which the mouse probe (D12/D13) was hybridised to protein extracts from a variety of cells lines and portions of 11 day mouse embryo (see text). Embryonic portions used are those depicted in figure 4.8 and described in the text. Amount of protein loaded was as follows: B16 - 10µg; HeLa - 34µg; Forelimb-23µg; Hindlimb - 23µg; Arch - 35µg; negative - 34µg; trunk - 45µg; whole 10 dpc - 35µg.

Figure 4.10

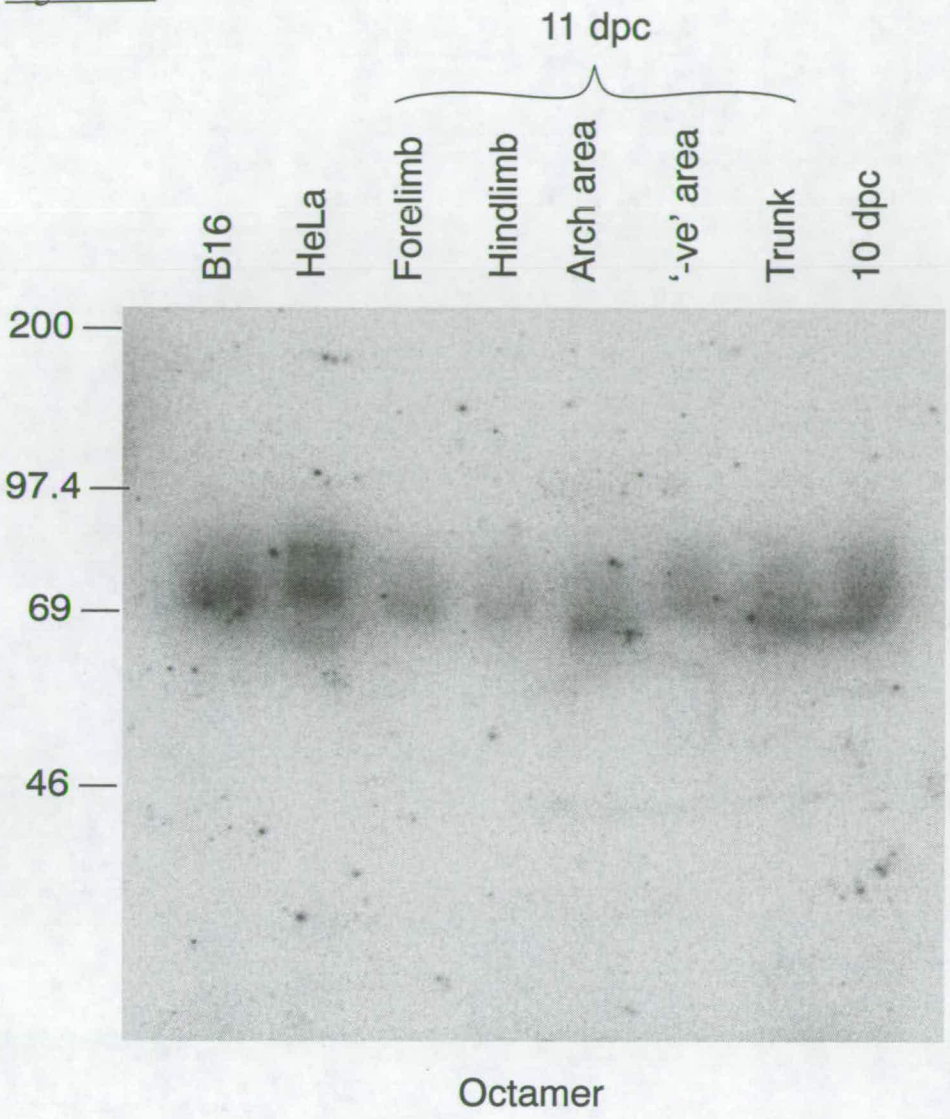


Figure 4.10 Southwestern blot showing hybridisation of a concatenated octamer probe (oligos B349/B351) hybridised to protein extracts from a variety of cell lines and portions of 11 dpc mouse embryo. Loading is as in figure 4.9. Conditions as before.

detected by the *Msx*-1 probes are sequence-specific and not detected by other DNA motifs.

4.2.5 Msx-*1 conserved sequence represses enhancer activity*

In vitro experiments were performed using the CAT (Chloramphenicol acetyl-transferase) reporter gene as a first step in assessing the regulatory function, if any, of the *Msx*-1 conserved sequence and the proteins it binds. These experiments were performed in collaboration with Dr. Alasdair MacKenzie. Having previously demonstrated identical protein-binding properties for the mouse and human sequences, these preliminary experiments were limited to use of the mouse sequence. The double stranded DNA fragment produced by annealing oligonucleotides D12 and D13 (Table 4.1) was used to create four constructs (figure 4.11) in two different reporter vectors. The first construct consists of plasmid pBLCAT3 (Luckow and Schutz, 1987) with a single copy of the mouse conserved sequence cloned upstream of a *tk* promoter driving the CAT gene. Basal transcription levels from the promoter will be raised by the *Msx*-1 sequence if it is capable of enhancer function and the appropriate binding factors are present in the host cells. The other three constructs are into the same vector, pCAT-Control (Promega), which has the CAT gene linked to a *tk* promoter and an SV40 enhancer providing high levels of CAT expression that will be reduced if the conserved sequence has a silencer function. These constructs carry 1, 2 or 3 copies of the mouse conserved element inserted downstream of the enhancer (which is itself downstream of the CAT gene). Transfection of these constructs into tissue-culture cell lines and subsequent determination of the relative CAT levels that they express provides information on the regulatory influence that the *Msx*-1 sequence has upon this gene. Figure 4.12 shows the results of transfection of these constructs into B16 and H3M cells. Results were standardised by dotblot measurement of transfected plasmid and averaged from two identical experiments (see legend to figure 4.12 for details).

A similar pattern of relative activity is seen in both cell lines, though the general level is lower in H3M than B16. In both cases the level of CAT activity in the absence of a reporter plasmid is barely detectable and effectively represents zero. Figure 4.12A&B (lanes 7 & 8) shows that in neither cell line is there significant activity from the unmodified pBLCAT plasmids: in H3M there is fractionally higher expression from the *tk* promoter (pBLCAT3) than the promoterless reporter (pBLCAT2), though this is undetectable in B16. There is no detectable effect of adding a single copy of the conserved element upstream of the *tk* promoter in the H3M cells and a very small effect in B16 (lane 6, figure 4.12B). The positive control plasmid (pCAT-Control; lane

Figure 4.11

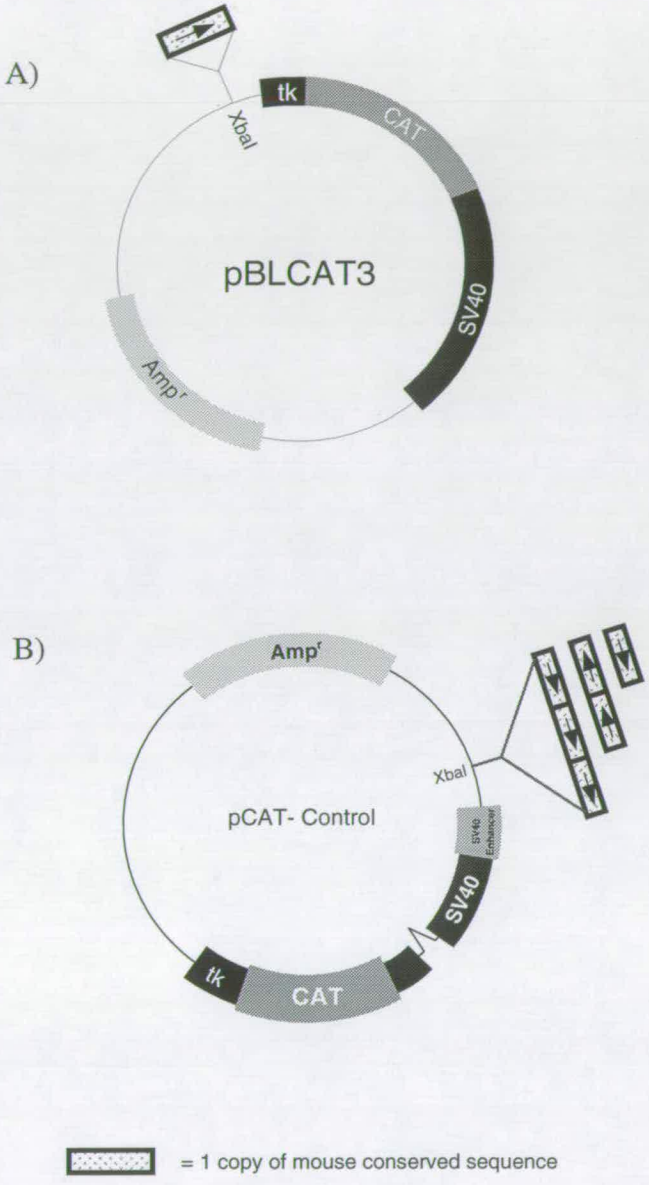


Figure 4.11 Constructs used in reporter gene assays. A) pBLCAT3 carrying a single copy of the mouse *Msx*-1 sequence upstream of its *tk* promoter. B) pCAT-Control carrying 1, 2 or 3 copies of the mouse *Msx*-1 sequence downstream of the SV40 enhancer, cloned in the orientation indicated.

Figure 4.12

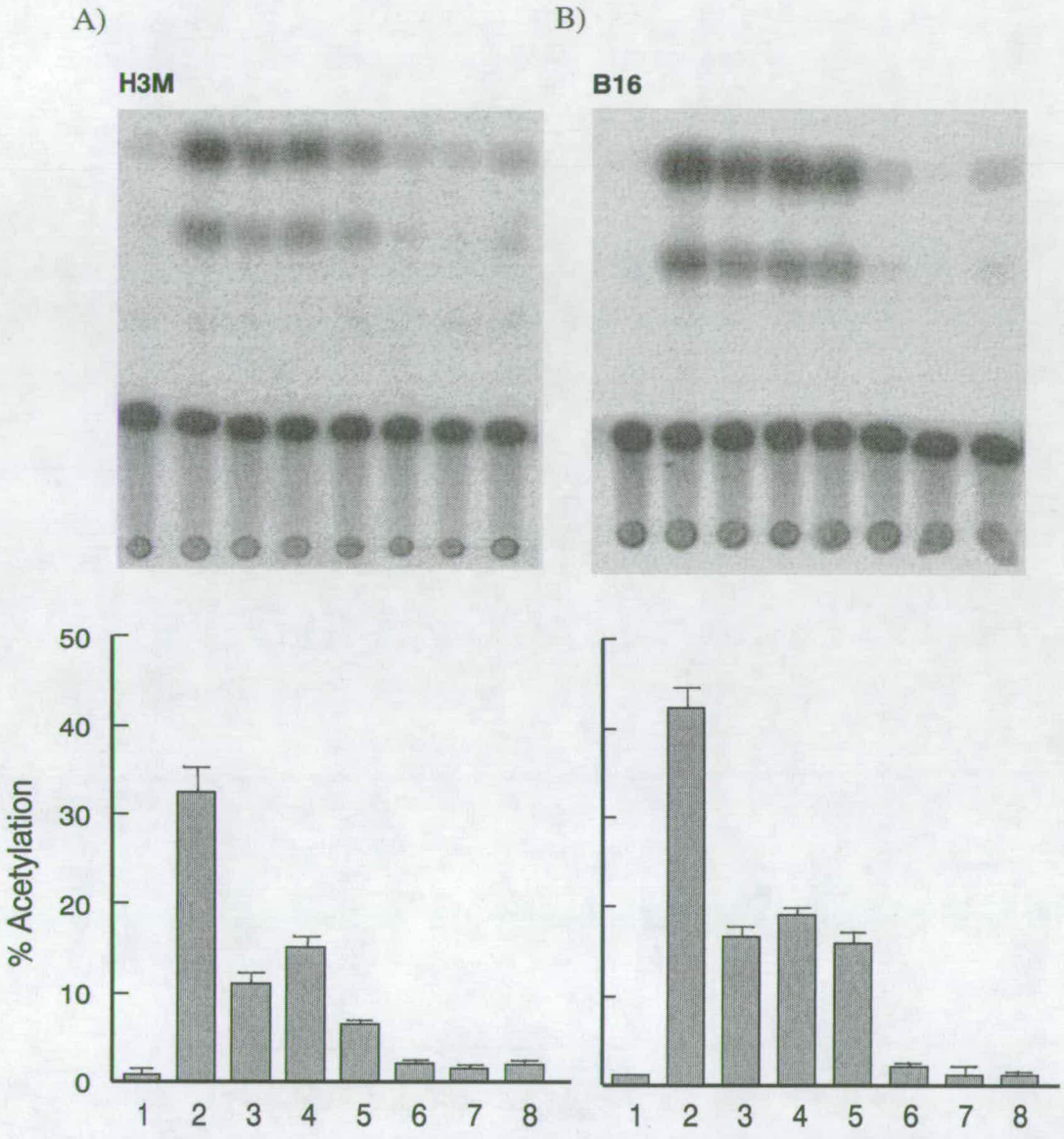


Figure 4.12 CAT assay results from H3M and B16 cells lines. The upper panel shows one of the two duplicate CAT assays performed for each cell line. The lower panel displays the results averaged from the two duplicate assays in histogram form (with bars showing maximum values). 4µg of plasmid DNA was added to each plate of ~10^5 cells and grown for 72 hours before harvesting. The ordinate is the percentage acetylation of the ^{14}C-Chloramphenicol added, calculated by summing the values for the two upper spots (acetylated forms) and dividing this by the sum of all 3 spots. Lane 1: No plasmid control, Lane 2: pCAT-Control, Lane 3: pCAT-Control+trimer (3 copies of the conserved sequence), Lane 4: pCAT-Control+dimer, Lane 5: pCAT-Control+monomer, Lane 6: pBLCAT3+monomer, Lane 7: pBLCAT3, Lane 8: pBLCAT2 (promoterless CAT). Measurements were performed using ImageQuant software after Phoshor Screen detection. Results were standardised by dotblot.

