THE STRUCTURE AND EVOLUTION OF A RODENT SERINE PROTEINASE INHIBITOR GENE COMPLEX.

JOHN DUGALD INGLIS

Ph.D

UNIVERSITY OF EDINBURGH

1990.
ACKNOWLEDGEMENTS

I would like to thank the following people for their assistance and patience. Its been a long time coming.....

Julia, without whom I couldn’t cope.

Bob, who can make even inhibitor evolution interesting.

Nick, who saved me from the Cleansing Department. Science will never forgive you.

Chris, who would run a tidy lab and is always nice.

Mark Hirst, Monaghan, P.A. Budd, Dinosaur Jones Jr., and R.R.Meehan, who all lived in a challenging environment.

Liz Graham and Duncan Davidson for in situ supervision.

Sandy, Norman and Douglas who turned pitch-black autorads into black and white prints.


And Jenny Helen Patricia, for being a hoot.

This project was carried out at the MRC Human Genetics Unit (1986-1989). I am grateful to all of the staff there, who made this a very enjoyable period of my life.
DECLARATION.

I declare that:

a) This thesis has been composed by myself, and

b) The work is either my own or that the worker/author involved is clearly stated.
ABSTRACT

The serine proteinase inhibitor (serpin) gene superfamily encodes a large group of proteins with diverse biological functions. The best characterised serpins are inhibitors found in mammalian serum. These modulate the proteolytic cascades responsible for fibrinolysis, tissue degradation and the inflammatory response. X-ray crystallography analysis indicates that an inhibitor presents an external peptide loop, the reactive centre, to its target proteinase during inhibition. The loop resembles an ideal substrate, but is cleaved relatively slowly. Synthetic or spontaneous mutations in the reactive centre region often alter the proteinase specificity of the inhibitor, without affecting the innate inhibitory mechanism.

A subfamily of rodent serpins, the Spi-2 genes, were previously shown to have undergone accelerated evolution in their reactive centre coding sequences. This may have been due to positive Darwinian selection for resistance to parasitic proteinases secreted as virulence factors. This project involved attempts to characterise the mode and tempo of this evolutionary phenomenon.

An RNA sequencing technique was developed to analyse reactive centre sequence divergence of an Spi-2 gene, contrapsin, in mouse species closely related to Mus domesticus. This indicated that the contrapsin reactive centre stabilised and became fixed for a specific inhibitory function following a period of rapid divergence during murid radiation. Analysis of two closely related rat Spi-2 genes suggests that non-synonymous point mutations may be the main mechanism of reactive centre sequence divergence.

A number of Spi-2 genomic sequences were isolated from inbred mouse strains. Analysis of these showed that the mouse Spi-2 locus consists of at least twelve closely linked genes, all with distinct reactive centres. Sequence analysis of the genes confirmed
the previous observation that the reactive centre region is hypervariable in this gene family. No obvious structural peculiarity of the genomic sequences which could account for the reactive centre divergence was observed. The amplification of the Spi-2 locus in the genus *Mus* appears to have been a recent and relatively complex process. A potential candidate for the primordial Spi-2 sequence was characterised at this stage.

Cloning and analysis of expressed Spi-2 genes in the mouse revealed expression of novel Spi-2 transcripts in the liver and in a chondrocytic teratocarcinoma cell line. The reactive centre of one cDNA sequence closely resembles that of a human serpin from a separate subfamily. An Spi-2 transcript showing several similarities to the unique human Spi-2 gene, α1-antichymotrypsin, was also isolated and may represent the functional rodent equivalent. *In situ* hybridisation to 16-day mouse embryos confirmed Spi-2 expression in cartilage, suggesting previously unsuspected functions for the duplicated genes. The isolation of two distinct Spi-2 subtypes from three separate rodent species suggests a model for the evolution of the Spi-2 family during rodent radiation.

In conclusion, the work described in this thesis confirms and extends the previous observations of serpin reactive centre hypervariability in rodents, and demonstrates that this process may be evolutionarily punctuated for expressed sequences. The mode of divergence appears to be mainly by non-synonymous point mutations, although closely linked divergent reactive centres suggest a potential role for restricted sequence conversion events. The process of acquisition of novel reactive centre specificities may be ongoing in the genus *Mus*, and is probably linked to gene duplication events.
Table of Contents

CHAPTER 1: Introduction ........................................................................................................ 1
Protein Inhibitors of Serine Proteinases ........................................................................... 3
The Kunitz Inhibitor Family ............................................................................................. 4
The Kazal Inhibitor Family ............................................................................................... 5
The Serine Proteinase Inhibitor (Serpin) Superfamily .................................................. 6
α1-Proteinase Inhibitor - the Archetypal Serpin ............................................................... 7
Tertiary Structure of Serpins ............................................................................................. 9
The Diversity of Serpin Function ..................................................................................... 12
The Nature of Rodent Serpin Gene Complexes ............................................................. 13
General Features of Multigene Families ......................................................................... 15
Mechanisms of Gene Duplication ................................................................................... 15
Gene Conversion in Multigene Family Evolution ............................................................. 16
Unusual Evolutionary Behaviour in the spi-2 Multigene Family ..................................... 17
Aims of the Project .......................................................................................................... 21
CHAPTER TWO: Materials and Methods ......................................................................... 23
Sequencing ....................................................................................................................... 23
Library Screening ............................................................................................................. 27
DNA Preparation ............................................................................................................. 29
RNA Isolation .................................................................................................................. 32
Subcloning ....................................................................................................................... 34
Southern and Northern Blotting ...................................................................................... 38
Nucleic Acid Probes ......................................................................................................... 41
Protein Analysis ................................................................................................................ 43
CHAPTER THREE: The Evolution of Contrapsin in the Genus Mus ................................. 45
Introduction ....................................................................................................................... 45
Northern Blot Analysis of Contrapsin Evolution in Mus .................................................. 49
Attempted PCR Analysis of Contrapsin Reactive Centre ................................................. 52
Sequencing of Contrapsin Reactive Centres in Mus ....................................................... 54
Evolution of the Contrapsin Gene in More Divergent Mice ........................................... 57
Divergence of Contrapsin in Ancient Mice ....................................................................... 61
RNA Sequencing of Albumin in Mus .............................................................................. 64
RNA Sequencing of α1-Proteinase Inhibitor in Mus ....................................................... 67
The Rat 4.12B Gene - A Recent Spi-2.1 Duplication Event ............................................. 70
Conclusions ...................................................................................................................... 74
### CHAPTER FOUR: The Genomic Structure of the Spi-2 Locus

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>75</td>
</tr>
<tr>
<td>Complexity of the Spi-2 Locus in Rodents</td>
<td>75</td>
</tr>
<tr>
<td>Analysis of Spi-2 Genomic Clones</td>
<td>77</td>
</tr>
<tr>
<td>Subcloning and Sequencing of DBA Spi-2 Reactive Centres</td>
<td>80</td>
</tr>
<tr>
<td>Genomic Analysis of DBA Spi-2 Sequences</td>
<td>86</td>
</tr>
<tr>
<td>Mapping of a DBA Spi-2 Gene to the spi-1/Spi-2 Complex</td>
<td>87</td>
</tr>
<tr>
<td>Screening of 129 Cosmid Library</td>
<td>89</td>
</tr>
<tr>
<td>Cosmid Mapping Strategies</td>
<td>90</td>
</tr>
<tr>
<td>Assembly of Spi-2 Contigs</td>
<td>95</td>
</tr>
<tr>
<td>Analysis of Spi-2 Contig Features</td>
<td>106</td>
</tr>
<tr>
<td>Analysis of Reactive Centre Sequences in the Contigs</td>
<td>107</td>
</tr>
<tr>
<td>Attempts to Isolate a Contrapsin-specific Genomic Sequence</td>
<td>114</td>
</tr>
<tr>
<td>Conclusions</td>
<td>115</td>
</tr>
</tbody>
</table>

### CHAPTER FIVE: Characterisation and Distribution of Expressed Spi-2 Sequences

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>116</td>
</tr>
<tr>
<td>Screening Genomic Clones for Expressed Sequences</td>
<td>117</td>
</tr>
<tr>
<td>DBA Genomic Clones</td>
<td>117</td>
</tr>
<tr>
<td>129 Strain Genomic Clones</td>
<td>120</td>
</tr>
<tr>
<td>Expressed Spi-2 Sequences in Apodemus sylvaticus Liver</td>
<td>123</td>
</tr>
<tr>
<td>Screening for Mouse α1-antichymotrypsin</td>
<td>127</td>
</tr>
<tr>
<td>Screening of EB22 cDNA Library for Spi-2 Sequences</td>
<td>127</td>
</tr>
<tr>
<td>Analysis of Putative Murine α1-antichymotrypsin</td>
<td>128</td>
</tr>
<tr>
<td>Analysis of Spi-2 Expression in the 16-Day Mouse</td>
<td>131</td>
</tr>
<tr>
<td>Screening for the Rat Arg-Ser Orthologue</td>
<td>136</td>
</tr>
<tr>
<td>Two Conserved Spi-2 Functions in the Family Muridae</td>
<td>137</td>
</tr>
<tr>
<td>Conclusions</td>
<td>143</td>
</tr>
</tbody>
</table>

### CHAPTER SIX: Expression of Cloned Spi-2 cDNAs

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>145</td>
</tr>
<tr>
<td>Construction of a Complete Spi-2 Coding Sequence</td>
<td>146</td>
</tr>
<tr>
<td>Attempts to Express Inhibitor Protein in E. coli</td>
<td>146</td>
</tr>
<tr>
<td>Inhibitor Expression by in vitro Translation</td>
<td>150</td>
</tr>
<tr>
<td>Attempts to Express Inhibitors in Mammalian Cells</td>
<td>152</td>
</tr>
<tr>
<td>Conclusions</td>
<td>153</td>
</tr>
</tbody>
</table>

### CHAPTER SEVEN: Discussion

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Evolution of Contrapsin</td>
<td>155</td>
</tr>
<tr>
<td>The Spi-2 Locus in the Laboratory Mouse</td>
<td>158</td>
</tr>
<tr>
<td>Expression of Spi-2 Genes in the Laboratory Mouse</td>
<td>160</td>
</tr>
<tr>
<td>Future Directions</td>
<td>163</td>
</tr>
</tbody>
</table>
CHAPTER ONE

Introduction

The term protease has been used for almost a century to describe enzymes which hydrolyse peptide bonds (Hedin, 1903). This vast group of molecules has since been divided into the Endopeptidases (or Proteinases) which cleave peptide bonds within aminoacid polymers, and the Exopeptidases which can only remove residues from, or cleave bonds close to, the ends of polypeptides. Proteinases can in turn be subdivided into four groups (Hartley, 1960) on the basis of their catalytic mechanisms. Originally these were defined as the serine, thiol, acid and metal proteinases; this grouping has become slightly amended to the serine, cysteine, aspartic and metallo proteinases, respectively (Barrett, 1980). Examples of each class have been sequenced at the protein level and their structure determined by X-ray crystallography.