2) gives high levels of CAT activity in both cell lines confirming that both transfection and CAT assay have been successful. Addition of a single copy of the conserved sequence element downstream of the SV40 enhancer in this plasmid (figure 4.11) reduces the expression level from this enhancer by 3 fold in H3M and 2.5 fold in B16 (lane 3). Addition, in a similar orientation, of three copies of the conserved element reduces expression by 5.1 fold in H3M and 2.7 fold in B16 (lane 5). Interestingly, two copies of the element, oppositely oriented in comparison to the other constructs, also represses but to a lesser extent than either one or three copies; 2.2 fold in both H3M and B16 (lane 4). This may reflect the opposite orientation of the *Msx*-1 sequence copies in the dimer construct as compared to the monomer or trimer (figure 4.11). Further experiments analysing the effects of element number and orientation are required to consolidate these early findings though it appears that the conserved element upstream of *Msx*-1 functions as a repressor binding site.

4.3 Discussion

In vitro analysis has shown that the conserved sequences found upstream of the mouse and human *Msx*-1 genes (section 3.3.3) are capable of sequence-specific interactions with nuclear proteins. The proteins bound by these two sequences appear to have similar properties as adjudged by the gel-retardation assay. Sequences from mouse and human *Msx*-1 bound proteins from mouse and human cell lines. Binding of the sequences to proteins separated by SDS-PAGE (South-western blotting) reveals that the sequences bind proteins of identical size and that binding to these proteins is specific to this sequence. A similar sequence found upstream of the mouse *Hox*d-9 gene (section 3.3.4) was shown, by gel retardation assay, to compete with the mouse *Msx*-1 sequence for specific complex formation with the same proteins. This was confirmed by South-western hybridisation which showed that the *Hox*D sequence bound proteins of identical size to those bound by the *Msx*-1 sequences. The *Hox*D sequence is itself highly conserved between mouse and human and resides in a region proposed to have transcriptional regulatory function (Figure 3.24; Renucci *et al.*, 1992), though despite the conservation this was not one of the regions chosen by Renucci and colleagues for study by gel-retardation experiments.

Besides the observed similarity in the proteins bound by these sequences there is variation, where both sequences bind an additional, smaller protein from HeLa cells. This second polypeptide may be encoded by a second gene, possibly related, different from that encoding the common, larger polypeptide. Alternatively it may be encoded by the same gene and represent a specific cleavage product derived from the larger

protein or the product of translation from an alternative transcript (such as from a different promoter or alternative splice form). Any of these alternatives may represent a human-specific variation as HeLa extracts were the only human protein source examined. Use of additional human cell lines and further characterisation of the factor will be required to distinguish between these various possibilities.

Production of protein extracts from various parts of the 11 dpc mouse embryo enabled me to address, in a relatively crude way, the question of whether the proteins that were binding the *Msx*-1 conserved sequence were expressed in a spatially restricted manner. We might expect that certain features of the complex spatial distribution pattern of *Msx*-1 transcripts in the mid-gestation embryo (section 1.4) are generated in response to the influence of transcriptional regulators that are themselves spatially restricted in their activity. Dissection of the mouse embryo was performed in such a way as to separate, as best as possible, the different expression domains that might be generated by a differential set of regulators. Expression in the limbs and face in particular might respond to regulators having a specific anterior-posterior distribution. However, as figure 4.9 shows, the dissection performed was not able to separate any regions having different levels of binding activity for the proteins concerned. The results presented support the view that proteins binding the *Msx*-1 conserved sequence are ubiquitously expressed in the 11 dpc embryo. This approach did not enable the examination of protein distribution in different tissue types, such as epithelium and mesenchyme. Features of its expression suggest that there may be differences in the regulatory capacity of these two tissues with respect to *Msx*-1. It might be possible to perform similar experiments using extracts of cultured tissue explants to further investigate this. If there were a difference in protein activity between epithelium and mesenchyme the crude dissection used here would not have detected it and indeed would have shown homogeneous distribution as was seen. The suggestion that *Msx*-1 plays a similar role in inductive epithelial-mesenchymal interactions in diverse parts of the body (section 1.4.3) lends support to the notion that expression responds to local, short range signals. Experiments to study the consequences of such signals on the regulation of transcription will require the use of *in vitro* recombination systems of the type used by Vainio *et al.* (1993) in their work on tooth development.

Preliminary experiments to assess the transcriptional activity of the *Msx*-1 conserved sequence and the proteins that bind to it have provided evidence that it is the binding site of a transcriptional repressor. Though these findings are based on work that requires further confirmation we might speculate on the potential role for such a regulatory interaction. While it is thought unlikely that widespread repression has a

role in the definition of the *Msx*-1 expression pattern it is possible that refinement of earlier expression may be accomplished by negatively acting signals. Temporal regulation of expression is also likely to be achieved, at least in part, by repressing earlier activation. In the limb, for example, cells leaving the progress zone are seen to down regulate *Msx*-1 rapidly, possibly as a result of transcriptional repression. Such repression might be antagonised by short range signals from AER producing the sharp boundary of expression in the distal limb. Functional studies in the embryo will be required to test such models.

The data presented suggest that a protein-binding DNA element has been identified upstream of the mouse *Msx*-1 gene. This element is conserved in sequence and protein-binding function in the upstream region of human *Msx*-1. A similar element, with identical protein-binding properties is found within the *Hox*D cluster, between the *Hox*d-9 and *Hox*d-10 genes. Proteins bound by the *Msx*-1 sequence are broadly expressed throughout the developing mouse embryo and in several tissue-culture cell types. *In vitro* reporter-gene studies suggest that the sequence is bound by a transcriptional repressor.

In the wake of these findings attempts were made to pinpoint more precisely the bases involved in protein-binding. Knowledge of the precise bases required for binding would enable assessment of whether indeed these *are* the positions invariantly conserved between the various *Msx*-1 and *Hox*D sites. With precise binding-site localisation it might also be possible to generate mutant sites with attenuated binding properties for use in functional studies. Localisation of binding is often done by 'footprinting' the target site - that is measuring the inhibitory effect of bound protein on the endonuclease DNase I, which will cleave a DNA molecule between each base excepting any portion of the DNA masked by a binding-protein. To perform such an assay it is necessary to produce radiolabelled target-site DNA saturated with protein, i.e. a sample in which the vast excess of DNA molecules are protein-bound. This proved impossible in this case due to the lack of pure protein (all work was done with crude nuclear extracts) and the presumed low level of the proteins in question within the extracts available. A possible way around this problem would be to attempt purification, or at least enrichment, of this protein from large quantities of crude extract by affinity chromatography. It may also be possible to footprint the protein-DNA complex following direct isolation from a gel-retardation gel, however this is dependent upon a highly stable complex. If the complex dissociated during elution the protein component would at least be greatly enriched for the required peptide possibly

enabling later reassociation and footprinting. Such work was not possible within the confines of this study but represents a potential route for the continuation of this project. One way in which the protein-binding site might be more accurately defined from the data presented is in comparison with the *Hox*d-9 sequence. As seen in figure 3.24 the homology between the *Msx*-1 sequence and the *Hox*d-9 sequence is shorter than that seen between each sequence and its cognate. Of the 44bp of similarity between the mouse and human *Msx*-1 only the 5' 24bp is similar to the *Hox*d-9 sequence with the similarity apparently reduced at the 3' end of this region. It is possible that this restricted similarity defines the common protein-binding site. Where the sequences diverge may reflect an additional protein binding site generating diversity of function between these sequences in the two genes (section 3.5).

Protein purification by affinity chromatography for the purposes described may yield sufficient material from which to obtain some peptide sequence. This could subsequently be used in the design of a degenerate probe enabling isolation of a cDNA encoding a binding protein. An alternative approach to this traditional 'purify→protein sequence→clone' route, taken in the characterisation of many genes encoding transcriptional regulators (e.g. Briggs *et al.*, 1986 & Kadonaga *et al.*, 1987), is the direct cloning of a DNA-binding protein from a cDNA expression library using the technology of detecting membrane-bound protein with a radiolabelled DNA target-site (Singh *et al.*, 1988; Vinson *et al.*, 1988). This method was tried twice, without success, at the end of this study using a B16 cDNA library (a gift from Dr. Ian Jackson) and the concatenated probes described (section 2.9.1) though the demonstrated feasibility of the South-western blotting provides optimism that this route to cloning the gene encoding the binding-protein will be possible in the future.

The recurrence of similar *cis*-acting sequences in the regulatory regions of multiple genes is seen in two sets of circumstances: 1) where co-ordinate regulation of several genes is required in response to an individual signal, or 2) where the genes are cognates in different species and there has been evolutionary conservation of a *cis*-acting element. It appears that the sequence identified upstream of *Msx*-1 may fall into both categories.

The original model of Britten and Davidson (1969) provided a conceptual framework upon which subsequent molecular models of co-ordinate regulation have been founded. They proposed that multiple genes may respond to a single regulator by way of possessing identical regulatory elements. This regulator would enable the conversion of a single signal (regulating the regulator) into more complex patterns of gene expression. Common binding sites for such regulators (transcription factors) have

since been found upstream of a number of sets of genes with related function, thereby putting them under co-ordinated control. Examples of such systems include the T cell receptor genes (Leiden, 1993), genes involved in melanin synthesis (Jackson *et al.*, 1994 and pers. comm.) and histone genes differentially expressed during the cell-cycle (van den Endt *et al.*, 1994). Comparison of the increasing number of promoter sequences available will no doubt reveal more examples of this type of regulation. The sequence similarity between the *Msx*-1 5' flanking sequence and the *Hox*d-9 5' flank prompts the suggestion made in section 3.5 that these genes are subject to common controlling influences and are co-ordinately regulated in some respect. This idea is supported by the observation that both sequences bind similar proteins.

Implicit in the assumption that conserved sequences highlight regulatory regions is that the specificity of the protein-DNA interactions has also been maintained during evolution, and therefore that the trans regulatory factors are highly conserved. Both mouse and human *Msx*-1 sequences bind proteins of a similar size from two different sources, one human and one mouse. This result is consistent with the observation that the two sequences are highly homologous and might be expected to have similar protein-binding properties. The presence of such similar sized proteins with similar DNA-binding specificity in both mouse and human cells implies that not only has the DNA sequence been conserved over the evolutionary interval separating the two species (~80 million years) but a specific protein that binds this sequence has also been highly conserved. A number of reports have been made in the last few years describing the remarkable conservation of *cis* and *trans* regulatory interactions between widely diverged species. Falb and Maniatis (1992) described the conservation of a bipartite *cis* element involved in regulation of the *Drosophila* and human alcohol dehydrogenase (*Adh*) genes. Cotransfection experiments showed that the *Drosophila* element was active in human cells and that activity from it was antagonised by addition of the *Drosophila* factor normally bound to it (AEF-1). The extreme conservation seen between genes of the *Drosophila* HOM-C and vertebrate *Hox* clusters (section 1.2) prompted experiments that have further demonstrated conservation across the vertebrate-invertebrate divide. Reciprocal experiments showed that an autoregulatory element from the *Drosophila* gene *Deformed* (*Dfd*), active in the head of the fly, provided similar spatial regulation in transgenic mice, and that conversely sequences upstream of the *Hox*b-4 gene (a *Dfd* cognate) directed expression of a transgene to a region overlapping the expression domain of *Dfd* in the *Drosophila* head (Awgulewitsch and Jacobs, 1992; Malicki *et al.*, 1992). Experiments such as these indicate an extremely high level of conservation within not only the components of the transcription regulating mechanisms but the systems into which these components are

organised. Future cloning of the proteins binding to the *Msx*-1 conserved sequence and transgenic experiments using the human regulatory regions in mouse will reveal to what extent the regulatory system governing *Msx*-1 expression has been conserved. At present it is impossible to assess whether there has been conservation in the regulatory mechanisms governing expression of the *Drosophila msh* gene, the archetype of the *Msx* family, as little work has been done on this gene since its initial description, though this situation may change soon (W. Gehring pers. comm.). In the meantime studies of *Msx*-1 in species such as chicken will test and extend the findings reported here. Comparisons, similar to that presented here, between the mouse *Msx*-1, -2 and -3 genes will be of key interest as these genes share certain features of their complex developmental expression patterns and will very likely have in common regulatory pathways that are responsible for this coincidental expression.

The next steps in the continuation of this project must involve demonstration of the regulatory activity of the sequence identified. Firstly regulatory activity *in vitro* can be assessed, particularly as cell lines have been identified in which putative regulatory proteins are expressed. Preliminary experiments of this type have been begun in collaboration with Dr. Alasdair MacKenzie (section 4.2.5). Early findings from CAT reporter studies are that the *Msx*-1 conserved sequence acts as the binding site for a transcriptional repressor, as presence of the site reduced activity of the SV40 enhancer. This putative repressor is presumably one of the peptides detected by South-western blotting.

As an extension to these findings *in vivo* functional analysis in transgenic animals is under way (by Dr. Alasdair MacKenzie) to determine the significance, if any, of this sequence to the developmental expression of *Msx*-1. Using as a starting point a construct in which 5kb of the 5' flank is fused to the lacZ reporter gene, recreating many features of the early *Msx*-1 expression pattern (Robert Hill, unpublished results), deletion analysis has been initiated to locate the regulatory elements within this region. Alongside this study specific deletion of the conserved element has been achieved and its consequences examined. Early results suggest that this element may interact with other sequences in the generation of the spatially restricted expression pattern observed.

Summary

Summary

I have used an evolutionary approach to the identification of DNA elements with a role in the transcriptional regulation of the murine homeobox gene *Msx*-1. This gene has a complex temporal and spatial pattern of transcription during murine embryogenesis and is proposed to have a role in the processes of pattern formation and morphogenesis. Comparison of the mouse and human genes along with sequence from their 5' flank has revealed regions of particularly high conservation between the two species. The supposition made is that a lack of modification since the common ancestor of mouse and human reflects a functional role for these sequences. Work was concentrated on a sequence found within the 5' flank of both genes as it was felt that this was most likely to represent a *cis*-acting transcriptional regulatory element.

Results presented provide evidence for a sequence-specific DNA-protein interaction occurring between a site in the 5' flank of the *Msx*-1 gene and a nuclear protein expressed throughout the mid-gestation mouse embryo. Gel-retardation assays showed binding to a nuclear factor and suggested that both mouse and human sequences had a conserved function. South-western blots demonstrated that both sequences interacted with the same proteins and that these proteins were conserved in size between mouse and human. The conservation of both the DNA and protein components of this interaction during the evolutionary interval separating mouse and human implies that they play an essential role in the function of the *Msx*-1 gene. This function is most likely concerned with the transcriptional regulation of the gene given the position of the DNA sequence in relation to the coding region. Search for similar sequences in the 5' flank of other genes revealed only one match, in the 5' flank of a homeobox gene of the *Hox*D cluster, *Hox*d-9. Using the same techniques this sequence was shown to have protein-binding properties identical to the *Msx*-1 sequence. There are similarities in the expression pattern and regulating influences acting upon *Msx*-1 and *Hox*d-9 (both AER responsive; section 3.5) and it is possible that this conserved sequence reflects response by both genes to a common molecular pathway, possibly stimulated by inductive interactions between the epithelium and mesenchyme. If this sequence does indeed act as a *cis* regulatory element then it may have influence over several genes in the *Hox*D cluster, most likely the five *Abd*-B homologues at the 5' end of the cluster expressed during development of the limb, at a stage similar to *Msx*-1. Preliminary *in vitro* experiments suggest that this sequence is the binding site of a transcriptional repressor and further studies are underway to clarify the role of this element in the embryonic expression of *Msx*-1.

Sequencing of the *Msx*-1 5'-flank from other organisms, in particular the chicken, will provide an extension to the findings reported here. The CpG island-like region discovered at the 5' end of the *Msx*-1 gene in mouse and human is proposed to be in a process of loss, examination of the chicken gene in this regard will provide a further test for this hypothesis. Discovery of sequence similarity in the chick to those regions shown to be conserved between mouse and human would provide compelling evidence of their functional significance and may provide a more precise definition of the active sequences. The most direct test available at present to determine the functional significance of this putative regulatory element involves the use of genetically modified mice. Work is under way to test *cis* regulatory activity in the *Msx*-1 5' flank using transgenic mice in which a lacZ reporter gene is fused to this sequence. Techniques of protein purification or direct expression cloning can be applied to isolate the gene (or genes) encoding *trans*-acting factor(s) that interact with *cis* elements pinpointed in these ways. Experiments would then be possible to establish the role of such genes in generating the *Msx*-1 expression pattern. This might enable the ultimate goal of these studies which is to piece together the complex network of interactions taking place in the developing vertebrate body and to understand the relationship between molecules and morphology.

References

References

Akam M., 1988. Homeotic genes and the control of segmental diversity. *Development* **104** 123-133

Akimenko M. -A. *et al.*, 1991. Characterisation of three Zebrafish genes related to *Hox*-7. Developmental patterning of the vertebrate limb. Eds Hinchliffe J. R., Hurle J. M. and Summerbell D. *NATO ASI Series* vol A205 61-64

Ali N. and Bienz M., 1991. Functional dissection of *Drosophila* Abdominal-B protein. *Mechanisms in Development* **35** 55-64

Altaba A. R. I. And Jessell T.M., 1991. Retinoic Acid modifies the pattern of cell-differentiation in the central-nervous-system of neurula stage *Xenopus* embryos. *Development* **112** 945-958

Altaba A. R. I. And Jessell T., 1991. Retinoic Acid modifies mesodermal matterning in early *Xenopus* embryos. *Genes and Development* **5** 175-187

Anderson G. M. and Freytag S. O., 1991. Synergistic activation of a human promoter *in vivo* by transcription factor Sp1. *Molecular and Cellular Biology* **11** 1935

Andrews N. C. and Faller D. V., 1991. A rapid micropreparation technique for extraction of DNA-binding proteins from limiting numbers of mammalian cells. *Nucleic Acids Research* **19** 2499

Angel P. *et al.*, 1987. Phorbol ester-inducible genes contain a common cis element recognised by a TPA-modulated trans-acting factor. *Cell* **49** 729-739

Antequera F. and Bird A., 1993a. CpG Islands. in *DNA Methylation: Molecular Biology and Biological Significance*. eds Jost J. P. and Saluz H. P. - Berghauser Verlag, Basel Switzerland

Antequera F. and Bird A., 1993b. Number of CpG islands and genes in human and mouse. *Proceedings of the National Academy of Sciences of the U.S.A* **90** 11995-11999

Appel B. And Sakonju S., 1993. Cell-type-specific mechanisms of transcriptional repression by the homeotic gene-products Ubx and Abd-A in *Drosophila* embryos. *EMBO Journal* **12** 1099-1109

Atchison M. L., 1988. Enhancers - Mechanisms of action and cell specificity. *Annual Review of Cell Biology* **4** 127-153

Attardi L. D. and Tjian R., 1993. *Drosophila* tissue-specific transceiption factor NTF-1 contains a novel isoleucine-rich activation motif. *Genes and Development* **7** 1341-1353

Auwerx J. and Sassone-Corsi P., 1992. AP-1 (Fos-Jun) regulation by IP-1: effect of signal transduction pathways and cell growth. *Oncogene* **7** 2271-2280

Awgulewitsch A. and Jacobs D., 1992. *Deformed* regulatory element from *Drosophila* functions in a conserved manner in transgenic mice. *Nature* **358** 341-344

Balling R. *et al.*, 1989. Craniofacial abnormalities induced by ectopic expression of the Homeobox gene *Hox*-1.1 in transgenic mice. *Cell* **58** 337-347

Banerji J. *et al.*, 1981. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27** 299-308

Barberis A. *et al.*, 1987. Mutually exclusive interaction of the CCAAT-binding factor and of a displacement protein with overlapping sequences of a histone gene promoter. *Cell* **50** 347-359

Bastian H. *et al.*, 1992. The murine even-skipped-like gene *Evx*-2 is closely linked to the *Hox*-4 complex, but is transcribed in the opposite direction. *Mammalian Genome* **3** 241-243

Bateson W., 1894. Materials for the Study of Variation: Treated with especial regard to discontinuity in the Origin of Species. Macmillan & Co., London.

Beeman R. W. *et al.*, 1993. Structure and function of the homeotic gene-complex (HOM-C) in the beetle, Tribolium castaneum. *Bioessays* **15** 439-444

Behringer R. R. *et al.*, 1987. Two 3' sequences direct adult erythroid-specific expression of human beta-globin genes in transgenic mice. *Proceedings of the National Academy of Sciences of the U.S.A.* **84** 7056-7060

Bell J. R. *et al.*, 1993. Genomic structure, chromosomal location and evolution of the mouse *Hox*-8 gene. *Genomics* **16** 123-131

Bender W. *et al.*, 1983. Molecular genetics of the *Bithorax* complex in *Drosophila melanogaster*. *Science* **221** 23-29

Bennett D. C., 1983. Differentiation in mouse melanoma cells: Initial reversibility and an on-off stochastic model. *Cell* **34** 445-453

Benton M. J., 1990. Phylogeny of the major tetrapod groups - morphological data and divergence dates. *Journal of Molecular Evolution* **30** 409-424

Benyajati C. *et al.*, 1992. Characterisation and purification of Adh distal promoter factor 2, Adf-2, a cell-specific and promoter specific repressor in *Drosophila*. *Nucleic Acids Research* **20** 4481-4489

Bermingham J. R. *et al.*, 1990. Different patterns of transcription from the two Antennapedia promoters during *Drosophila* embryogenesis. *Development* **109** 553-566

Bernstine E. G. *et al.*, 1973. Alkaline phosphatase activity in mouse teratoma. *Proceedings of the National Academy of Sciences of the U.S.A.* **70** 3899-3903

Bieberich C. J. *et al.*, 1990. Evidence for positive and negative regulation of the *Hox*-3.1 gene. *Proceedings of the National Academy of Sciences of the U.S.A.* **87** 8462-8466

Biggin M. D. and Tjian R., 1989. A purified *Drosophila* homeodomain represses transcription *in vitro*. *Cell* **58** 433-440

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

Blum, M. *et al.* 1992. Gastrulation in the mouse: the role of the homeobox gene *goosecoid*. *Cell* 69 1097-1106.

Bouvagnet P. F. *et al.*, 1987. Multiple positive and negative 5' regulatory elements control the cell-type-specific expression of the embryonic skeletal myosin heavy-chain gene. *Molecular and Cellular Biology* **7** 4377-4389 [published erratum in *Molecular and Cellular Biology* **8** 1010]

Bowen B. *et al.*, 1980. The detection of DNA-binding proteins by protein blotting *Nucleic Acids Research* **8** 1-20

Brand A. H. *et al.*, 1985. Characterization of a "silencer" in yeast: a DNA sequence with properties opposite to those of a transcriptional enhancer. *Cell* **41** 41-48

Breier G. *et al.*, 1986. Sequential expression of murine homeo box genes during F9-EC cell-differentiation. *EMBO Journal* **5** 2209-2215

Briggs M. R. *et al.*, 1986. Purification and biochemical characterisation of the promoter-specific transcription factor, Sp1. *Science* **234** 47-52

Britten R. J. and Davidson E. H., 1989. Gene regulation for higher cells: a theory. *Science* **165** 349-358

Brown J. M. *et al.*, 1993. Experimental analysis of the control of expression of the homeobox gene *Msx*-1 in the developing limb and face. *Development* **119** 41-48

Bruggemeier U. *et al.*, 1990. Nuclear Factor-1 acts as a transcription factor on the MMTV promoter but competes with steroid-hormone receptors for DNA-binding. *EMBO Journal* **9** 2233-2239

Bucher P., 1993. The Eukaryotic Promoter Database, EPD. EMBL Nucleic acid sequence data library, Release 37. Postfach 10.2209, D-6900 Heidelberg.