Certain structural features are common to all four groups. The catalytic cleft lies along the surface of the enzyme. The target peptide sequence lies along this cleft during cleavage, and residues lining the cleft are important in determining which peptide sequences are bound and cleaved. This has given rise to a terminology for substrate-binding subsites, and their complementary substrate residues (Berger and Schlechter, 1970). Residues in the substrate are numbered P1, P2, P3 in an amino terminal direction away from the peptide bond cleaved by the proteinase. Residues towards the carboxyl end are numbered P1', P2', P3', etc. Sites responsible for binding the substrate in the enzymes cleft are numbered S1, S2, S3 towards the substrate amino terminus, and S1', S2', S3' towards the carboxyl end.
Each type of catalytic mechanism may have arisen more than once during evolution. Where this is the case, and it is established for at least the serine and cysteine proteinase families, one subfamily appears to have been more successful than the others and is consequently more widely dispersed. The classic example is that of the chymotrypsin and subtilisin superfamilies in the serine proteinases. Examples of the chymotrypsin superfamily have been found in most major prokaryotic and eukaryotic groups, whereas subtilisin related enzymes have only been found in bacteria. The two groups have acquired a very similar catalytic mechanism by convergent evolution (Hartley, 1970).

The mechanism used by the serine proteinases has now been fairly well established. The critical serine residue, Ser-227 in chymotrypsin, was shown to be esterified by diisopropyl fluorophosphate when this compound inactivated chymotrypsin (Hartley, 1960). Similar experiments demonstrated the importance of a histidine residue, His-60. The hydroxyl group of the serine residue appears to attack the carbonyl carbon atom in the peptide substrate, giving a tetrahedral intermediate. The histidine residue supplies a basic environment to facilitate the reaction. Breakdown to an acyl-enzyme and release of an alcohol or amine is followed by water-mediated hydrolysis of the acyl-enzyme complex to yield enzyme and products. The Asp-102 residue was at one time considered to have an essential role in the process, forming the "charge-relay" system, but the discovery of serine proteinases which do not possess an appropriate aspartate residue now challenges this view (Lazure et al., 1984).

Serine proteinases have a huge range of specificities. As mentioned above, residues lining the catalytic groove dictate the nature of substrate peptides bound, and the extent of the interactions depends upon the particular proteinase. In general, sites towards the amino terminus of the active site are more important in serine proteinases. An example is trypsin, which has an absolute requirement for an arginine or lysine residue in the
substrate at the P1 position. Trypsin will hydrolyse most peptides at these residues, as would be expected for a broad range digestive enzyme. However, many other serine proteinases which also cleave at basic residues have developed increased specificity in the S2 and S3 positions, thus restricting their potential target range and refining their role. This can result in enzymes which will cleave only one or a few particular peptide bonds in vivo. An extreme example is nerve-growth-factor γ-subunit endopeptidase which has no general proteolytic activity and is specific for nerve growth factor proprotein cleavage (Berger and Shooter, 1977). Other groups of serine proteinases show specificity for cleavage at large hydrophobic P1 residues such as tryptophan and phenylalanine (eg the chymotrypsin family), or small aliphatic P1 residues such as valine and alanine (eg the elastase family). Again, these specificities may depend largely upon the P1 residue, or be attenuated by the surrounding substrate sequence

**Protein Inhibitors of Serine Proteinases**

Inhibitors of proteinases in vivo are normally proteins which act by binding tightly to the catalytic cleft as a potential substrate, but only being cleaved very slowly. Described kinetically, a typical enzyme:inhibitor interaction will have a very high $k_{cat}/K_m$, typically $10^4$ to $10^6$ moles/second, indicating that the inhibitor is an excellent substrate. However, both $k_{cat}$ and $K_m$ have values several orders of magnitude lower than typical substrate cleavages, making the hydrolysis of the reactive site bond extremely slow (Laskowski and Kato, 1980). Examinations of the factors which retard cleavage have employed extensive X-Ray crystallographic analysis of inhibitor:proteinase complexes and of the isolated components (Read and James, 1986). These have demonstrated that, as expected, most inhibitors have a reactive centre (the polypeptide loop containing the P1 residue) complementary to the catalytic cleft of their target enzyme. They also possess rigid structures, presenting the reactive centre region as an exposed
loop, the degree of inhibitor rigidity increasing with increasing target specificity. This rigidity is often contributed by multiple disulphide bonds in the smaller inhibitors, but can be attained by other structural means, such as extensive β-pleated sheet structure around the reactive centre loop. At least ten families of protein inhibitors of serine proteinases have been characterised and some typical examples will be described.

*The Kunitz Inhibitor Family*

The pancreatic trypsin inhibitor (PTI, also described as Kunitz) family contained the first inhibitor to be characterised and sequenced, and its structure, free and in complex, studied by X-ray crystallography (Kunitz and Northrop, 1936; Huber et al., 1974). Bovine PTI, the most studied example, is a small inhibitor (58 amino acids) found in most bovine organs. Its reactive site contains Lys-Ala at the Pi-Pi' position, and as expected it is an efficient trypsin inhibitor. The 3D structure of PTI in complex with bovine trypsin, trypsinogen, and kallikrein has been determined (Read and James, 1986). The inhibitor is pear shaped, with the reactive site on an exposed loop at the narrow end of the molecule. The loop conformation is not radically altered during complex with proteinase; the reactive site loop and catalytic cleft fit one another in a classical "lock and key" fashion. PTI contains three disulphide bonds which help to stabilise a small hydrophobic core region, and a β-pleated sheet region is located immediately to the carboxyl side of the reactive site loop region. Members of the PTI(Kunitz) family have been found in mammals, reptiles, molluscs and sea anemones; the family may predate the radiation of multicellular organisms (Laskowski and Kato, 1980). Examples exist of tandemly repeated copies of PTI (Kunitz) subunits within larger proteins; eg inter-α-trypsin inhibitor, a protein of 160,000 kD, which contains two PTI-like inhibitor domains (Morii and Travis, 1985). In the case of the Red Sea turtle egg-white inhibitor, chelonianin, the first of two tandemly linked inhibitor domains is a typical PTI (Kunitz) structure, while
the second is unrelated, and may represent a new inhibitor family (Kato et al., 1980).

*The Kazal Inhibitor Family*

A second very large inhibitor family is the Pancreatic Secretory Trypsin Inhibitor (Kazal) [PSTI(Kazal)] family. Examples of this have also been well characterised by sequencing and X-ray crystallography (Kazal et al., 1948; Weber et al., 1981). This inhibitor looks quite different from the PTI (Kunitz) family, forming a wedge-shaped disc, with the reactive site at the narrower edge. As in PTI there are three disulphide bonds, but they are in nonhomologous positions. Despite an entirely different tertiary structure, the reactive site loop of PSTI is very similar in its conformation to that of PTI. Studies of enzyme inhibitor complexes show that the loops of both families contact the proteinase catalytic cleft in a very similar manner; a clear case of two different structural solutions to a biological problem (Read and James, 1986). The PSTI (Kazal) family is also an ancient one, predating vertebrate radiation, with examples in leeches (Krejci and Fritz, 1976).

Several other families of small inhibitors have been described, each apparently unrelated (at least by recent evolutionary descent) to the others (Laskowski and Kato, 1980; Read and James, 1986). They share the common characteristics of the PTI (Kunitz) and PSTI (Kazal) families: defined secondary structure despite their small size, presence of a reactive centre loop at the surface of the molecule, frequent tandem repetition of inhibitor units within a larger protein, and retention of activity (though against novel proteinase targets) despite substitution of the P1 residue by chemically different amino acids.

*The Serine Proteinase Inhibitor (Serpin) Superfamily*

A mechanistically distinct class of inhibitors, first characterised in mammalian plasma, displays several differences from the numerous families of small inhibitors
described above (reviewed in Travis and Salvesen, 1983). These inhibitors are considerably larger (about 400 amino acids), with a single inhibitory site, and do not occur as repetitive subunits within a protein. Unlike classical (small) inhibitors, they do not obey the equilibrium reaction; upon cleavage they become completely inactivated as inhibitors. They also appear to form a covalent bond with the proteinase during complex formation; this structure is resistant to SDS or urea treatment. It is not clear whether this bonding occurs during complex formation, or as a result of denaturing agent treatment. Irreversible modification of the serpin molecule by proteinase, accompanied by active proteinase release, occurs in many instances. This often occurs with greater efficiency during interactions between non-cognate inhibitors and proteinases.