Buettner R. *et al.*, 1993. An alternatively spliced mRNA from the AP-2 gene encodes a negative regulator of transcriptional activation by AP-2. *Molecular and Cellular Biology* **13** 4174-4185

Bulfone A. *et al.* ,1993. The mouse *Dlx*-2 (*Tes*-1) gene is exrpessed in spatially restricted domains of the forebrain, face and limbs in midgestation mouse embryos. *Mechanisms of Development* **40** 129-140 (*erratum* published in *Mechanisms of Development* **42** 187)

Bullock W. O. *et al.*, 1987. XL1-Blue - A high-efficiency plasmid transforming *rec*A *Escherichia coli* strain with β-Galactosidase selection. *Biotechniques* **5** 376-379

Buratowski S. *et al.*, 1989. Five intermediate complexes in transcription initiation by RNA polymerase II. *Cell* **56** 549-561

Bürglin T. R. *et al.*, 1989. *Caenorhabditis elegans* has scores of homeobox-containing genes. *Nature* **341** 239-243

Carpenter E.M. *et al.*, 1993. Loss of *Hox*a-1 (*Hox*-1.6) function results in the reorganization of the Murine hindbrain. *Development* **118** 1063-1075

Carrasco A. E. *et al.*, 1984. Cloning of a *Xenopus laevis* gene expressed during early embryogenesis coding for a peptide region homologous to *Drosophila* homeotic genes. *Cell* **37** 409-414

Carrington J. L. and Fallon J F., 1988. Initial limb budding is independent of apical ectodermal ridge activity: evidence from a *limbless* mutant *Development* **104** 361-367

Catron K. M. *et al.*, 1993. Nucleotides flanking a conserved TAAT core dictate the DNA-binding specificity of 3 murine homeodomain proteins. *Molecular and Cellular Biology* **13** 2354-2365

Chan S. -W. and Mann R. S., 1993. The segment identity functions of *Ultrabithorax* are contained within its homeodomain and carboxy-terminal sequences. *Genes and Development* **7** 796-811

Chisaka O. and Capecchi M. R., 1991. Regionally restricted developmental defects resulting from targeted disruption of the mouse homeobox gene *Hox*-1.5. *Nature* **350** 473-479

Cho K. W. Y. *et al.*, 1991. Overexpression of a homeodomain protein confers axis-forming activity to uncommitted *Xenopus* embryonic-cells. *Cell* **65** 55-64

Clerc R. G. *et al.*, 1988. The B-cell-specific Oct-2 protein contains POU box- and homeobox-type domains. *Genes and Development* **2** 1570-1581

Coelho C. N. D. *et al.*, 1991. Expression of the chicken homeobox-containing gene Ghox-8 during embryonic chick limb development. *Mechanisms of Development* **34** 143-154

Coelho *et al.*, 1992. *GHox*-7: a chicken homeobox-containing gene expressed in a fashion consistent with a role in patterning events during embryonic chick limb development. *Differentiation* **49** 85-82

Coelho C. N. D. *et al.*, 1993a. Ectoderm from various regions of the developing chick limb bud differentially regulates the expression of the chicken homeobox-containing genes Ghox-7 and Ghox-8 by limb mesenchymal cells. *Developmental Biology* **156** 303-306

Coelho C. N. D. *et al.*, 1993b. The expression pattern of the chicken homeobox-containing gene Ghox-7 in developing Polydactylous limb buds suggests its involvement in Apical Ectodermal Ridge-directed outgrowth of limb mesoderm and in programmed cell-death. *Differentiation* **52** 129-137

Cohen R. B. *et al.*, 1986. Partial purification of a nuclear protein that binds to the CCAAT box of the mouse α_1-globin gene. *Molecular and Cellular Biology* **6** 821-832

Cohen S. M. And Jurgens G., 1990. Mediation of *Drosophila* head development by gap-like segmentation genes. *Nature* **346** 482-484

Colantuoni V. *et al.*, 1987. Negative control of liver-specific gene expression: cloned human retinol-binding protein gene is repressed in HeLa cells. *EMBO Journal* **6** 631-636

Colberg-Poley A. M. *et al.*, 1985. Expression of murine genes containing homeo box sequences during visceral and parietal endoderm differentiation of embryonal carcinoma stem cells. *Cold Spring Harbor Symposium of Quantitative Biology* **50** 285-290

Courey A. J. and Tjian R., 1988. Analysis of Sp1 *in vivo* reveals multiple domains, including a novel glutamine-rich activation motif. *Cell* **55** 887-898

Courey A. J. *et al.*, 1989. Synergistic activation by the glutamine-rich domains of human transcription factor Sp1. *Cell* **59** 827-836

Croston G. E. *et al.*, 1991. Sequence-specific antirepression of histone H1-mediated inhibition of basal RNA polymerase II transcription. *Science* **251** 643-649

Culotta V. C. *et al.*, 1989. Fine mapping of a mouse metallothionein gene metal response element. *Molecular and Cellular Biology* **9** 1376-1380

Curran T. and Franza B. R. Jr., 1988. Fos and Jun: the AP-1 connection. *Cell* **55** 395-397

Czerny T. *et al.*, 1993. DNA-Sequence Recognition by Pax proteins - bipartite structure of the Paired domain and its binding site. *Genes and Development* **7** 2048-2061

Davidson D. R. and Hill R. E., 1991. *Msh* like genes: a family of Homeobox genes with wide ranging expression during vertebrate development. *Seminars in Developmental Biology* **2** 405-412

Davidson D. *et al.*, 1988. A gene with sequence similarity to *Drosophila engrailed* is expressed during the development of the neural tube and vertebrae in the mouse. *Development* **104** 305-316

Davidson D. R. *et al.*, 1991. Position-dependent expression of 2 related homeobox genes in developing vertebrate limbs. *Nature* **352** 429-431

de Villiers J. and Schaffner W., 1981. A small segment of polyoma virus DNA enahnces the expression of a cloned beta-globin gene over a distance of 1400 base pairs. *Nucleic Acids Research* **9** 6251-6264

Dearolf C. R. *et al.*, 1990. Transcriptional regulation of *Drosophila* segmentation gene *fushi tarazu* (*ftz*). *Bioessays* **12** 109-113

deBoer E. *et al.*, 1988. The human beta-globin promoter; nuclear protein factors and erythroid specific induction of transcription. *EMBO Journal* **7** 4203-4212

denDennen J. T. *et al.*, 1989. Nucleotide sequence of the rat γ-crystallin gene region and comparison with an orthologous human region. *Gene* **78** 201-213

Desplan C. *et al.*, 1985. The *Drosophila* developmental gene, *engrailed*, encodes a sequence-specific DNA-binding activity. *Nature* **318** 630-635

Dessain S. *et al.*, 1992. Antp-type homeodomains have distinct DNA binding specificities that correlate with their regulatory functions in embryos. *EMBO Journal* **11** 991-1002

Diamond M. I. *et al.*, 1990. Transcription factor interactions: selectors of positive or negative regulation from a DNA element. *Science* **249** 1266-1272

Dikstein R. *et al.*, 1992. c-abl has a sequence-specific enhancer binding activity. *Cell* **69** 751-757

Dolecki G. J. *et al.*, 1986. Stage-specific expression of a homeo box-containing gene in the non-segmented sea-urchin embryo. *EMBO Journal* **5** 925-930

Dollé P. *et al.*, 1989. Coordinate expression if the murine *Hox*-5 complex homeobox containing genes during limb pattern formation. *Nature* **342**, 767-772

Dollé P. *et al.*, 1991. *Hox*-4 genes in the morphogenesis of mammalian genitalia. *Genes & Development* **5** 1767-1776

Dollé P. *et al.*, 1993a. Local alterations of *Krox*-20 and *Hox* gene-expression in the hindbrain suggest lack of rhombomere-4 and rhombomere-5 in homozygote null *Hox*a-1 (*Hox*-1.6) mutant embryos. *Proceedings of the National Academy of Sciences of the U.S.A.* **90** 7666-7670

Dollé P. *et al.*, 1993b. Disruption of the *Hox*d-13 gene induces localised heterochrony leading to mice with neotenic limbs. *Cell* **75** 431-441

Dorn A. *et al.*, 1987. A multiplicity of CCAAT box-binding proteins. *Cell* **50** 863-872

Driever W. and Nüsslein-Volhard C., 1988. A gradient of bicoid protein in *Drosophila* embryos. *Cell* **54** 83-93

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

Duboule D., 1992. The Vertebrate Limb - A model system to study the *Hox*/HOM gene network during development and evolution. *Bioessays* **14** 375-384

Duboule D. & Dollé P., 1989. The structural and functional-organisation of the murine *Hox* gene family resembles that of *Drosophila* homeotic genes. *EMBO Journal* **8**, 1497

Dunaway M. and Dröge P., 1989. Transactivation of the *Xenopus* rRNA gene promoter by its enhancer. *Nature* **341** 657-659

Duret L. *et al.*, 1993. Strong conservation of non-coding sequences during vertebrates evolution: potential involvement in post-transcriptional regulation of gene expression. *Nucleic Acids Research* **21** 2315-2322

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

Dynan W. S., 1989. Modularity in promoters and enhancers. *Cell* **58** 1-4

Eid R. *et al.*, 1992. Analysis of LacZ reporter genes in transgenic embryos suggests the presence of several cis-acting regulatory elements in the murine *Hox*b-6 gene. *Developmental Dynamics* **196** 205-216

Ekker S. C. *et al.*, 1991. Optimal DNA-sequence recognition by the Ultrabithorax homeodomain of *Drosophila*. *EMBO Journal* **10** 1179-1186

Ekker M. *et al.*, 1992a. Co-ordinate embryonic expression of three Zebrafish engrailed genes *Development* **116** 1001-10

Ekker M. *et al.*, 1992b. Regional expression of three homeobox transcripts in the inner ear of zebrafish embryos. *Neuron* **9** 27-35

Ekker S. C. *et al.*, 1992c. Differential DNA sequence recognition is a determinant of specificity in homeotic gene action. *Neuron* **11** 4059-4072

Evans R. M., 1988. The steroid and thyroid hormone receptor superfamily. *Science* **240** 889-895

Falb D. and Maniatis T., 1992. A conserved regulatory unit implicated in tissue-specific gene expression in *Drosophila* and man. *Genes and Development* **6** 454-465

Feinberg A. P. and Vogelstein B., 1983. A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. *Analytical Biochemistry* **132** 6-13

Fjose A. *et al.*, 1985. Isolation of a homeo box-containing gene from the *engrailed* region of *Drosophila* and the spatial-distribution of its transcripts. *Nature* **313** 284-289

Flamant F. *et al.*, 1987. An embryonic DNA-binding protein specific for the promoter of the retrovirus long terminal repeat. *Molecular and Cellular Biology* **7** 3548-3553

Fletcher C. *et al.*, 1987. Purification and characterisation of OTF-1 a transcription factor regulating cell-cycle expression of a human histone H2b gene. *Cell* **51** 773-781

Florence B. *et al.*, 1991. DNA-binding specificity of the fushi tarazu homeodomain. *Molecular and Cellular Biology*. **11** 3613-3623

Fraser S. *et al.*, 1990. Segmentation in the chick-embryo hindbrain is defined by cell lineage restrictions. *Nature* **344** 431-435

Fried M., 1989. Measurement of protein-DNA interaction parameters by electrophoresis mobility shift assay. *Electrophoresis* **10** 366-376

Fried M. and Crothers D. M., 1981. Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Research* **9** 6505-6525

Friedman A. D. *et al.*, 1989. CCAAT/enhancer binding protein activates the promoter of the serum albumin gene in cultured hepatoma cells. *Genes and Development* **3** 1314-1322

Galang C. K. And Hauser C. A., 1992. Cooperative DNA-binding of the highly conserved human Hox-2.1 homeodomain gene-product. *New Biologist* **4** 558-568

Garabedian M. J. *et al.*, 1986. A tissue-specific transcription enhancer from the *Drosophila* yolk protein 1 gene. *Cell* **45** 859

Garber R. L. *et al.*, 1983. Genomic and cDNA clones of the homeotic locus *Antennapedia* in *Drosophila*. *EMBO Journal* **2** 2027-2036

Garcia-Bellido A., 1975. Genetic control of wing disc development in *Drosophila*. *Cell Patterning: CIBA Foundation Symposium* **29** 161-178

Garcia-Fernandez J. and Holland P. W. H., 1994. Archetypal organization of the Amphioxus *Hox* gene-cluster. *Nature* **370** 563-566

Garner M. M. and Revzin A., 1981. A gel elctrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the *Escherichia coli* lactose operon regulatory system. *Nucleic Acids Research* **9** 3407

Gaub M. P. *et al.*, 1987. The chicken ovalbumin promoter is under negative control which is relieved by steroid hormones. *EMBO Journal* **6** 2313-2320

Gaunt S. J. *et al.*, 1988. Spatially restricted domains of homeo-gene transcripts in mouse embryos - relation to a segmented body plan. *Development* **104** 169-179

Gaunt S. J. *et al.*, 1993. Expression of the mouse *goosecoid* gene during mid-embryogenesis may mark mesenchymal cell lineages in the developing head, limbs and body wall. *Development* **117** 769-778

Gavin B. J. *et al.*, 1990. Expression of multiple novel *Wnt*-1/*Int*-1-related genes during fetal and adult-mouse development. *Genes and Development* **4** 2319-2332

Gehring W. J., 1987. The homeobox: Structural and Evolutionary aspects. in, Molecular approaches to Developmental Biology, eds Firtel R. A. and Davidson E. H.; Liss, New York. 115-129

Genetics Computer Group, 1991. Program manual for the GCG package, Version 7, April 1991, 575 Science Drive, Madison, Wisconsin, USA 53711

Gey G. O. *et al.*, 1952. *Cancer Research* **12** 264

Ghosh D., 1990. A relational database of transcription factors. *Nucleic Acids Reasearch* **18** 1749-1756

Ghosh D., 1992. TFD - the Transcription Factor Database. *Nucleic Acids Research* **20** Suppl. 2091-2093

Ghosh D., 1993. Status of the Transcription Factor Database (TFD). *Nucleic Acids Research* **21** 3117-3118

Gibson T. J. *et al.*, 1993. Proposed structure for the DNA-binding domain of the helix-loop-helix family of eukaryotic gene regulatory proteins. *Protein Engineering* **6** 41-50

Giguere V. *et al.*, 1986. Functional domains of the human glucocorticoid receptor. *Cell* **46** 645-652

Giguere V. and Evans R. M., 1990. Identification of receptors for Retinoids as members of the Steroid and Thyroid-Hormone receptor family. *Methods in Enzymology* **189** 223-232

Gillies S. D. *et al.*, 1983. A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. *Cell* **33** 717-728

Gilmour D. S. *et al.*, 1989. *Drosophila* nuclear proteins bind to regions of alternating C and T residues in gene promoters. *Science* **245** 1487-1490

Gluzman Y., ed. 1985. Eukaryotic Transcription: The role of *cis*- and *trans*-acting elements in initiation. Cold Spring Harbor Laboratory Press, New York.

Goding C. R. *et al.*, 1987. Multiple transcription factos interact with the adenovirus-2 EII late promoter: evidence for a novel CCAAT recognition factor. *Nucleic Acids Research* **15** 7761-7780

Goldberg M. L. *et al.*, 1989. The *Drosophila zeste* locus is non-essential. *Genetics* **123** 145-155

Gorman C. M. *et al.*, 1982. Recombinant genomes which express chloramphenicol acetyltransferase in mammalian cells. *Molecular and Cellular Biology* **2** 1044-1051

Govind S. and Steward R., 1993. Coming to grips with *Cactus*. *Current Biology* **3** 351-354

Graham A. *et al.*, 1989. The murine and *Drosophila* homeobox gene complexes have common features of organisation and expression. *Cell* **57** 367-378

Graves B. J. *et al.*, 1986. Homologous recognition of a promoter domain common to the MSV LTR and the HSV tk gene. *Cell* **44** 565-576

Green J. M. *et al.*, 1987. Multiple basal elements of a human hsp70 promoter function differently in human and rodent cell lines. *Molecular and Cellular Biology* **7** 3646-3655

Grosveld F. *et al.*, 1987. Position-independent, high-level expression of the human beta-globin gene in transgenic mice. *Cell* **51** 975-985

Grueneberg D. A. *et al.*, 1992. Human and *Drosophila* homeodomain proteins that enhance the DNA-binding activity of Serum Response Factor. *Science* **257** 1089-1095

Grunstein M. and Hogness D. S., 1975. Colony Hybridisation: A method for the isolation of cloned DNAsthat contain a specific gene. *Proceedings of the National Academy of Sciences of the U.S.A.* **72** 3961-3965

Gumucio D. L. *et al.*, 1991. γ-Globin gene regulation: evolutionary approaches. in *The regulation of hemoglobin switching* G. Stamatoyannopoulos and A. W. Nienhuis, eds. Johns Hopkins University Press, Baltimore

Ha I. *et al.*, 1993. Multiple functional domains of human transcription factor-IIB: distinct interactions with two general transcription factors and RNA polymerase II. *Genes and Development* **7** 1021-1032

Han K. H. *et al.*, 1989. Synergistic activation and repression of transcription by *Drosophila* homeobox proteins. *Cell* **56** 573-583

Hanahan D., 1985. Techniques for transformation of *E. coli.* in *DNA cloning: A practical approach Volume 1* (ed D. M. Glover) IRL Press Oxford

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

Hanes S. D. and Brent R., 1991. A genetic model for the interaction of the homeodomain recognition helix with DNA. *Science* **251** 426-430

Hardison R. and Miller W., 1993. Use of long sequence alignments to study the evolution and regulation of mammalian globin gene clusters. *Molecular Biology and Evolution* **10** 73-102

Hart C. P. *et al.*, 1987. Seqeucne analysis of the murine Hox-2.2, -2.3 and -2.4 homeoboxes: evolutionary and structural comparisons. *Genomics* **1** 182-195

Harvey R. P. and Melton D. A., 1988. Microinjection of synthetic *XHox*-1A messenger-RNA disrupts somite formation in developing *Xenopus* embryos. *Cell* **53** 687-697

Hatzopoulos *et al.*, 1988 in Transcription and Splicing, B.D. Hames & D.M. Glover (eds) Oxford: IRL Press

Hauser C. A. *et al.*, 1985. Expression of homologous homeo-box-containing genes in differentiated human teratocarcinoma cells and mouse embryos. *Cell* **43** 19-28

Hayashi S. and Scott M. P., 1990. What determines the specificity of action of *Drosophila* homeodomain proteins? *Cell* **63** 883-894

Henikoff, S., 1984. Unidirectional digestion with exonuclease III creates targetted breakpoints for DNA sequencing. *Gene* **28** 351-359

Henry J. J. and Grainger R. M., 1990. Early tissue interactions leading to embryonic lens formation in *Xenopus laevis*. *Developmental Biology* **141** 149-163

Herbomel P. *et al.*, 1983. Structure and function of the promoter enhancer region of polyoma and SV40. *Molecular Biology Reports* **9** 153

Herr W. and Clarke J., 1986. The SV40 enhancer is composed of multiple functional elements that can compensate for one another. *Cell* **45** 461-470

Herschbach B. M. and Johnson A. D., 1993. Transcriptional repression in eukaryotes. *Annual review of Cell Biology* **9** 479-509

Hewitt J.E. *et al.*, 1991. Structure and sequence of the human Homeobox gene-HOX7. *Genomics* **11** 670-678

Hill R. E. *et al.*, 1989. A new family of mouse homeo box-containing genes - molecular structure, chromosomal location, and developmental expression of *Hox*-7.1. *Genes and Development* **3** 26-37

Hinnebusch A. G. and Liebman S. W., 1991. Protein synthesis and translational control in Saccharomyces cerevisiae. in *Volume I. Molecular and Cellular Biology of the Yeast Saccharomyces: Genome Dynamics, Protein Synthesis and Energetics.* eds Broach J. R., Pringle J. R. & Jones E. W. Cold Spring Harbor Laboratory Press p. 627

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

Holland P. W. H., 1991. Cloning and Evolutionary analysis of Msh-like Homeobox genes from mouse. zebrafish and ascidian. *Gene* **98** 253-257

Holland P., 1992. Homeobox genes in Vertebrate evolution. *Bioessays* **14** 267-273

Holland P. W. H. and Hogan B. L. M., 1986. Phylogenetic distribution of *Antennapedia*-like homeo boxes. *Nature* **321** 251-253

Holland P. *et al.*, 1992a. Development and Evolution - Mice And Flies head to head. *Nature* **358** 627-628

Holland P. W. H. *et al.*, 1992b. An Amphioxus homeobox gene - Sequence conservation, spatial expression during development and insights into Vertebrate evolution. *Development* **116** 653-661

Hollenberg S. M. and Evans R. M., 1988. Multiple and cooperative trans-activation domains of the human glucocorticoid receptor. *Cell* **55** 899-906

Hood L. *et al.*, 1992. Model genomes: the benefits of analysing homologous human and mouse sequences. *Trends in Biotechnology* **10** 19-22

Hornbruch, A. and Wolpert, L., 1986. Positional signalling by Hensen's node when grafted to the chick limb bud. *Journal of Embryology and Experimental Morphology* **94** 257-265.

Huang J. D. *et al.*, 1993. The interplay between multiple enhancer and silencer elements defines the pattern of *decapentaplegic* expression. *Genes and Development* **7** 694-704

Imagawa M. *et al.*, 1987. Transcription factor AP-2 mediates induction by two different signal-transduction pathways: protein kinase C and cAMP. *Cell* **51** 251-260

Imler J. L. *et al.*, 1987. Negative regulation contributes to tissue specificity of the immunoglobulin heavy-chain enhancer. *Molecular and Cellular Biology* **7** 2558-2567

Ingham P. W., 1988. The molecular genetics of embryonic pattern formation in *Drosophila. Nature* **355** 25-34.

Ingraham H. A. *et al.*, 1988. A tissue-specifc transcription factor containing a homeodomain specifies a pituitary phenotype. *Cell* **55** 519-529

Izpisúa-Belmonte J.-C. *et al.*, 1991. Expression of the homeobox *Hox*-4 genes and the specificiation of position in chick wing development. *Nature* **350** 585-589.

Izpisúa-Belmonte J.-C. *et al.*, 1992. Expression of *Hox*-4 genes in the chick wing links pattern formation to the epithelial-mesenchymal interactions that mediate growth. *EMBO Journal* **11** 1451-1458.

Jabs, E.W. *et al.*, 1993. A mutation in the homeodomain of the human MSX2 gene in a family affected with autosomal-dominant Craniosynostosis. *Cell* **75** 443-450

Jackson I. J. *et al.*, 1994. Genetics and Molecular Biology of mouse pigmentation. *Pigment Cell Research* **7** 73-81

Jacob F. and Monod J., 1961. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology* **3** 318-356

Jantzen H. M. *et al.*, 1987. Cooperativity of glucocorticoid response elements located far upstream of the tyrosine aminotransferase gene. *Cell* **49** 29-38

Jaynes J. B. and O'Farrell P. H., 1988. Activation and repression of transcription by homeodomain-containing proteins that bind a common site. *Nature* **336** 744-749

Jegalian B. G. And De Robertis E. M., 1992. Homeotic transformations in the mouse induced by overexpression of a Human HOX3.3 transgene. *Cell* **71** 901-910

Jiang J. and Levine M., 1993. Binding affinities and co-operative interactions with bHLH activators delimit threshold responses to the *dorsal* gradient morphogen. *Cell* **72** 741-751

Jiang J. *et al.*, 1993. Conversion of a *dorsal*-dependent silencer into an enhancer: evidence for *dorsal* corepressors. *EMBO Journal* **12** 3201-3209

Johnson P. F. *et al.*, 1987. Identification of a rat liver nuclear protein that binds to the enhancer core element of three animal viruses. *Genes and Development* **1** 133-146

Johnston M. and Dover J., 1988. Mutational analysis of the GAL4-encoded transcriptional activator protein of *Saccharomyces cerevisiae*. *Genetics* **120** 63-74

Jowett A. K. *et al.*, 1993. Epithelial-mesenchymal interactions are required for Msx-1 and Msx-2 gene-expression in the developing murine molar tooth. *Development* **117** 461-470

Joyner A. L. *et al.*, 1985. Expression during embryogenesis of a mouse gene with sequence homology to the *Drosophila engrailed* gene. *Cell* **43** 29-37

Kadonaga J. T. and Tjian R., 1986. Affinity purification of sequence specific DNA-binding proteins. *Proceedings of the National Academy of Sciences of the U.S.A* **83** 5889-5893

Kadonaga J. T. *et al.*, 1987. Isolation of cDNA encoding transcription factor Sp1 and functional analysis of the DNA binding domain. *Cell* **51** 1079-1090

Kaufman T. C. *et al.*, 1980. Cytogenetic analysis of chromosome 3 in *Drosophila melanogaster*: the homeotic gene complex in polytene chromosome interval 84A-B. *Genetics* **94** 115-133

Keegan L. *et al.*, 1986. Separation of DNA binding from the transcription-activating function of a eukaryotic regulatory protein. *Science* **231** 699-704

Kennedy G. C. and Rutter W. J., 1992. Pur-1, a zinc-finger protein that binds to purine-rich sequences, transactivates an insulin promoter in heterologous cells. *Proceedings of the National Academy of Sciences of the U.S.A.* **89** 11498-11502