Isolation and partial amino acid sequencing of two inhibitors, α1-antitrypsin and antithrombin III, and a third protein, chicken ovalbumin, revealed that all three were related to one another. The three proteins had apparently diverged some 500 million years ago (Hunt and Dayhoff, 1980), and formed the first characterised members of a novel protein superfamily, the serpins (serine proteinase inhibitors). Several other inhibitors from mammalian serum were rapidly characterised at the DNA level, and it has become clear that the serpins form a very large, complex family, with a convoluted evolutionary history (Carrell et al., 1987). Proteins from several unexpected sources have been found to be members of the serpin family, some of them, like ovalbumin, with no detectable inhibitory activity towards any serine proteinase. The recent report of an inhibitory serpin in an insect raises the question of how ancient the family really is (Kanost et al., 1989). Before examining some facets of the evolutionary behaviour of the family, the biology of the most studied serpin, α1-proteinase inhibitor (α1-PI, formerly α1-antitrypsin) will be described in more detail.
**α1-Proteinase Inhibitor - the Archetypal Serpin**

α1-Proteinase Inhibitor is a globular serum protein of 394 amino acid residues (Carrell et al., 1982). It is primarily synthesised in the liver parenchymal cells, where it forms a substantial fraction of the mRNA population (0.5%). It is a major component of plasma (~3mg/ml), only less abundant than albumin and the immunoglobulins. The physiological role of α1-PI is thought to be the inhibition of neutrophil elastase, with which it forms a highly stable complex \( k_{\text{ass}} = 10^{-7} \text{ sec}^{-1} \text{ in vitro} \), the most rapid serpin/proteinase association rate known. [Travis and Salvesen, 1983]). It also inhibits a wide range of other proteinases, and may play a significant secondary role *in vivo* in the control of certain physiological proteinases, in addition to neutrophil elastase (Heidtmann and Travis, 1986). The primary function of α1-PI appears to be the prevention of uncontrolled degradation of elastin fibres, especially in the lung, by neutrophil elastase. Mutations in the α1-PI gene, or conditions in which inhibitor levels are reduced (eg smoking-induced α1-PI oxidation), cause pulmonary emphysema (Laurell and Eriksson, 1963). A number of bacterial and other non-serine proteinases rapidly inactivate α1-PI as an ideal substrate (Johnson et al., 1986). This involves a single cleavage in the reactive centre region causing irreversible inactivation, with no detectable inhibition of the proteinase.

Several mutations of the human α1-PI gene have been described (Carrell et al., 1989). The majority of these, which abolish or drastically affect correct translation and folding of the protein, tend to give rise to a phenotype of emphysema of variable severity. Mutations affecting glycosylation also occur, with altered proteins being inefficiently secreted from the liver and causing liver cirrhosis as a second symptom (Jeppsson, 1976). All of these types of mutation tend to give a classical null effect for the mutated allele. However a second class of α1-PI mutants exists. These are mutated
at the P1 residue, altering the target proteinase specificity of the inhibitor, but not the structural ability of the protein to efficiently inhibit a proteinase. The classic example for α1-PI is the Pittsburgh mutation (Owen et al., 1983). The normal P1 residue in α1-PI is methionine, an amino acid most efficiently recognised by elastase. In α1-PI Pittsburgh, the methionine residue has mutated to arginine, altering the specificity of the inhibitor towards the tryp tic proteinases. The inhibitory efficiency of α1-PI Pittsburgh towards elastase was decreased by four orders of magnitude, while inhibition of trypsin-like proteinases (thrombin, plasmin, kallikrein) increased to the same degree. A normal (α1-PI) level of expression of the mutant, coupled with the new specificity, would be expected to cause major physiological effects. In fact the individual with this mutation suffered from an eventually fatal inability to form blood clots, due to the inappropriate antithrombin activity of α1-PI Pittsburgh. This ability of the serpins to undergo complete change of proteinase specificity, via mutation of the P1 residue, without losing the general structural features which cause anti-proteinase activity, is very similar to that of the smaller inhibitor molecules discussed above (Laskowski and Kato, 1980).

A number of studies have been carried out to examine the effect of site-directed mutation of the P1 residue on target specificity (Rosenberg et al., 1984; Courtney et al., 1984; Jallat et al., 1986; Holmes et al., 1987). These have confirmed that the primary determinant of an inhibitor’s proteinase specificity is the P1 residue, and that this can be modified in a predictable manner, without loss of innate inhibitory activity. Mutants with "improved" α1-PI activity have been produced, replacing methionine with leucine or valine. These proteins retain very similar biochemical specificity to α1-PI, but are not susceptible to the methionine oxidation and consequent inactivation thought to occur in smoking-induced emphysema. They may prove to be useful therapeutic agents for emphysema sufferers (Carrell and Travis, 1985).
Tertiary Structure of Serpins

Attempts to characterise the three dimensional structure of the serpins have not been very successful. Intact, unmodified, inhibitor crystals have proved difficult to produce, despite several attempts. The reason for this may be that the intact serpin structure is in a highly stressed conformation. One line of evidence which supports this theory comes from successful crystallographic analysis of elastase-cleaved α1-PI (Loebermann et al., 1984) which revealed that the P1 methionine and P1' serine residues lie at opposite ends of the molecule, about 70 Angstroms apart. A representation of the structure is shown in Figure 1.1. Since these residues would be adjacent in native (unmodified) inhibitor, it is clear that a major conformational change has occurred in the protein as a result of proteinase:inhibitor interaction.

Attempts to crystallise ovalbumin, a distantly related serpin, have also proved difficult, except for proteinase-treated ovalbumin, plakalbumin (Wright, 1984). The structural instability of intact serpins may prove to be the rule for most family members. However there are preliminary reports of the production of native antithrombin III crystals (Carrell et al., 1987), but these have not yet yielded detailed structural data. There are no known reports of X-ray analysis of serpin:proteinase complexes.

The α1-PI X-ray analysis (Loeberman et al, 1984), revealed other notable features of the tertiary structure. Flanking the reactive centre loop are two highly conserved regions of amino acid sequence. These form two well defined strands of β-pleated sheet in two of the three β-sheets in the protein.
Figure 1.1 Representation of the three dimensional structure of modified α1-PI (Loebermann et al., 1984). α-helices are shown as cylinders, and β-pleated sheet strands as ribbon arrows. The residues Met-358 and Ser-359 which comprise the P1-P1’ residues are indicated by asterisks at opposite ends of the modified molecule. The β-barrel structure discussed below is the cup-like structure at mid-left of the molecule, formed by the two minor β-sheets and surrounding α-helices.
The reactive centre region itself forms a strand of the major β-sheet A in the modified inhibitor, but presumably exists as a surface loop in the native molecule. This flanking of the reactive centre by defined secondary structure is also reminiscent of the small inhibitor families, and may echo the requirement for a stressed conformation in the bait region.

The concept of two forms of serpin, stressed (S) and relaxed (R) has been developed by Carrell and others (Carrell and Owen 1985; Carrell et al., 1987, Bruch et al., 1988). Considerable increases in the heat stability of a number of serpins occur following proteolytic modification of the reactive centre, presumably due to rearrangements similar to the one which occurs in modified α1-PI. The modified R form:proteinase complex may exhibit novel epitopes to enable clearance of complexes from the circulation by specific receptors (Pizzo et al., 1988), and may also play a role in localised hormone release by certain serpins (Pemberton et al.,1988, and see below). This overall view of serpin structure, passing from a strained intact inhibitor to a relaxed, irreversibly inactivated, modified form, has been challenged recently (Gettins 1989). Nuclear magnetic resonance studies of proteolytically modified ovalbumin showed no marked conformational change from the intact state upon cleavage. Gettins suggests that the major change seen in some serpins may be due to disruption of internal salt bridges exposed by cleavage of the reactive centre loop rather than to the direct release of a strained conformation by loop cleavage.

Another feature of the cleaved α1-PI structure is the existence of a β-barrel structure at one end of the molecule. This is formed by the two β-sheets not containing the cleaved reactive centre strand, and can be seen at the mid-left of the molecule in Figure 1.1. The barrel structure is surrounded by most of the basic residues in the molecule, and echoes the β-barrel structures found for example in transthyretin, a major plasma
thyroxine-binding protein (Blake et al., 1974). An indication that this structure may be involved in hormone transport in some serpins is presented below.

*The Diversity of Serpin Function*

Most serpins characterised to date have a defined anti-proteinase activity, which in several cases had been defined biochemically before the advent of gene cloning. Most characterised mammalian serum inhibitors belong to the serpin superfamily. However, several serpins have been found in unexpected roles.

The rat glia-derived neurite-promoting factor (GdNPF) is a 43kD glycoprotein involved in neurite outgrowth and displays inhibitory activity to several serine proteinases. Cloning of the gene by hybridisation-selected translation revealed that GdNPF was a serpin, with 25 - 40% identity to other mammalian inhibitors (Sommer et al., 1987). The action of GdNPF in promotion of neurite extension is shared by other serine proteinase inhibitors (including synthetic tri-peptides), but other effects of GdNPF, such as peripheral nervous system regeneration, granule cell neuron migration, and astrocyte proliferation, are unique. These may be due to effects of the products of GdNPF proteolysis, rather than to proteinase inhibition.

Thyroxine binding globulin (TBG), the major vertebrate thyroxine-binding protein, was cloned using polyclonal anti-TBG antibodies to screen a λ-gt11 expression library (Flink et al., 1986). The cDNA sequence showed a high degree of similarity to serpin sequences, particularly the functional inhibitors α1-PI and α1-antichymotrypsin. TBG has no known inhibitory activity, but appears to undergo the major conformational change typical of several other serpins when treated with proteinase (Pemberton et al., 1988). The thyroxine binding site is proposed to be the β-barrel domain described above, since transthyretin, which binds two thyroxine molecules, contains two β-barrel
domains (Blake et al., 1974). A second hormone-binding protein, corticosteroid binding globulin (CBG) (Hammond et al., 1987), also proved to be a serpin, again with no obvious anti-proteinase activity. Pemberton et al. (1988) propose that the retention of the switch from S to R in these proteins is required to allow altered delivery of hormone at inflammatory sites where proteinase activity is increased.

The gene for a 38kD protein from cowpox virus responsible for lesion haemorrhage has been cloned using a series of deletion mutants (Pickup et al., 1986). The sequence showed 30% identity to several serpins, with greater conservation of internal hydrophobic residues. Viruses containing the sequence produce localised lesion haemorrhage upon infection, but the specificity of the inhibitor is unknown, although the P1 residue is probably alanine. The inhibitor is presumed to perturb the normal serine proteinase-mediated coagulation cascade. The origins of this gene are uncertain but may well have arisen by horizontal evolution via gene transfer from the host to the virus, as is thought to occur for viral oncogenes.