Kenyon C. and Wang B., 1991. A cluster of Antennapedia-class homeobox genes in a nonsegmented animal. *Science* **253** 516-517

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

Kessel M. *et al.*, 1990. Variations of cervical-vertebrae after expression of a Hox-1.1 transgene in mice. *Cell* **61** 301-308

Kessel M. And Gruss P., 1991. Homeotic transformations of murine vertebrae and concomitant alteration of Hox codes induced by Retinoic Acid. *Cell* **67** 89-104

Kim J. L. *et al.*, 1993a. Co-crystal structure of TBP recognizing the minor groove of a TATAelement. *Nature* **365** 520-527

Kim Y. *et al.*, 1993b. Crystal structure of a yeast TBP/TATA-box complex. *Nature* **365** 512-520

Kimura A. *et al.*, 1993. Chicken ovalbumin upstream promoter-transcription factor (COUP-TF) represses transcription from the promoter of the gene for ornithine transcarbamylase in a manner antagonistic to hepatocyte nuclear factor-4 (HNF-4). *Journal of Biological Chemistry* **15** 11125-11133

Kingston R. E. *et al.*, 1987. Heat-inducible human factor that binds to a human hsp70 promoter. *Molelular and Cellular Biology* **7** 1530-1534

Kirov N. *et al.*, 1993. Conversion of a silencer into an enhancer - evidence for a co-repressor in dorsal-mediated repression on *Drosophila*. *EMBO Journal* **12** 3193-3199

Kissinger C. R. *et al.*, 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

Klug A., 1993. Opening the gateway *Nature* **365** 486-487

Kollias G. *et al.*, 1987. The human beta-globin gene contains a downstream developmental specific enhancer. *Nucleic Acids Research* **15** 5739-5747

Koop B. F. and Hood L., 1994. Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. *Nature Genetics* **7** 48-53

Krasnow M. A. *et al.*, 1989. Transcriptional activation and repression by Ultrabithorax proteins in cultured *Drosophila* cells. *Cell* **57** 1031-1043

Kress *et al.* C., 1990. *Hox*-2.3 upstream sequences mediate LacZ expression in intermediate mesoderm derivatives of transgenic mice. *Development* **109** 775

Krieg P. A. and Melton D. A., 1987. An enhancer responsible for activating transcription at the midblastula transition in *Xenopus* development. *Proceedings of the National Academy of Sciences of the U.S.A.* **84** 2331-2335

Krumlauf R., 1992. Evolution of the vertebrate *Hox* homeobox genes. *Bioessays* **14** 245-252

Krumlauf R., 1993. *Hox* genes and pattern-formation in the branchial region of the vertebrate head. *Trends in Genetics* **9** 106-112

Kuhl D. *et al.*, 1987. reversible silencing of enhancers by sequences derived from the human IFN-alpha promoter. *Cell* **50** 1057-1069

Kuner J. M. *et al.*, 1985. Molecular-cloning of *engrailed* - a gene involved in the development of pattern in *Drosophila melanogaster*. *Cell* **42** 309-316

Kuzoira M. A. and McGinnis W., 1989. A homeodomain substitution changes the regulatory specificity of the *Deformed* protein in *Drosophila* embryos. *Cell* **59** 563-571

Kuzuoka M. *et al.*, 1994. Murine homeobox-containing gene, *Msx*-1 - Analysis of Genomic organization, promoter structure, and potential autoregulatory *cis*-acting elements. *Genomics* **21** 85-91

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

Lammer E. J. *et al.*, 1985. Retinoic Acid embryopathy. *New England Journal of Medicine* **313** 837-841

Lane D. *et al.*, 1992. Use of Gel Retardation to analyze protein-nucleic acid interactions *Microbiological Reviews* **56** 509-528

Laney J. D. and Biggin M. D., 1992. *zeste*, a nonessential gene, potently activates *Ultrabithorax* transcription in the *Drosophila* embryo. *Genes and Development* **6** 1531-1541

Larsen F. *et al.*, 1992. CpG islands as gene-markers in the Human genome. *Genomics* **13** 1095-1107

Laudet V. *et al.*, 1993. Ancestry and diversity of the HMG-box superfamily. *Nucleic Acids Research* **21** 2493-2501

Laughon A. and Scott M. P., 1984. Sequence of a *Drosophila* segmentation gene: protein structure homology with DNA-binding proteins. *Nature* **310** 25-31

Le Douarin, 1982. The Neural Crest. Cambridge University Press

Le Mouellic H. *et al.*, 1992. Homeosis in the mouse induced by a null mutation in the *Hox*-3.1 gene. *Cell* **69** 251-264

Lee H. -S. *et al.*, 1992. DNA sequence requirements for generating paused polymerase at the start of hps70. *Genes and Development* **6** 284-295

Leiden J. M., 1993. Transcriptional regulation of T cell receptor genes. *Annual Review of Immunology* **11** 539-570

Lenormand J. L. *et al.*, 1993. Identification of a *cis*-acting element responsible for muscle-specific expression in the c-*mos* proto-oncogene. *Nucleic Acids Research* **21** 695-702

Leuther K. K. *et al.*, 1993. Genetic evidence that an activation domain of GAL4 does not require acidity and may form a β sheet. *Cell* **72** 575-585

Levine M. *et al.*, 1984. Human DNA-sequences homologous to a protein coding region conserved between homeotic genes of *Drosophila*. *Cell* **38** 667-673

Lewis E. B., 1978. A gene complex controlling segmentation in *Drosophila*. *Nature* **276**, 565-570

Lockett T. J. and Ashburner M., 1989. Temporal and spatial utilisation of the alcohol dehydrogenase gene promoters during the development of *Drosophila melanogaster*. *Developmental Biology* **134** 430-437

Lu Q. *et al.*, 1993. $(CT)_n(GA)_n$ repeats and heat shock elements have distinct roles in chromatin structure and transcriptional activation of the *Drosophila hsp26* gene. *Molecular and Cellular Biology* **13** 2802-2814

Luckow B. and Schütz G., 1987. CAT constructions with multiple unique restriction sites for the functional analysis of eukaryotic promoters and regulatory elements. *Nucleic Acids Research* **15** 5490

Lufkin T. *et al.*, 1991. Disruption of the *Hox*-1.6 homeobox gene results in defects in a region corresponding to its rostral domain of expression. *Cell* **66** 1105-1119

Lumpkin O. J. and Zimm B. H., 1982. Mobility of DNA in gel-electrophoresis. *Biopolymers* **21** 2315-2316

Lumsden A. and Keynes R., 1989. Segmental patterns of neuronal development in Chick hindbrain. *Nature* **337** 424-428

Lüscher B. *et al.*, 1989. Regulation of the transcription factor AP-2 by the morphogen Retinoic Acid and by second messengers. *Genes and Development* **3** 1507-1517

Lyons G. E. *et al.*, 1992. Multiple sites of *Hox*-7 expression during mouse embryogenesis - comparison with Retinoic Acid Receptor messenger-RNA localization. *Molecular Reproduction and Development* **32** 303-314

Ma H. *et al.*, 1991. AGL1-AGL6, an Aribodopsis gene family with similarity to floral homeotic and transcription factor genes. *Genes and Development* **5** 484-495

MacKenzie A. *et al.*, 1991a. The homeobox gene *Hox*-7.1 has specific regional and temporal expression patterns during early murine embryogenesis, especially tooth development *in vivo* and *in vitro*. *Development* **111** 269-285

Mackenzie A. *et al.*, 1991b. *Hox*-7 expression during murine craniofacial development. *Development* **113** 601-611

Mackenzie A. *et al.*, 1992. Expression patterns of the homeobox gene. *Hox*-8. in the mouse embryo suggest a role in specifying tooth initiation and shape. *Development* **115** 403-420

Malicki J. *et al.*, 1990. Mouse *Hox*-2.2 specifies thoracic segmental identity in *Drosophila* embryos and larvae. *Cell* **63** 961-967

Malicki J. *et al.*, 1992. A human HOX4B regulatory element provides head-specific expression on *Drosophila* embryos. *Nature* **358** 345-347

Maniatis T. *et al.*, 1987. Regulation of inducible and tissue-specific gene expression. *Science* **236** 1237-1245

Mantovani R. *et al.*, 1988. An erythroid specific nuclear factor binding to the proximal CACCC box of the beta-globin gene promoter. *Nucleic Acids Research* **16** 4299-4313

Marini J. C. *et al.*, 1982. Bent helical structure in kinetoplast DNA. *Proceedings of the National Academy of Sciences of the U.S.A.* **79** 7664-7668

Martin G. R., 1981. Isolation of a pluripotent cell line from early mouse embryos cultured in a medium conditioned by teratocarcinoma stem cells. *Proceedings of the National Academy of Sciences of the U.S.A.* **78** 7634-7638

Mastrangelo I. A. *et al.*, 1991. DNA looping and Sp1 multimer links - a mechanism for transcriptional synergism and enhancement. *Proceedings of the National Academy of Sciences of the U.S.A.* **88** 5670-5674

Matsuo I. *et al.*, 1991. Binding of a factor to an enhancer element responsible for the tissue-specific expression of the chicken alphaA-crystallin gene. *Development* **113** 539-550

McGinnis N. *et al.*, 1990. Human *Hox*-4.2 and *Drosophila Deformed* encode similar regulatory specificities in *Drosophila* embryos and larvae. *Cell* **63** 969-976

McGinnis W. *et al.*, 1984a. A conserved DNA-sequence in homeotic genes of the *Drosophila* Antennapedia and bithorax complexes. *Nature* **308** 428-433

McGinnis W. *et al.* ,1984b. A homologous protein-coding sequence in *Drosophila* homeotic genes and its conservation in other metazoans. *Cell* **37** 403-408

McGinnis W. *et al.*, 1984c. Molecular cloning and chromosom mapping of a mouse DNA sequence homologous to homeotic genes of *Drosophila*. *Cell* **38** 675-680

McGinnis W., 1985. Homeo box sequences of the Antennapedia class are conserved only in higher animal genomes. *Cold Spring Harbor Symposium of Quantitative Biology* **50** 263-270

McGinnis W. and Krumlauf R., 1992. Homeobox genes and axial patterning. *Cell* **68** 283-302

McKnight S. and Tjian R., 1986. Transcriptional selectivity of viral genes in mammalian cells. *Cell* **46** 795-805

McLain K. *et al.*, 1992. Ectopic expression of *Hox*-2.3 induces craniofacial and skeletal malformations in transgenic mice. *Mechanisms of Development* **39** 3-16

Mercola M. *et al.*, 1983. Transcriptional enhancer elements in the mouse Immunoglobulin heavy chain ocus. *Science* **221** 663-665

Merika M. and Orkin S. H., 1993. DNA-binding specificity of GATA family transcription factors. *Molecular and Cellular Biology* **13** 3999-4010

Mermod N. *et al.*, 1989. The proline-rich transcriptional activator of CTF/NF-1 is distinct from the replication and DNA binding domain. *Cell* **58** 741-753

Meyer B. I. and Gruss P., 1993. Mouse *Cdx*-1 expression during gastrulation. *Development* **117** 191-203

Miller J. *et al.*, 1985. Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. *EMBO Journal* **4** 1609-1614

Miskimins W. K. *et al.*, 1985. Use of a protein-blotting procedure and a specific DNA probe to identify nuclear proteins that recognize the promoter region of the Transferrin receptor gene. *Proceedings of the National Academy of Sciences of the U.S.A.* **82** 6741-6744

Mitchell P. J. *et al.*, 1991. Transcription factor AP-2 is expressed in neural crest cell lineages during mouse embryogenesis. *Genes and Development* **5** 105-119

Mitchell P. J. and Tjian R., 1989. Transcriptional regulation in mammalian cells by sequence-specific DNA-binding proteins. *Science* **245** 371-378

Mlodzik M. and Gehring W. J., 1987. Expression of the *caudal* gene in the germ line of *Drosophila*: formation of an RNA and protein gradient during early embryogenesis. *Cell* **48** 465-478

Molven A. *et al.*, 1990. Expression of a homeobox gene-product in normal and mutant Zebrafish embryos - Evolution of the tetrapod body plan. *Development* **109** 279

Molven A. *et al.*, 1991. Genomic structure and restricted neural expression of the zebrafish wnt-1 (int-1) gene. *EMBO Journal* **10** 799-807

Monaghan A. P. *et al.*, 1991. The Msh-like homeobox genes define domains in the developing vertebrate eye. *Development* **112** 1053-1061

Monaghan A. P. *et al.*, 1993. Postimplantation expression patterns indicate role for the mouse Forkhead/HNF-3 alpha, beta and gamma genes in determination of the definitive endoderm, chordamesoderm and neuroectoderm. *Development* **119** 567-578

Morgan B.A. *et al.*, 1992. Targeted misexpression of *Hox*-4.6 in the avian limb bud causes apparent homeotic transformations. *Nature* **358** 236-239.

Müller M. M. *et al.*, 1984. A homeo-box-containing gene expressed during oogensis in *Xenopus*. *Cell* **39** 157-162

Müller H.-P. *et al.*, 1989. An enhancer stimulates transcription in *trans* when attached to the promoter *via* a protein bridge. *Cell* **58** 767 (erratum published *Cell* 1989 **59** 405)

Munaim S. I. *et al.*, 1988. Developmental changes in fibroblast growth factor in the chick embryo limb bud. *Proceedings of the National Academy of Sciences of the U.S.A.* **85** 8091-8093

Murphy P. *et al.*, 1989. Segment-specific expression of a homeobox-containing gene in the mouse hindbrain. *Nature* **341** 156-159

Murre C. *et al.*, 1989a. A new DNA-binding and dimerization motif in immunoglobulin enhancer binding, *daughterless*, MyoD and *myc* proteins. *Cell* **56** 777-783

Murre C. *et al.*, 1989b. Interactions between heterologous helix-loop-helix proteins generate complexes that bind specifically to a common DNA-sequence. *Cell* **58** 537-544

Myers R. M. *et al.*, 1986. Fine structure genetic analysis of a beta-globin promoter. *Science* **232** 613-618

Neuhaus D. *et al.*, 1992. Solution structures from two zinc-finger domains from SWI5 obtained using two-dimensional 1H nuclear magnetic resonance spectroscopy. A zinc-finger structure with a third strand of beta-sheet. *Journal of Molecular Biology* **228** 637-651

Niswander L. and Martin G.R., 1992. FGF-4 expression during gastrulation, myogenesis, limb and tooth development in the mouse. *Development* **114** 755-768

Niswander L and Martin G. R., 1993a. FGF-4 regulates expression of *Evx*-1 in the developing mouse limb. *Development* **119** 287-294

Niswander L *et al.*, 1993b. FGF-4 replaces the apical ectodermal ridge and directs outgrowth and patterning of the limb. *Cell* **75** 579-587.

Noden D. M., 1988. Interactions and fates of avian craniofacial mesenchyme. *Development* **103** (Supplement) 121-140

Nomenclature Comittee of the International Union of Biochemists, 1985. European Journal of Biochemistry. **150** 1-5

Ohkuma Y. *et al.*, 1990. *engrailed*, a homeodomain protein, can repress *in vitro* transcription by competition with the TATA box-binding protein transcription factor-IID. *Proceedings of the National Academy of Sciences of the U.S.A.* **87** 2289-2293

Ondek B. *et al.*; 1988. The SV40 enhancer contanis two distinct levels of organization. *Nature* **333** 40-45

Otting G. *et al.*, 1990. Protein DNA contacts in the structure of a homeodomain DNA complex determined by Nuclear-Magnetic-Resonance spectroscopy in solution. *EMBO Journal* **9** 3085-3092

Pabo C. O. and Sauer R. T., 1984. Protein-DNA recognition. *Annual Review of Biochemistry.* **53** 293-321

Parr B.A. *et al.*, 1993. Mouse *Wnt* genes exhibit discrete domains of expression in the early embryonic CNS and limb buds. *Development* **119** 247-261

Pevny L. *et al.*, 1991. Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. *Nature* **349** 257-260

Poole S. J. *et al.*, 1985. The *engrailed* locus of *Drosophila*: structural analysis of an embryonic transcript. *Cell* **40** 37-43

Prestridge D. S., 1991. SIGNAL SCAN: a computer program that scans DNA sequences for transcriptional elements. *CABIOS* **7** 203-206

Prestridge D. S. and Stormo G., 1993. SIGNAL SCAN 3.0: new database and program features. *CABIOS* **9** 113-115

Ptashne M., 1986. Gene regulation by proteins acting nearby and at a distance. *Nature* **322** 697-701

Ptashne M., 1988. How eukaryotic transcripitonal activators work. *Nature* **335** 683-689

Pugh B. and Tjian R., 1991. Transcription from a TATA-less promoter requires a multisubunit TFIID complex. *Genes and Development* **5** 1935-1945

Püschel A. W. *et al.*, 1991. Separate elements cause lineage restriction and specify boundaries of *Hox*-1.1 expression. *Development* **112** 279-287

Qian Y. Q. *et al.*, 1989. The structure of the *Antennapedia* homeodomain determined by NMR-spectroscopy in solution - comparison with prokaryotic repressors. *Cell* **59** 573-580

Qian S. *et al.*, 1991. The *bx* region enhancer, a distant *cis*-control element of the *Drosophila Ubx* gene and its regulation by *hunchback* and other segmentation genes. *EMBO Journal* **10** 1415-1425

Quong M. W. *et al.*, 1993. A new transcriptional-activation motif restricted to a class of helix-loop-helix proteins is functionally conserved both yeast and mammalian cells. *Molecular and Cellular Biology* **13** 792-800

Ramsay R. G. *et al.*, 1991. Increase in specific DNA binding by carboxyl truncation suggests a mechanism for activation of Myb. *Oncogene* **6** 1875-1879

Ramírez-Solis R. *et al.*, 1993. *Hox*b-4 (*Hox*-2.6) mutant mice show homeotic transformation of a cervical vertebra and defects in the closure of the sternal rudiments. *Cell* **73** 279-294

Ransone L. J. *et al.*, 1990. Domain swapping reveals the modular nature of *fos*, *jun* and CREB proteins. *Molecular and Cellular Biology* **10** 4565-4573

Rastinejad F. and Blau H. M., 1993. Genetic complementation reveals a novel regulatory role for 3' untranslated regions in growth and differentiation. *Cell* **72** 903-917

Renucci A. *et al.*, 1992. Comparison of mouse and human HOX-4 complexes defines conserved sequences involved in the regulation of *Hox*-4.4. *EMBO Journal* **11** 1459-1468

Richman J. M. and Tickle C., 1989. Epithelia are interchangeable between facial primordia of chick embryos and morphogenesis is controlled by the mesenchyme. *Developmental Biology* **136** 201-210

Richman J. M. And Tickle C., 1992. Epithelial Mesenchymal interactions in the outgrowth of limb buds and facial primordia in chick-embryos. *Developmental Biology* **154** 299-308

Rigby P. W. *et al.*, 1977. Labelling deoxyribonucleic acid to high specific activity *in vitro* by nick translation with DNA Polymerase I. *Journal of Molecular Biology* **113** 237-251

Rivera V. M. *et al.*, 1993. A growth facor-induced kinase phosphorylates the serum repsonse factor at a site that regulates its DNA-binding activity. *Molecular and Cellular Biology* **13** 6260-6273

Robert B. *et al.*, 1989. *Hox*-7, a mouse homeobox gene with a novel pattern of expression during embryogenesis. *EMBO Journal* 8 91-100

Robert B. *et al.*, 1991. The Apical Ectodermal Ridge regulates *Hox*-7 and *Hox*-8 gene-expression in developing chick limb buds. *Genes and Development* **5** 2363-2374

Roeder R. G., 1991. The complexities of eukaryotic transcription initiation: regulation of preinitiation complex assembly. *Trends in Biochemical Sciences* **16** 402-408

Ros M. A. *et al.*, 1992. Apical Ridge dependent and independent mesodermal domains of Ghox-7 and Ghox-8 expression in chick limb buds. *Development* **116** 811-818

Rosenfeld M. G., 1991. POU domain transcription factors: pou-er-ful developmental regulators. *Genes and Development* **5** 897-907

Ruiz i Altaba A. and Jessel T., 1991. Retinoic acid modifies mesodermal patterning in early *Xenopus* embryos. *Genes and Development* **5** 175-187

Saitou N. and Nei M., 1987. The Neighbour-joining method: A new method for reconstructing phylogentic trees. *Molecular Biology and Evolution* **4** 406-425

Sambrook J. *et al.*, 1989. Molecular cloning, a laboratory manual: 2nd Edition. Cold Spring Harbor Laboratory Press.

Sanger F. *et al.*, 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the U.S.A.* **74** 5463-5467

Santoro C. *et al.*, 1988. A family of human CCAAT-box-binding proteins active in transcription and DNA replication: cloning and expression of multiple cDNAs. *Nature* **334** 218-224

Satokata I.and Maas R., 1994. *Msx*-1 deficient mice exhibit claft palate and abnormalities of craniofacial and tooth development. *Nature Genetics* **6** 348-355

Saudek V. *et al.*, 1991. The solution structure of a leucine-zipper motif peptide. *Protein Engineering* **4** 519-529

Saunders J. W. Jr, 1948. The proximo-distal sequence of origin of the parts of the chick wing and the role of the ectoderm. *Journal of Experimental Zoology* **108** 363-403

Saunders, 1980 *Developmental Biology*. Macmillan New York

Saxén L. and Sariola H., 1987. Early organogenesis of the Kidney. *Pediatric Nephrology* **1** 385-392

Schaffner W., 1989. How do different transcription factors binding the same DNA sequence sort out their jobs? *Trends in Genetics* **5** 37-39

Schibler U. and Sierra F., 1987. Alternative promoters in devlopmental gene-expression. *Annual Review of Genetics* **21** 237-257

Schierwater B. *et al.*, 1991. Homeoboxes in Cnidarians. *Journal of Experimental Zoology* **260** 413-416

Schneuwly S. *et al.*, 1986. Structural organization and sequence of the homeotic gene *Antennapedia* of *Drosophila melanogaster*. *EMBO Journal* **5** 733-739

Schneuwly S. *et al.*, 1987a. Molecular analysis of the dominant homeotic *Antennapedia* phenotype. *EMBO Journal* **6** 201-206

Schneuwly S. *et al.*, 1987b. Redesigning the body plan of *Drosophila* by ectopic expression of the homeotic gene *Antennapedia*. *Nature* **325** 816-818

Schubert F. R. *et al.*, 1993. The *Antennapedia*-type homeobox genes have evolved form 3 precursors separated early in metazoan evolution. *Proceedings of the National Academy of Sciences of the U.S.A.* **90** 143-147

Schugart K. *et al.*, 1989. Duplication of large genomic regions during the evolution of vertebrate homeobox genes. *Proceedings of the National Academy of Sciences of the U.S.A.* 86 7067-7071

Scott M. P., 1983. The molecular organization of the *Antennapedia* locus of *Drosophila*. Cell 35 763-776

Scott M. P. and Weiner A. J., 1984. Structural relationships between genes that control development: sequence homology between the *Antennapedia*, *Ultrabithorax* and *fushi tarazu* loci of *Drosophila*. *Proceedings of the National Academy of Sciences of the U.S.A.* **81** 4115-4119

Scott M. P. *et al.*, 1990. The structure and function of the homeodomain. *Biochimica et Biophysica Acta - Reviews on Cancer* **989** 25-48

Scott M. P., 1992. Vertebrate homeobox gene nomenclature. *Cell* **71** 551-553

Seed B. and Sheen J.-Y., 1988. A simple phase-extraction assay for chloramphenicol acyltransferase activity. *Gene* **67** 271-277

Sen R. and Baltimore D., 1986. Multiple nuclear factors interact with the immunoglobulin enhancer sequences. *Cell* **46** 705-716

Serfling E. *et al.*, 1985. Enhancers and eukaryotic gene transcription. *Trends in Genetics* **1** 224-230

Sham M. H. *et al.*, 1992. Analysis of the murine *Hox*-2.7 gene: conserved alternative transcripts with differential distributions on the nervous system and the potential for shared regulatory regions. *EMBO Journal* **11** 1825-1836

Shehee W. R. *et al.*, 1989. Nucleotide sequence of the BALB/c mouse β-globin complex. *Journal of Molecular Biology* **205** 41-62

Shepherd J. C. W. *et al.*, 1984. Fly and Frog homoeo domains show homologies with yeast mating type regulatory proteins. *Nature* **310** 70-71

Sigler P. B., 1988. Acid blobs and negative noodles. *Nature* **333** 210-212

Simeone A. *et al.*, 1990. Sequential activation of HOX2 homeobox genes by Retinoic Acid in human Embryonal Carcinoma-cells. *Nature* **346** 763-766

Singh H. *et al.*, 1988. Molecular cloning of an enhancer binding protein: isolation by screening of an expression library with a recognition site DNA. *Cell* **52** 415-423

Smith D. L. and Johnson A. D., 1992. A molecular mechanism for combinatorial control in yeast - MCM1 protein sets the spacing and orientation of the homeodomains of an α-2 dimer. *Cell* **68** 133-142

Soeller W. C. *et al.*, 1993. Isolation of cDNAs encoding the *Drosophila* GAGA transcription factor. *Molecular and Cellular Biology* **13** 7961-7970

Somma P. *et al.*, 1990. The housekeeping promoter from the mouse CpG island HTF9 contains multiple protein-binding elements that are functionally redundant. *Nucleic Acids Research* **19** 2817-2824

Song K. *et al.*, 1992. Expression of *Hox*-7.1 in myoblasts inhibits terminal differentiation and induces cell transformation. *Nature* **360** 477-481

Southern E. M., 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. Journal of Molecular Biology **98** 503-517

Strathern J. N. *et al.*, 1981. (eds) The Molecular Biology of the yeast *Saccharomyces*: Life cycle and Inheritance. Cold Spring Harbor Laboratory Press.