A cDNA isolated from an insect (Manduca sexta) fat-body cDNA library has been sequenced, and encodes a serpin with inhibitory activity (Kanost et al., 1989). The inhibitor is anti-elastase in type, with an alanine at the P1 position, and shows anti-elastase and antichymotryptic activity. Comparison of the sequence to mammalian serpins indicates that the progenitor gene existed before the divergence of vertebrates and invertebrates. To date, this is probably the most distant characterised member of the superfamily from the mammalian serum inhibitors.

The Nature of Rodent Serpin Gene Complexes

A previous report from this laboratory (Hill et al., 1984) described the sequences of two serpins derived from a mouse liver cDNA library. One of these was the putative
murine equivalent of \( \alpha_1 \)-Proteinase Inhibitor, and the second was the serum inhibitor contrapsin (Takahori and Sinohara, 1982), which has an anti-trypsin type activity. There is no known equivalent of contrapsin in human serum, but screening of a human liver cDNA library enabled isolation of a contrapsin-related sequence expressed at moderately high levels (0.1% of clones). This sequence proved to be \( \alpha_1 \)-antichymotrypsin, a well characterised human serum serpin with an antichymotryptic spectrum of activity, principally active against cathepsin G and mast cell chymase (Reilly et al., 1982; Wintroub et al., 1982). It is an acute phase reactant, with levels rising rapidly in response to infection or inflammation (four-fold in eight hours (Fey and Fuller, 1987) The human and mouse sequences showed 70% identity at the nucleotide level, with high conservation of the amino acid residues thought to comprise the internal hydrophobic domains of the protein. However, the reactive centre regions were completely dissimilar, with one residue in sixteen in common between the two proteins. This apparently accounts for the marked difference in specificity between the mouse and human proteins. Contrapsin and \( \alpha_1 \)-antichymotrypsin would appear to have acquired different functions since rodent/primate divergence, specifically by reactive centre divergence.

Further analysis of the contrapsin and \( \alpha_1 \)-PI genes in the mouse (Hill et al., 1985) demonstrated that each was a member of a separate clustered multigene family. The two gene families, designated \( \text{Spi-1} \) (\( \alpha_1 \)-PI) and \( \text{Spi-2} \) (contrapsin) are closely linked on chromosome 12, and would appear to have arisen as a series of duplications from a single primordial serpin sequence. In man, the situation is simplified; the \( \alpha_1 \)-PI gene (and a related sequence) lies within 120 kb of the single \( \alpha_1 \)-antichymotrypsin sequence on chromosome 14 (P. Kearney, personal communication). The \( \alpha_1 \)-PI/\( \alpha_1 \)-antichymotrypsin divergence clearly predates rodent/primate divergence since both gene types are found in both lineages. However the evolution of the \( \text{Spi-1} \) and \( \text{Spi-2} \) loci after rodent/primate
divergence has been very different in man and mouse. An intriguing evolutionary phenomenon which has occurred in the rodent Spi-2 gene complex will be discussed after a brief review of some mechanisms of multigene family evolution.

**General Features of Multigene Family Evolution**

The unusual evolutionary behaviour of multigene families has been studied in many different organisms at several different levels and has given rise to a number of basic concepts of molecular evolution. Following an initial gene duplication event providing two linked related sequences a series of processes act simultaneously upon the genes. These may include amplification and deletion of genes or parts of genes by unequal crossing over, sequence dispersion to unlinked loci, inter- and intra-family sequence conversion, slippage replication, selection and drift. The processes all exert varying degrees of influence upon individual genes within a family and upon the homogeneity of the family (Arnheim, 1983; Maeda and Smithies, 1986; Dover and Strachan, 1987).

**Mechanisms for Gene Duplication**

The initial event which produces two linked copies of a gene can occur in a number of ways. Chromosome breakage and reunion can rescue two copies of a unique sequence onto one chromosome. This process can be random, as no recombination signals have been discovered for certain families which appear to have arisen in this way. An example is the human haptoglobin locus where two separate, apparently random, breakages have generated both intragenic and intergenic duplications (Maeda and Smithies, 1986; Bensi et al., 1985). A second powerful mechanism involves duplication by homologous recombination between repeated sequences flanking the original gene. Recombination between two truncated Kpn1 repeated sequences appears to have been responsible for the duplication of the primordial foetal haemoglobin gene in man as remnants of the
repeat sequence are located between the duplicated genes, and at the 5' and 3' boundaries of the duplication (Shen et al., 1981). Alu repeated sequences appear to have mediated the initial duplication event in the human growth hormone/human chorionic somatomammotrophin in an identical fashion (Barsh et al., 1983).

A repeat-dependent homologous recombination mechanism can equally well expand or contract an extant multigene family, and mediate intragenic duplication or deletion. However, the principal mechanism for expansion or contraction of the copy number of a multigene family would appear to be unequal crossing over between the previously duplicated sequences. In the β-globin locus this has been demonstrated both within genes, yielding chimaeric proteins, and between genes, causing duplications and deletions (Flavell et al., 1978; Jones et al., 1981). Unequal crossing over events frequently generate fluctuations in copy number in certain multigene families (Hill et al., 1986; Geliebter and Nathenson, 1987). The isolation of a sequence by dispersion to a separate chromosome, or by physical separation from related sequences on the same chromosome often leads to the development of two related, but evolutionarily independent, subfamilies. This has presumably occurred in the unlinked α-globin and β-globin multigene families (Jeffreys, 1982) and the linked but isolated Igh-V subfamilies in the mouse (Riblet et al., 1987; Tutter, 1988).

Gene Conversion in Multigene Family Evolution

Gene conversion, the non-reciprocal transfer of sequence from one gene to another related one, is a major mechanism in multigene family evolution. The process can cause the phenomenon of concerted evolution, where genes within a species are more closely related to one another than to their orthologues in a separate species (Arnheim, 1983). Gene conversion was first demonstrated at the nucleotide level for the two human foetal globin genes, Gγ and Aγ, thought to have duplicated twenty to forty million years ago,
yet which on a single chromosome were found to possess pronounced sequence similarity over a 1500 basepair region (Slightom et al., 1980). The process of gene conversion can also generate novel polymorphism in coding sequences by highly localised conversion events, as demonstrated for the H-2k\textsuperscript{bml} mouse major histocompatibility gene (Schulze et al., 1983; Weiss et al., 1983), and accelerate sequence divergence of recently duplicated genes, as shown for the human immunoglobulin $\text{V}_k$ genes (Bentley and Rabbits, 1983).

Two contrasting examples of fairly recent multigene family amplifications in the mouse are the mouse major urinary proteins, or MUPS (Bishop et al., 1982; Clark et al., 1984) and the kallikrein family (Mason et al., 1983; Evans et al., 1987). The MUP amplification unit appears to consist of a class 1 (active) and class 2 (pseudogene) sequence in a divergent transcriptional orientation within a 45 kb genomic domain. The pseudogene component appears to have become inactivated before the amplification occurred (Ghazal et al., 1985), although it may have acquired a secondary biological function (Clark et al., 1985). The kallikrein family contains a similar number of genes and pseudogenes (about fifteen of each) to the MUP locus, but the duplication mechanism has resulted in tandem linkage of all the sequences, with inactivation presumably occurring after duplication (Drinkwater et al., 1987; Evans et al., 1987).

**Unusual Evolutionary Behaviour in the Spi-2 Multigene Family**

During further analysis of Spi-2 genes in rodents, (Hill and Hastie, 1987) the isolation of two rat Spi-2 cDNAs from liver libraries revealed additional highly unusual evolutionary behaviour. As mentioned above, the murine contrapsin and human $\alpha$-antichymotrypsin genes which had been presumed to be orthologous showed considerable divergence at the nucleotide and amino acid levels in their reactive centre regions. This had led to distinct inhibitory specificities, apparently in the period since
rodent/primate divergence. The rat sequences, \textit{Spi-2.1} and \textit{Spi-2.2}, were sufficiently related to one another and to contrapsin to fall within the \textit{Spi-2} family; and presumably to have arisen from an \textit{Spi-2} family amplification event following rodent/primate divergence. The \textit{Spi-2.2} sequence shares several characteristics of the human \textit{\alpha 1}-antichymotrypsin gene; in terms of induction following acute phase inflammation, transcript size, and especially chemical nature of the P1 residue, this gene appears to encode a functional rat equivalent of the human gene. In contrast, the \textit{Spi-2.1} gene resembles the mouse contrapsin gene, with a similar size of transcript, a high constitutive level of expression, and the probable specificity of the P1 residue. However, if the rat \textit{Spi-2.1} represented the true orthologue of the mouse contrapsin, it appeared to have undergone extremely rapid nucleotide evolution in the reactive centre region since \textit{Spi-2.1} and \textit{Spi-2.2} divergence, and also in the shorter evolutionary period following mouse/rat divergence. Contrapsin and \textit{Spi-2.1} display less nucleotide identity in the reactive centre region than contrapsin and \textit{\alpha 1}-antichymotrypsin.

To establish in more detail the quantitative nature of the sequence divergence, analysis of the four sequences, contrapsin, \textit{Spi-2.1}, \textit{Spi-2.2}, and \textit{\alpha 1}-antichymotrypsin was carried out using the method of Li, Wu, and Luo (1985). This sophisticated algorithm estimates rates of synonymous and asynonymous substitution in related sequences, applying values derived from mammalian genes for the different probabilities of different codon changes. The program calculates a rate of about $6 \times 10^{-9}$ substitutions per site per year for pseudogene sequences in mammals. A rate of $0.8$ to $1 \times 10^{-9}$ is typical for non-synonymous (ie amino acid altering) substitutions, and a five-fold higher value ($4.65 \times 10^{-9}$) for synonymous (translationally silent) nucleotide substitutions. This is in accord with molecular evolutionary theory which would predict that synonymous substitutions should occur at a rate close to that of substitutions in (non-selected) pseudogenes, and
considerably more frequently than possibly deleterious nonsynonymous substitutions.