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

Struhl G. *et al.*, 1989a. The gradient morphogen bicoid is a concentration dependent transcription factor. *Cell* **57** 1259-1273

Struhl K. 1989a. Helix-turn-helix, zinc-finger and leucine-zipper motifs for eukaryotic transcriptional regulatory proteins. *Trends in Biochemical Sciences* **14** 137-140

Stuart J. J. *et al.*, 1991. A defiency of the homeotic complex of the beetle *Tribolium*. *Nature* **350** 72-74

Sturm R. A. *et al.*, 1988. The ubiquitous octamer-binding protein Oct-1 contains a POU domain with a homeobox subdomain. *Genes and Development* **2** 1582-1599

Su W. *et al.*, 1991. DNA looping between sites for transcriptional activation: self-association of DNA-bound Sp1. *Genes and Development* **5** 820-826

Summerbell D. *et al.*, 1973. Positional information in chick limb morphogenesis. *Nature* **244** 492-496

Summerbell D., 1974. A quantitative analysis of the effect of excision of the AER from the chick limb bud. *Journal of Embryology and experimental Morphology* **32** 651-660

Sutherland J. A. *et al.*, 1992. Conserved motifs in Fos and Jun define a new class of activation domain. *Genes and Development* **6** 1810-1819

Suzuki T. and Suzuki Y., 1988. Interaction of composite protein complex with the fibroin enhancer sequence. *Journal of Biological Chemistry* **263** 5979-5986

Szyf M. *et al.*, 1990. A DNA signal from the Thy-1 gene defines *de novo* methylation patterns in embryonic stem cells. *Molecular and Cellular Biology* **10** 4396-4400

Tagawa M. *et al.*, 1990. Expression of a novel DNA-binding protein with zinc-finger structure in various tumor cells. *Journal of Biological Chemistry* **265** 20021-20026

Takahashi Y. and Le Douarin N., 1990. cDNA cloning of a quial homeobox gene and its expression in neural crest-derived mesenchyme and lateral plate mesoderm. *Proceedings of the National Academy of Sciences of the U.S.A.* **87** 7482-7486

Tamarin A. *et al.*, 1984. Analysis of upper beak defects in chicken embryos following treatment with retinoic acid. *Journal of Embryology and Experimental Morphology* **84** 105-123

Tartof K. D. and Hobbs C. A., 1987. Improved media for growing plasmid and cosmid clones. *Focus (Bethesda Research Laboratories Inc)* **9:2** 12

Tasset D. *et al.*, 1990. Distinct classes of transciptional activation domains function by different mechanisms. *Cell* **62** 1177-1187

Tautz D., 1988. Regulation of the *Drosophila* segmentation gene *hunchback* by two maternal morphogenetic centres. *Nature* **332** 281-284

Thaller C. and Eichele G., 1987. Identification and spatial distribution of retinoids in the developing chick limb bud. *Nature* **327** 625-628

Thisse C. and Thisse B., 1992. Dorso-ventral development of the *Drosophila* embryo is controlled by a cascade of transcriptional regulators. *Development* Supplement 173

Thompson M. A. *et al.*, 1992. Nerve growth factor-induced derepression of peripherin gene-expression is associated with alterations in proteins binding to a negative regulatory element. *Molecular and Cellular Biology* **12** 2501-2513

Tickle C. *et al.*, 1975. Positional signalling and specification of digits in chick limb morphogenesis. *Nature* **254** 199-202

Tickle C. *et al.*, 1982. Local application of Retinoic acid in the limb bud mimics the action of the polarizing region. *Nature* **296** 564-566

Towbin H. *et al.*, 1979. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: Procedure and some applications. *Proceedings of the National Academy of Sciences of the U.S.A.* **76** 4350-4354

Treisman J. *et al.*, 1989. A single amino-acid can determine the DNA-binding specificity of homeodomain proteins. *Cell* **59** 553-562

Treisman J. *et al.*, 1991. The paired box encodes a second DNA-binding domain in the paired homeodomain protein. *Genes and Development* **5** 594-604

Treisman J. *et al.*, 1992a. The homeodomain - A new face for the Helix-turn-Helix. *Bioessays* **14** 145-150

Treisman R., 1990. The SRE: a growth factor responsive transcriptional regulator. *Seminars in Cancer Biology* **1** 47-58

Treisman R., 1992b. The Serum Response Element. *Trends in Biochemical Sciences* **17** 423-426

Treisman R. *et al.*, 1992c. Spatial flexibility in ternary complexes between SRF and its accessory proteins. *EMBO Journal* **11** 4631-4640

Tsukiyama T. *et al.*, 1994. ATP-dependent nucleosome disruption at a heat-shock promoter mediated by binding of GAGA transcription factor. *Nature* **367** 525-532

Usuda Y. *et al.*, 1991. Affinity purification of transcription factor IIA from HeLa cell nuclear extracts. *EMBO Journal* **10** 2305-2310

Vainio S. *et al.*, 1993. Identification of BMP-4 as a signal mediating secondary induction between epithelial and mesenchymal tissues during early tooth development. *Cell* **75** 45-58

van Assenfeldt G. B. *et al.*, 1989. The β-Globin dominant control region activates homologous and heterologous promoters in a tissue-specific manner. *Cell* **56** 969-977

van den Endt F. M. *et al.*, 1994. Cell-cycle controlled histone H1, H3 and H4 genes share unusual arrangements of recognition motifs for HiNF-D supporting a coordinate promoter binding mechanism. *Journal of Cell Physiology* **159** 515-530

van Holde K., 1994. Properly preparing promoters. *Nature* **367** 512-513

Vasseur-Cognet M. and Lane M. D., 1993. CCAAT/enhancer binding protein alpha (C/EBP alpha) undifferentiated protein: a developmentall regulated nuclear protein that binds to the C/EBP alpha gene promoter. *Proceedings of the National Academy of Sciences of the U.S.A.* **90** 7312-7316

Verrijzer C. P. *et al.*, 1992. The DNA binding specificity of the bipartite POU domain and its subdomains. *EMBO Journal* **11** 4993-5003

Vershon A. K. and Johnson A. D., 1993. A short, disordered protein region mediates interactions between the homeodomain of the yeast α-2 protein and the MCM1 protein. *Cell* **72** 105-112

Vieira J. and Messing J., 1982. The pUC plasmids - an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. *Gene* **19** 259-269

Vieira J. and Messing J., 1987. Production of single-stranded plasmid DNA. *Methods in Enzymology* **153** 3

Vinson C. R. *et al.*, 1988. *In situ* detection of sequence-specific DNA-binding activity specified by a recombinant bacteriophage. *Genes and Development* **2** 801-806

Virbasius C. A. *et al.*, 1993. NRF-1, an activator involved in nuclear-mitochondrial interactions, utilizes a new DNA-binding domain conserved in a family of developmental regulators. *Genes and Development* **7** 2431-2445

Wahli W. and Martinez E., 1991. Superfamily of Steroid Nuclear receptors - positive and negative regulation of gene expression. *FASEB Journal* **5** 2243-2249

Walker M. D. *et al.*, 1983. Cell-specific expression controlled by the 5'-flanking region of the insulin and chymotrypsin genes. *Nature* **306** 557-561

Wall L. *et al.*, 1988. The human β-globin gene 3' enhancer contains multiple binding sites for an erythroid-specific protein. *Genes and Development* **2** 1089-1100

Walsh M. J. *et al.*, 1992. Characterization of DNA-Protein interactions within a distal regulatory element upstream of a mammalian housekeeping promoter. *Journal of Biological Chemistry* **267** 7026-7035

Wang B. B. *et al.*, 1993. A homeotic gene-cluster patterns the anteroposterior body axis of *C.elegans*. *Cell* **74** 29-42

Wedden, 1987. Epithelial mesenchymal interactions in the development of chick facial primordia and the target of retinoid action. *Development* **99** 341-351

Wedden *et al.*, 1989. Expression pattern of homeobox-containing genes during chick embryogenesis. *Development* **105** 639-650

Weigel D. and Jäckle H., 1990. The *forkhead* domain - a novel DNA-binding motif of eukaryotic transcription factors. *Cell* **63** 455-456

West M. *et al.*, 1992. Characterisation and purification of a novel transcriptional repressor from HeLa cell nuclear extracts recognising the negative regulatory element region of human immunodeficiency virus-1 long terminal repeat. *Journal of Biological Chemistry* **267** 24948-24952

Westerfield M. *et al.*, 1992. Specific activation of mammalian *Hox* promoters in mosaic transgenic zebrafish. *Genes and Development* **6** 591-598

Wharton R. P. and Ptashne M., 1985. Changing the binding-specificity of a repressor by redesigning an α-helix. *Nature* **316** 601-605

Whiting J. *et al.*, 1991. Multiple spatially specific enhancers are required to reconstruct the pattern of *Hox*-2.6 gene expression. *Genes and Development* **5** 2048-2059

Wilbur W. J. and Lipman D. J., 1983. Rapid similarity searches of nucleic acid and protein databanks. *Proceedings of the National Academy of Sciences of the U.S.A* **80** 726-730

Wilkinson D. G. *et al.*, 1989. Segmental expression of *Hox*-2 homeobox-containing genes in the developing mouse hindbrain. *Nature* **341** 405-409

Wilkinson D. G., 1993. Molecular mechanisms of segmental patterning in the vertebrate hindbrain and neural crest. *Bioessays* **15** 499-505

Williams G. T and Morimoto R. I., 1990 Maximal stress-induced transcription from the Human *HSP70* promoter requires interactions with the basal promoter elements independent of rotational alignment. *Molecular and Cellular Biology* **10** 3125-3136

Williams T. *et al.*, 1988. Cloning and expression of AP-2, a cell-type-specific transcription factor that activates inducible enhancer elements. *Genes and Development* **2** 1557-1569

Williams T. And Tjian R., 1991a. Analysis of the DNA-binding and activation properties of the human transcription factor AP-2. *Genes and Development* **5** 670-682

Williams T. And Tjian R., 1991b. Characterization of a dimerization motif in AP-2 and its function in heterologous DNA-binding proteins. *Science* **251** 1067-1071

Winslow G. M. *et al.*, 1989. Transcriptional activation by the *Antennapedia* and *fushi tarazu* proteins in cultured *Drosophila* cells. *Cell* **57** 1017-1030

Wolgemuth D. J. *et al.*, 1989. Transgenic mice overexpressing the mouse homeobox-containing gene *Hox*-1.4 exhibit abnormal gut development. *Nature* **337** 464-467

Wright C. V. E., 1989. Interference with function of a homeobox gene in *Xenopus* embryos produces malformations in the anterior spinal-cord. *Cell* **59** 81-93

Wynne J. And Treisman R., 1992. SRF And MCM1 have related but distinct DNA-binding specificities. *Nucleic Acids Research* **20** 3297-3303

Yasuda Y. *et al.*, 1986. Developmental abnormalities induced by all-trans retinoic acid in fetal mice. 1. Macroscopic findings. *Teratology* **34** 37-49

Yanish Perron C. *et al.*, 1985. Improved M13 phage cloning vectors and host strains: nucleotide sequence of the M13mp18 and pUC19 vectors. *Gene* **33** 103-119

Yokouchi Y. *et al.*, 1991. Chicken homeobox gene *Msx*-1: Structure, expression in limb buds and effect of Retinoic Acid. *Development* **113** 431-444 (Note: *Msx*-1 here is a chicken cognate of mouse *Msx*-2/*Hox*-8)

Zappavigna V. *et al.*, 1991. *Hox*-4 Genes encode transcription factors with potential auto-regulatory and cross-regulatory capacities. *EMBO Journal* **10** 4177-4188

Zhao J. J. *et al.*, 1993. The mouse *Hox*-1.3 gene is functionally equivalent to the *Drosophila Sex combs reduced* gene. *Genes and Development* **7** 343-354

Zon L. I. *et al.*, 1991. Expression of GATA-binding proteins during embryonic development in Xenopus laevis. *Proceedings of the National Academy of Sciences of the U.S.A.* **88** 10642-10646