However, the calculated rates of nonsynonymous substitution in the reactive centre regions (Region 2) of the Spi-2 sequences were four-fold higher than the rates of synonymous substitution in the more conserved upstream sequences (defined as Region 1) using the Li program. Actual substitution rates were not always calculable for Region 2, due to the degree of sequence divergence. The rate of nonsynonymous substitution for the contrapsin and rat Spi-2.1 genes was 6-8 times that of the overall substitution rate for two mouse pseudogenes (Hill and Hastie, 1987). The explanation proposed for these extremely rapid evolutionary rates, in the region of the protein most important for functional specificity, was postulated to be the occurrence of positive Darwinian selection for amino acid altering mutations in the reactive centre. Rapid evolution of novel inhibitory specificities against exogenous proteinases secreted by parasites as virulence factors could provide an appreciable selective advantage to a host, causing rapid fixation of favorable nonsynonymous mutations. The relatively high synonymous rates in the reactive centre region, where these are calculable, indicate that the situation may be complex.

The difficulty of aligning reactive centre sequences from different genes may exceed the ability of the computer program to determine realistic values.

A separate observation supporting the idea that the rodent Spi-2 sequences might have a role in defence against parasites was recently reported (Mohda et al., 1988). A component of mouse serum found associated with schistosome worm isolates from infected mice was shown by immunoprecipitation to be contrapsin. The exact nature of the interaction is not clear, but the secretion of serine proteinases by schistosomes is thought to play a major role in tissue invasion by this parasite. Contrapsin with altered electrophoretic mobility, perhaps representing a proteinase-modified form, was detected in the serum of infected compared to control mice.
Following the observation by Hill and Hastie, a similar instance of reactive centre hypervariability was reported for the avian ovomucoid inhibitors (Laskowski et al., 1987). The protein sequences of ovomucoid inhibitory domains from one hundred avian species were established, and the most variable amino acid found to be the specificity-determining P1 residue. These small Kazal-type inhibitors have been subjected to crystallographic analysis in complex with several serine proteinases; the importance of the P1 residue is not in doubt. No nucleotide sequence analysis has been reported for these proteins, therefore relative synonymous and nonsynonymous substitution rates cannot be deduced, but it appears probable that a similar phenomenon is occurring in the reactive centre regions of the ovomucoid inhibitors (Leigh Brown, 1987).

The proposition that the cause of accelerated reactive centre sequence divergence in the Spi-2 genes is due to positive selection has been challenged recently (Graur and Li, 1988). Using an index of protein mutability (Graur, 1985), the authors claim that the intrinsic amino acid composition of the Spi-2 sequences in the reactive centre regions will permit increased rates of evolution, relative to the remainder of the molecule. They also postulate that the reactive centre region is free of compositional constraint, and state that a positive selection scenario is not necessarily required to account for the sequence data reported by Hill and Hastie. Graur and Li predict that the reactive centre regions will tend to be highly polymorphic within populations, since innate reactive centre mutability with no selection would not tend to homogenise a particular sequence within a species. They suggest that a positive selection mechanism will tend to rapidly drive advantageous mutations to fixation, reducing intraspecies polymorphism (Nei and Graur, 1984). The extreme polymorphism of, for example, the MHC gene cluster in most mammals contradicts this view, at least for some gene classes (Steinmetz and Hood, 1983; Klein, 1986)).
Aims of the Project

Comparison of rodent Spi-2 genes revealed the highly unusual phenomenon of accelerated evolution in the reactive centre region (Hill and Hastie, 1987). It was of great interest to determine whether this was a general phenomenon for this gene family in rodents, the mechanisms which might promote accelerated evolution, and the actual divergence rates involved.

One approach involved determining the nature of the contrapsin reactive centre sequence in murids closely related to the laboratory mouse. At this stage, the status of the contrapsin reactive centre sequence in rodents was uncharacterised, except for the original Swiss inbred mouse strain (Hill et al., 1984) and the putative orthologue in the rat, Spi-2.1 (Hill and Hastie, 1987). A strictly neutralist prediction would be that this region would continue to evolve rapidly over comparatively short periods of time, due to its predicted intrinsic mutability (Graur and Li, 1988). Additionally, if the accelerated sequence divergence were due to a hypermutational mechanism peculiar to the Spi-2 reactive centre, analysis of closely related murine species might reveal intermediates in the process. However, if the divergence had been due to a positive selection for a particular novel endogenous reactive centre specificity, then stabilisation of the sequence might have occurred at some stage during murine radiation. This type of "internal race" would be more likely to result in stabilisation than selection by exogenous parasitic proteinases, where continuous rapid co-evolution might be expected (Clarke, 1976).

A second area of research involved the nature of the Spi-2 complex in rodents. It was clear that the family had become amplified since primate/rodent divergence and, to some extent, since mouse/rat divergence (Hill et al., 1985). How many genes the mouse possessed at the Spi-2 locus, their mode of amplification, their intragenic structure, and the range of specificities in functional inhibitors encoded within the locus, were ques-
tions which could be addressed by genomic analysis. If accelerated evolution was occurring in several family members specifically within the reactive centre region, was a rodent-specific intron/exon structure responsible? Which genes in the family had acquired a specific physiological role, and was their evolution similarly accelerated? And finally, did the mouse possess a functional, or evolutionary, orthologue of the (ancestral?) single-copy human Spi-2 gene, α1-antichymotrypsin. This thesis describes attempts to answer some of these questions.
CHAPTER TWO

Materials and Methods

Chemicals - unless otherwise stated, all chemicals were obtained from BDH Reagents Ltd.

Enzymes - unless otherwise stated, all enzymes were obtained from BCL Ltd.

1. SEQUENCING

1.1 RNA Sequencing

For RNA sequencing in the project, modifications of the dideoxy chain termination technique (Sanger et al., 1977) were always used. Template material was normally total liver RNA, but occasionally liver mRNA (oligo d(T) method - section 4.2). Initially no standard protocols were available, and a method was independently developed. Later, published protocols were obtained (Weismann, 1986; Geliebter, 1987) which were broadly similar to the methods used in the project. Specific oligonucleotides complementary to 3'coding regions of specific mRNAs were used as primers, and the same criteria applied regarding length and composition as for DNA sequencing.

Primers for sequencing were radiolabelled with $^{32}$P γ-ATP using T4 polynucleotide kinase (Amersham International). This precluded the requirement for inclusion of a radioactive nucleotide during the extension reaction. This often gave better results, probably due to the fact that only poly-nucleotides initiated from the labelled primer were detectable, thus reducing background signal due to nonspecific priming. This can represent a particular problem when directly sequencing a highly complex population of nucleic acids. However, unlabelled primer and free radioactive nucleotides were used in
initial stages of the project and often gave reasonably low background.

10 μl of total RNA (in a volume of 4 to 5 μl) was heat-denatured at 90°C for three minutes and then briefly spun in a microfuge. One μl of RNAsin (ribonuclease inhibitor - Promega Biotech) was added, followed by 25 ng of pre-labelled oligonucleotide primer (in a volume of 5 μl) (Section 7.3) Four μl of 5X RT buffer was added (5X RT buffer = 250 mM Tris-HCl pH 8.0, 50 mM MgCl₂, 200 mM KCl and 50 mM DTT) and the reaction mix left to anneal at 37°C for twenty minutes. One μl of reverse transcriptase (BCL) was added and 3.5 μl aliquots were added to each of four reaction tubes: A, C, G, T which contained 2.5 μl of the appropriate deoxy/dideoxy nucleotide mixes. The final concentrations for these were

A' dATP 0.2 mM ddATP 0.15 mM dCTP, dGTP, dTTP 0.2 mM
C' dCTP 0.2 mM ddCTP 0.15 mM dATP, dGTP, dTTP 0.2 mM
G' dGTP 0.4 mM ddGTP 0.15 mM dATP, dCTP, dTTP 0.2 mM
T' dTTP 0.2 mM ddTTP 0.30 mM dATP, dCTP, dGTP 0.2 mM

The tubes were then incubated at 42°C for 10 minutes. In the initial reactions, 1 μl of a chase mix (1 mM dATP, dCTP, dGTP, dTTP) was added and the temperature raised to 50°C for a further 5 minutes but this did not significantly improve the sequence data and was dispensed with. The reactions were stopped by the addition of 4 μl formamide dye mix (100 ml formamide, 0.1 g xylene cyanol, 0.1 g bromophenol blue, 4 ml 500 mM EDTA) and heated to 90°C for three minutes immediately before loading onto a polyacrylamide gel (Section 1.3).

1.2 dsDNA Sequencing

Double-stranded DNA sequencing was modified and used for the project in preference to standard M13 single-stranded sequencing for two reasons. Since isolated
genomic clones often contained two or more related genes, subcloning was often required prior to detailed analysis and was carried out routinely for the individual Spi-2 sequences. Secondly, plasmid subclones from genomic inserts were frequently several kilobases in size, for ease of isolation, and to ensure that the entire 3' end of the gene was present. Inserts of more than 1 to 2 kb can be difficult to propagate in M13. Double stranded DNA sequencing is less efficient overall than single-stranded M13 sequencing, but in this project, where directed sequencing was often possible, was quite adequate.

3 μg of CsCl purified, supercoiled, plasmid DNA was added to 40 μl of denaturation soln. (0.2M NaOH, 2mM EDTA) and left at RT for five minutes. 4 μl of neutralizing soln. (0.4M NaOAc pH 4.5) was then added, followed by 90 μl of EtOH. The DNA was then spun for 15 minutes, and the pellet washed (80% EtOH) and dried. It was then resuspended in 5 μl of endlabelled oligonucleotide primer (=25 ng ), 3 μl of TE and 2 μl of 5x Sequenase sequencing Buffer. After vortexing, template and primer were annealed at 37°C for 20 minutes. Two μl of diluted ( 1:7 ) Sequenase was then added and gently mixed. Three μl of the RM was then added to each of four ddNTP reaction tubes which already contained 2 μl of the appropriate termination mix.

<table>
<thead>
<tr>
<th></th>
<th>dATP</th>
<th>dCTP</th>
<th>dGTP</th>
<th>dTTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>A'</td>
<td>8μM</td>
<td>80μM</td>
<td>dATP</td>
<td>dCTP</td>
</tr>
<tr>
<td>C'</td>
<td>8μM</td>
<td>8μM</td>
<td>dCTP</td>
<td>dCTP</td>
</tr>
<tr>
<td>G'</td>
<td>8μM</td>
<td>8μM</td>
<td>ddGTP</td>
<td>dCTP</td>
</tr>
<tr>
<td>T'</td>
<td>8μM</td>
<td>8μM</td>
<td>ddTTP</td>
<td>dATP</td>
</tr>
</tbody>
</table>

The tubes were spun to mix and incubated at 37°C for five minutes. Four μl of formamide dye mix was added to each tube and samples heated to 80°C for 3 minutes just before gel loading.

This protocol was adapted for sequencing intact phage DNA or supercoiled cosmid
DNA in suitable circumstances. Both adaptions followed the plasmid sequencing protocol for reagents. Cosmid sequencing required more template DNA (> 10 μg) to yield comparable results but was otherwise very similar to plasmid sequencing. Phage DNA also required more template, and in addition heat denaturation rather than alkali denaturation was used as the DNA was already linear. Both modifications were useful on occasion, but never provided the quality of data that plasmid sequencing yielded.

1.3 Polyacrylamide Urea Gel Electrophoresis

Sequencing gels were run on 50 cm glass plates in TBE buffer (TBE - section 6.2). One glass plate was coated with MPTS (30 μl gamma-methacryloxypropyltrimethoxy silane [Sigma] in 10 ml EtOH + 300 μl 10% glacial acetic acid). This treatment ensures adhesion of the gel to the plate. The second plate was silane-coated with 2% dimethylchlorosilane in trichloroethane [BDH] to ensure that the gel adhered to the MPTS plate only. The plates were taped together, separated by 0.5 mm PTFE spacers and the gel mix added by pipette. The gel mix contained (per 200 ml):

- 84g urea
- 30ml 40% acrylamide stock (38% acrylamide, 2% bisacrylamide)
- 40ml 5X TBE, to 200ml with ddH₂O
- 180 μl 10% ammonium persulphate
- 90 μl TEMED added per 40 ml of gel mix immediately before pouring.

After setting, the gel was put onto a vertical gel apparatus, 1X TBE buffer added, and the comb carefully removed. The wells were rinsed with buffer immediately before running to remove urea, and three ul of reaction mix added per well. Samples were electrophoresed at 27 watts for 90 to 100 minutes for a standard run. Fixing was carried out in 10% methanol 10% acetic acid for 10 minutes, and the gel was then washed thoroughly (> 5 min) in running water and dried at 80°C. Overnight exposure to Kodak XAR-5 film was generally sufficient to read 150 to 200 nucleotides without ambiguity. Sequences
were normally read directly into \textit{vi} files on the Vax and were then processed using the \textit{dnaprog} facility, written by P. Jeppesen (MRC HGU). Later sequences were occasionally read in to the Staden system, using a Graflex digitiser.

2. LIBRARY SCREENING

2.1 \textit{Plasmid and Cosmid Library Screening}

There were no major differences in screening plasmid or cosmid libraries during the project, although plating density was often higher for cosmids. Screening was normally carried out on 20 x 20 cm petri-dishes as this allowed plating of large numbers (2 \times 10^4 to 1 \times 10^5) of bacterial colonies. The method of Grunstein and Hogness was used (Maniatis 1982). The library was plated on a 20 x 20 cm Nitrocellulose (Schleicher and Schuell) or Hybond-N (Amersham) filter which was laid dry on a freshly poured L-Agar (containing selective antibiotic) 20 x 20 cm dish. After incubation at 37°C overnight, replica "slave" filters (normally two) were taken from the master filter, marking paired filters with waterproof ink to enable accurate keying of autoradiographs to the master plate. All three filters were then reincubated on L-Agar and antibiotic for four to five hours, to permit regrowth of the colonies.

Slave filters were then processed by laying onto saturated 3MM (Whatman) filter paper in 20 cm dishes:

\begin{itemize}
  \item Membrane Lysis \hspace{1cm} 10\% SDS soln \hspace{1cm} 2 minutes
  \item DNA Denaturation \hspace{1cm} 0.5M NaOH, 1.5M NaCl \hspace{1cm} 3 minutes
  \item Neutralization \hspace{1cm} 1.5M Tris-Cl, 1.5M NaCl (pH 5.5) \hspace{1cm} 3 minutes
  \item Filter Wash \hspace{1cm} 2xSSC \hspace{1cm} 3 minutes
\end{itemize}
After removing bacterial debris by gentle agitation, filters were blotted dry and then fixed by baking (Nitrocellulose: 2 hrs, 80°C) or UV irradiation (Hybond-N: 3 minutes on UV-transilluminator). Filters were prehybridised and hybridised in Quick-Hybe as described in Southern Blotting (Section 7.). After auto-radiography for the required period (1-3 days at -70°C), films for both slave filters were transposed and compared. Duplicate spots were selected, and relevant areas isolated from the master filter (stored at 4°C during probing). Filter fragments were put into 1 ml L-Broth/15% Glycerol and vortexed to solubilise colonies. They could then be stored frozen indefinitely. Secondary screens were carried out by plating small (1 x 10^-3) aliquots of the primary pick stock. These were generally plated on gridded Nitrocellulose filters in 9 cm petri-dishes. The process of replicating, regrowing, denaturing, and probing was carried out as for the primary screen. Depending on the density of colonies on the secondary plates, individual colonies could be selected at this stage, or a tertiary screen performed.

2.2 Phage Library Screening

Phage Library screening was similar in methodology to plasmid screening. Libraries were plated on 20 x 20 cm plates at similar densities to cosmids (maximum of 1 x 10^5 plaques per plate), normally on L-Agar medium. Fresh E. coli of the appropriate strain (see section 5), previously grown to stationary phase in L- Broth + 10mM MgCl_2 + 0.2% Maltose, were spun down and resuspended in 0.5 volumes of 10mM MgCl_2. Phage stocks (previously titred) were added to 1 ml of the bacterial suspension in a 50 ml Falcon Tube and incubated at room temperature for 15 minutes. The cells were then diluted with 40 ml of molten Top L-Agar (45°C), swirled to mix, poured over a dried L-Agar petri dish, and incubated at 37°C overnight. Nitrocellulose or Hybond-N filters were laid, dry, onto the top agar, and keyed with a needle containing waterproof ink. The first filter was left on the plate for 1 minute, subsequent filters for 2 to 3 minutes.
The filters were then processed exactly as for plasmid screening, omitting the 10% SDS Membrane Lysis step. After hybridisation, and selection of duplicate positives, relevant areas of top agar were cut out and put into 1 ml of SM (50 mM Tris, 5 mM MgCl₂) in an eppendorf. This stock could be stored at 4°C for several months. Aliquots of the stock were then used for secondary screens, as for plasmids, on 9 cm petri-dishes.

3. DNA PREPARATION

3.1 Plasmid DNA Preparation

Purification of closed-circular supercoiled plasmid DNA from bacterial cultures is a central technique in molecular genetics. For analysis of several independent colonies from a cloning experiment, a rapid, efficient method is required to isolate small (µg), relatively pure amounts of plasmid DNA. For the preparation of milligram quantities of highly pure, RNA and E.coli DNA-free plasmid for sequencing, T7 or SP6 polymerase transcription, or vector construction, a more elaborate protocol was employed.

Small scale plasmid preparation.

Bacterial colonies were added to 15 ml of L-Broth (see Section 5) plus appropriate antibiotic and shaken at 37°C overnight in 50 ml Falcon tubes. The tubes were spun at 3000 rpm for 10 minutes to pellet the bacteria, and pellets then resuspended in 1 ml of TELT with 1 mg Lysozyme (Sigma) TELT is 50 mM Tris-Cl pH 7.5, 62.5 mM EDTA, 2.5 M LiCl, 0.4% Triton-X100. After transfer to Eppendorf tubes, the samples were boiled for 1 minute. After 5 minutes incubation on ice, the tubes were spun in a microfuge for 15 to 30 minutes to pellet E.coli debris and denatured chromosomal DNA. Between 0.5 and 0.8 ml of supernatant was removed, and the plasmid DNA precipitated with 0.6 volumes of Isopropanol. After spinning for 15 minutes, the plasmid pellet was
washed with 80% ethanol, dried and resuspended in 50 μl TE. This method yielded 20 to 50 μg (with pUC based plasmids) of DNA, of sufficient purity for restriction enzyme analysis, and for use in nick-translation or random-priming reactions (Section 7). It was not always sufficiently pure for sequence analysis. It was also unsuitable for cosmid minipreps, since boiling tends to irreversibly denature circular DNA molecules of larger (50 kb) size.

Large scale plasmid preparation.

Bacterial colonies were inoculated into 250 ml of Terrific Broth or into 500 ml of L-Broth plus appropriate antibiotic, and shaken at 37°C overnight. The culture medium was decanted to 250 ml Sorvall centrifuge tubes and spun at 8000 rpm for 10 minutes. The pellets were resuspended in 5 ml of cold GTE (50 mM glucose, 50 mM Tris pH7.5, 10 mM EDTA), and 5 ml cold GTE + 10 mg lysozyme added. After 15 minutes on ice, 20 ml of alkaline SDS (0.25 N NaOH, 1% SDS) was stirred in, to denature chromosomal DNA. After a further 10 minutes, 15 ml of high salt soln.(3M Potassium Acetate, 2M Glacial Acetic Acid pH 4.5) was added to precipitate *E. coli* debris and chromosomal DNA. The supernatant was filtered into a 50 ml Falcon tube through muslin to remove the coarse precipitate, and 0.6 volumes of isopropanol added. Samples were left for 15 minutes at room temperature, and centrifuged at 3000 rpm for 10 minutes to pellet the plasmid DNA. Pellets were washed with 80% ethanol, lyophilized until almost dry, and redissolved in 3.8 ml TE. After addition of 4.2 g CsCl (BCL) and 200 μl ethidium bromide (10 mg/ml, Sigma) tubes were left in the dark at room temperature for 30 minutes to allow precipitation of any remaining bulk protein. This was removed by spinning at 3000 rpm for 5 minutes, and the optical density of the DNA solutions then adjusted to between .930 and .945 by addition of CsCl. The samples were spun at 40000 rpm overnight in a Sorvall Ultracentrifuge. Plasmid bands were removed with a needle or
fine-tipped pastette, and ethidium extracted with butanol. DNA was then precipitated by addition of 2.5 volumes of 80% ethanol, DNA spooled when possible, washed in 80% ethanol, dried, and resuspended in TE.

This method was suitable for cosmid DNA preparation, and a scaled down version was used (minus the CsCl gradient steps) to prepare minipreps of cosmid DNA when these were required. Plasmid DNA purified by CsCl gradient centrifugation was capable of essentially any enzymic manipulation: sequencing, RNA polymerase transcription and for use in novel constructs.

3.2 Bacteriophage DNA Preparation

As with plasmid preparations, two forms of bacteriophage DNA preparation protocol were used; a small scale, rapid method yielding 1 to 10 μg of DNA and a larger scale protocol yielding 50 to 100 μg DNA. These were very similar to each other and only the large scale one will be described in detail.

Plating bacteria of a suitable strain were grown in L-Broth + 10 mM \( \text{MgCl}_2 \) + 0.1% maltose at 37°C overnight. These were spun down at 3000 rpm and resuspended in 0.5 volume 10 mM \( \text{MgCl}_2 \). About \( 5 \times 10^5 \) phage were added to 1 ml of cells, and incubated at RT for 15 minutes. Molten top-agarose (40 mls at 42°C) was added to the bacteria and plated onto a 20 cm x 20 cm pre-poured L-Agarose petri-dish. This was incubated at 37°C overnight, without inversion. The plate was inspected for complete lysis of bacteria; if this had occurred, 20 mls of SM was added to the plate, which was then agitated at room temperature for five hours. The SM was pipetted into a 50 ml Falcon tube and spun at 3000 rpm for 20 minutes to pellet agarose aggregates. The supernatant was carefully decanted, 0.1 mg/ml DNase and 0.1 mg/ml RNAse added, and incubated at 25°C for 30 minutes. An equal volume of 20% PEG 6000 (Polyethylene glycol MW
6000, Sigma), 2M NaCl was added to the tube to precipitate phage particles, and the samples left on ice for 2 hrs. Prolonged precipitation times (overnight) increased final yields. Loose precipitates were decanted into 13 ml tubes and spun at 10000 rpm for 20 minutes to pellet the phage. The pellet was then resuspended in 4 ml 50mM Tris-Cl pH 7.5 10 mM MgCl₂ and extracted three times with chloroform to remove all traces of PEG. The phage particles were then disrupted by the addition of EDTA, pH 8.0 to a final concentration of 20 mM followed by rapid extraction with phenol. Following a second phenol extraction and two chloroform extractions, the phage DNA was precipitated by addition of NaCl to 0.2M and two volumes of ethanol. DNA was spooled onto a glass rod when possible, washed with 80% EtOH and resuspended in TE. This bacteriophage DNA was sufficiently pure for most enzymic manipulations, and was used for direct DNA sequencing on occasion.

Oligonucleotide Synthesis

Oligonucleotides were synthesised and purified according to manufacturers instructions on an Applied Biosystems 381A Oligonucleotide Synthesiser.

4. RNA ISOLATION

Analysis of rodent liver RNA formed an important part of this project. I was fortunate to have access to large stocks of RNA already prepared by R. Hill and R. Meehan. These included all the inbred Mus domesticus and Rattus stocks used in the project. J. Ansell (Zoology Dept. Edinburgh University) kindly provided Mus caroli animals. F. Berger (University of South Carolina) kindly provided samples of total liver RNA from M. hortulanus, M. cookii, M. cervicolor, M. pahari, M. saxicola, and M. minutoides. Specimens of Apodemus sylvaticus were trapped at the Bush Estate, Roslin
using Longworth Traps, with the assistance of R. Hill and J. Ansell.

4.1 Total Mouse Liver RNA

This protocol was carried out as previously published (Hill et al., 1985). Whole liver was quickly removed from freshly killed mice and homogenized in 20 ml of 8M Guanidinium Hydrochloride (BRL) in TE, for 45 seconds, using a motorized tissue mixer. The suspension was centrifuged for 10 minutes at 8000 rpm, and the supernatant carefully removed. Half a volume of cold ethanol was added, and after storage at -20°C for 30 minutes, RNA was pelleted by centrifugation at 10000 rpm for 15 minutes. The pellet was resuspended in 15 ml of 6M Guanidinium Hydrochloride in TE, and reprecipitated with 0.5 volumes cold EtOH. After 30 minutes, the RNA was repelleted, and the 6M extraction repeated. The RNA pellet was resuspended in 10 mls of DEPC Water (0.1% Diethylpyrocarbonate in double distilled water, autoclaved). Resuspension often required use of a Dounce homogeniser. NaCl was added to a concentration of 0.2M, and RNA precipitated with 2.5 volumes of ethanol. This precipitation was repeated, and the RNA was finally resuspended in 5 mls of DEPC water and stored at -70°C after measuring its concentration by absorbance at 260 nm. The RNA was sufficiently free of inhibitors at this stage to be used for reverse transcriptase sequencing.

4.2 Mouse Liver poly (A)+RNA

This followed the method of Aviv and Leder (1972). Total RNA (1 to 2 mg) was diluted to 1 mg/ml and heated at 65°C for 5 minutes. An equal volume of 2X loading buffer (20mM Tris.Cl pH 7.5, 0.5M NaCl, 1mM EDTA, 0.1% SDS) was added to the RNA. An oligo (dT) cellulose (Sigma) column (1 ml volume) was poured and equilibrated with loading buffer. The RNA was then applied to the column and the eluate reheated, and reapplied. The column was then washed with 5 volumes of loading buffer,
and then with buffer containing 0.1M NaCl. The flowthrough buffer was retained at this stage. Elution buffer (10mM Tris.Cl pH 7.5, 1mM EDTA, 0.1% SDS) was then added (3 volumes) and 0.5 ml fractions collected. Each fraction was checked for RNA by spotting onto a dried agarose/ethidium bromide plate and visualising with UV. Positive fractions were pooled and precipitated with ethanol, and the poly (A)+ pellet resuspended in DEPC water at 1µg/µl, and stored at -70°C.

5. SUBCLONING

This section deals with general DNA subcloning resources and methods which were used with numerous variations during the project. The main reference for these is Maniatis et al (1982) Molecular Cloning, A Laboratory Manual, but protocols and methods were frequently updated and modified during the project.

5.1 Cloning Vectors

The most frequently used plasmid during this work was pBS (formerly Bluescribe, Stratagene) a 3 kb general-purpose cloning vector derived from pUC 19 (Messing, 1983). It possesses a number of useful features including high copy number in E. coli, the lac Z blue-white complementation test for inserts, the pUC19 Multiple Cloning Site containing 13 unique cloning sites, and Phage T3 and T7 promoter sequences flanking the polylinker, enabling production of specific single stranded RNA transcripts. Other less frequently used subcloning vectors were pSP64, pTZ18R, and for expression experiments, pJLA502 (gift of J. McCarthy), ptac85 (gift of J.Myles), and pSVL (Pharmacia). Clones isolated from the DBA cDNA library and from B. Hill were in pBR322, these were transferred to Bluescribe where necessary. The 129/St mouse cosmid library (gift of L. Stubbs, ICRF) was cloned in pcosEMBL2 (Ehrich et al., 1987). The DBA mouse
genomic library (from R. Hill) was cloned in λEMBL3. The bacteriophage cDNA libraries (R. Hill and Clontech) were cloned in λgt10.

5.2 Bacterial Strains

The most frequently used strain of *E. coli* used was JM83 (Messing, 1983), as it is readily made transformation competent by a number of procedures, has a stable phenotype, and does not require IPTG to induce the lac Z promoter on the Bluescribe plasmid. Its genotype is: *E coli* JM83 ara, *lacpro*, strA, thi, o80d, lac Z *M15

For screening phage libraries, *E. coli* LE392 and BNN102 were used as recommended in Maniatis (1982). For *E. coli* expression of foreign proteins, several bacterial strains were employed, notably CAG629 (gift from J. Myles), a protease deficient strain. Its genotype is CAG629: lacZ (am), trp (am), pho(am), supC<sup>ts</sup>, mal, rpsL, lon, htpR

5.3 Bacterial Media

*E. coli* were normally grown on LB (Luria-Bertani) agar plates or in LB medium.

LB medium, per litre (pH 7.5): 10g bacto-tryptone (Difco), 5g bacto-yeast extract (Difco), 10g NaCl. LB Agar: LB medium plus 1.5% Agar (Difco). LB Top Agar: LB medium plus 0.7% Agarose (Sigma).

H agar plates were used when the lac Z blue-white test was used to detect inserts, in conjunction with 0.02% BCIG (5-bromo-4-chloro-3-indolyl-β-D-galactoside; stock solution = 2% in dimethylformamide). H agar, per litre (pH 7.3): 10g bacto-tryptone (Difco), 8g NaCl.

Terrific Broth (Tartoff and Hobbs, 1987) was used latterly, for large scale plasmid preparations as it gave considerably higher yields of DNA than LB medium. This is presumably due to the supplementation of phosphate, and of an additional rich carbon source, glycerol. Terrific Broth, per litre: 15g bacto-tryptone, 30g bacto-yeast extract,
5 ml Glycerol. 1/10 volume of 1M $K_2HPO_4$ added before inoculation.

Antibiotics to select for plasmid maintainance were used at the following concentrations:
Ampicillin (Sigma) 100μg/ml, Tetracyclin (Sigma) 25μg/ml, Kanamycin (Sigma) 30μg/ml. Stocks were 1000X and stored at -20°C.

5.4 Subcloning DNA Fragments

Subcloning of DNA fragments into plasmid vectors for further analysis was one of the most frequently performed procedures during the project. DNA fragments of interest were either isolated by agarose electrophoresis and purified before ligation or "shot-gunned" into a plasmid mini-library and the correct clone identified by plasmid miniprep or by Grunstein-Hogness hybridisation. In either case, the purity of the insert DNA was the most important factor in the success of ligations, as T4 DNA ligase (BCL) appears to be strongly inhibited by agarose contamination. This phenomenon is much more pronounced in blunt-end ligations, which never became routine. Elu-tips (Schleicher and Schuell), Spin-X columns, and DEAE-paper (Whatman) were all successful at times, but never totally reliable for blunt-ended DNA. For "sticky-end" (overlapping cohesive termini) ligations, use of GTG or LMP Agarose (Seakem) and careful sequential phenol (twice), and chloroform (twice), extractions usually provided DNA of sufficient purity. Vector to insert ratios varied depending on insert size, but was usually 1:5 on a molar basis. 25 ng of linear vector was ligated to insert in a 10μl reaction volume with 1μl T4 DNA ligase. Ligation was at room temperature for 2 to 4 hours. Ligations could be stored indefinitely at -20°C before transformation.
5.5 Bacterial Transformation

_E. coli_ were normally made competent for DNA uptake by the method of Simianis (Hanahan, 1985), and could be stored for at least six months at -70°C at an efficiency of 5 x 10^6 colonies per µg of supercoiled plasmid DNA. The strain JM83 was normally used (Section 5.2) but other strains adapted well to this method. Cells were grown to an A_260 of 0.5 in LB medium + 10mM MgCl₂. After chilling on ice and pelleting at 3000rpm for 10 min, cells were resuspended in 0.3 volumes of RF1 solution (RF1 = 100mM Rubidium chloride, 45mM manganese chloride, 30 mM potassium acetate, 10mM calcium chloride, and 15% w/v glycerol, pH 5.8) and left on ice for 30 minutes. The cells were respun and resuspended in 0.1 volumes of RF2 solution (RF2 = 10mM MOPS, 10 mM rubidium chloride, 75 mM calcium chloride, and 15% w/v glycerol, pH 6.8). After 15 min on ice, the cells were flash-frozen in liquid nitrogen in 200µl aliquots and stored at -70°C. For transformations, the cells were thawed at room temperature and placed on ice. Ligated DNA (< 10µl) was added to the cells and left to adsorb on ice for up to 1 hour. After a 90 second heatshock at 42°C, the cells were chilled on ice for 1 minute, 0.8 mls of LB medium added, and the cells allowed to develop antibiotic resistance for 1 hour at 37°C. Suitable aliquots of the transformation were then plated on selective media and grown overnight at 37°C.

This transformation procedure proved very reliable throughout this work and methods producing more highly competent _E. coli_ (eg Hanahan, 1983) were rarely resorted to.
6. SOUTHERN AND NORTHERN BLOTTING

6.1 Genomic DNA Digests

Genomic DNA was stored at 4°C until required. Routinely, 5μg of DNA per gel lane was digested with restriction enzyme, before electrophoresis. Enzyme digests were carried out at 37°C according to manufacturers instructions. The standard DNA concentration for digestion was 0.1 to 0.2 mg/ml, with an enzyme excess of 3 units per microgram. Digestion was for a minimum of four hours, for mammalian DNA, and normally overnight. The standard digestion conditions were 50 mM Tris-Cl (pH 7.5), 100 mM NaCl, 10 mM MgCl₂ and 5 mM spermidine. Reactions were halted with Stop Mix (30% Ficoll, 0.25% Orange G, 0.5M EDTA, and 10X TAE buffer) and stored at room temperature before loading.

6.2 Agarose Gel Electrophoresis - DNA

Agarose Gels were either run in TAE (Tris-acetate) or TBE (Tris-borate) buffer. 20X TAE = 0.8M Tris-Cl, 0.4M NaOAc, 20mM EDTA, pH adjusted to 8.2 with Glacial Acetic Acid. 10X TBE = 1M Tris-Cl, 0.8M Boric Acid, 20mM EDTA, pH 8.3. For Southern blotting, large (25 cm x 20 cm, 250 - 300 ml) gels were run at 40V overnight in TAE buffer. Up to 24 lanes per gel could be loaded, with phage λ DNA digested with HindIII used as a size marker. The concentration of agarose was varied, depending on the size of DNA bands expected, but was normally about 0.8% in 1X TAE. In the cosmid partial mapping experiments the gels were about 0.4% agarose in TAE. Mini gels (20 ml volume) were normally run in TBE as this buffer is less prone to overheating at higher voltages; these gels had ten lanes and were used for rapid subcloning analysis. The size marker for mini gels was PhiX 174 DNA digested with HaeIII.
6.3 Southern Transfer

This broadly follows the method of Southern (1975). Following photography of a gel, the DNA was denatured by immersion in three gel volumes of denaturing solution (0.5M NaOH, 1.5M NaCl) for forty minutes. This was followed by thirty minutes in neutraliser solution (1M Tris-Cl, 1.5M NaCl pH 5.5). The gel was then laid on a transfer apparatus, a wick of Whatmans 17MM filter paper with a reservoir of 20X SSC (20X SSC = 3M NaCl, 0.3M Na citrate , pH 7.4.) A nitrocellulose (Schleicher and Schuell) or Hybond-N (Amersham) membrane was carefully placed on the gel, avoiding any air bubbles. Two sheets of Whatmans 3MM filter paper were placed over the filter, followed by paper towels to a depth of 5 cm. A weight (about 1 kg) was placed on the assembly, and transfer carried out for at least eight hours. The membrane was then baked for two hours at 80°C (nitrocellulose) or UV irradiated for three minutes (Hybond-N). For mini-gel blotting, the denature step was reduced to fifteen minutes, the neutraliser step omitted, and transfer time reduced to thirty minutes.

6.4 RNA Electrophoresis and Northern Blotting

RNA was also electrophoresed in agarose, with the inclusion of 18% formaldehyde to ensure that secondary structure in the RNA (a cause of mobility anomalies) was minimised. Up to 10μg of total RNA or 2μg of polyA+ RNA was loaded per gel track. RNA was denatured for 10 minutes at 45°C in 2 volumes of 1.5X Sample Buffer (300μl deionised formamide, 108μl formaldehyde solution, 6μl 1M phosphate buffer). The gel was prepared in 10mM sodium phosphate buffer (1M NaPO₄ buffer pH 6.5 = 70% 1M NaH₂PO₄ + 30% 1M Na₂HPO₄. 1.5g of agarose was melted in 82 mls of buffer, and after cooling to 55°C, 18ml of formaldehyde solution (BDH) was added and the gel poured quickly. 0.1 volumes of Loading Mix (0.1% bromophenol blue in Stop Mix) was added to the RNA samples, and the gel run at 40V (15V overnight), after loading. The
low voltage was necessary as the gels were prone to buffer failure at high current or temperature and this degraded the RNA. Buffer failure was detected by change of the bromophenol blue to a yellow colour. The gel was soaked in 1 litre of 2X SSC for 2 hours to remove the Formaldehyde, and any marker lanes were removed for ethidium staining. The remainder of the gel was blotted on a Northern transfer apparatus (as Southern but with 10X SSC in the reservoir) for at least eight hours. Filters were then processed as for Southern blots.

6.5 Hybridisation

Pre-hybridisation and hybridisation for Southern and Northern filters were carried out in Quick-Hybe hybridisation mix. This consisted of 5X SSC, 0.1% bovine serum albumin, 0.1% ficoll, 0.1% polyvinylpyrrolidine, 0.5% SDS, 0.5% sodium pyrophosphate and 100μg ml denatured salmon sperm DNA. It was stable for several months at room temperature. Filters were sealed in plastic bags with about 1 ml Quick-Hybe per 20cm² of filter. After prehybridisation at the appropriate temperature for at least twenty minutes, the probe was carefully added to the bag, and hybridised for an appropriate period. For genomic blots, total RNA blots, and primary library screens this was normally overnight, but under more favourable conditions could be reduced to two hours. The bag was then carefully opened in a 2X SSC wash (4X SSC for oligonucleotides) at the hybridisation temperature, the probe discarded to drain, and the filters washed at the appropriate stringency until the signal reached an acceptable level. Filters were then wrapped in clingfilm, keyed with radioactive pen, and autoradiographed at -70°C with Kodak XAR-5 film. Films were developed in a Fuji RGB-2 Automatic Film Processor.
7. **NUCLEIC ACID PROBES**

7.1 *Nick Translation of DNA*

This technique was used in early stages of the project. It is essentially the method of Rigby (1977). DNA (100 ng to 300 ng) incubated with DNA polymerase I from *E. coli* and bovine pancreatic DNAse I in the presence of a radioactively labelled nucleotide, and the three other unlabelled nucleotides, will incorporate label as the polymerase removes and replaces tracts of DNA initiating at single stranded nicks produced by the DNAse. The reaction is carried out at 15°C which permits template extension by Pol I, but prevents excessive DNA degradation by DNAse. Reactions were carried out in 10 μl with 1 μl 10X Salts (500 mM Tris-Cl (pH 7.8), 50 mM MgCl₂, 0.2 mM dATP, dCTP, dGTP, and 100 μg/ml bovine serum albumin), 100 ng DNA in 5 μl ddH₂O, 3 μl (=30 μCi) ³²P dTTP, and 1 μl of DNA polymerase I/DNAse I mixture (0.4 units/μl). This was incubated at 15°C for 1 hour, and incorporation of radioactivity measured by TCA precipitation. Successful reactions had 100 μl of denatured salmon sperm (10 mg/ml stock) added to them and were boiled for ten minutes immediately before addition to the prehybridised filter.

7.2 *Random Priming of DNA*

This technique was used later in the project as it was more convenient (BCL Random Priming Kit), more reliable, and gave higher specific activity of probe than nick translation. The technique is based on the method of Feinberg and Vogelstein (1983), and uses random hexanucleotides to act as primers for polymerase chain extension in the presence of radiolabelled nucleotides. As little as 10 ng of DNA could be used, even in an agarose fragment. Routinely, 50 ng of DNA (plasmid or insert), was denatured in 10 μl TE for 10 minutes by heating to 100°C. The sample was chilled on ice, and 5 μl of
random-prime buffer/dNTPs was added, followed by 4μl $^{32}$P dCTP, and 1μl Klenow fragment of DNA Polymerase I. Random-prime buffer consists of hexanucleotides in 4X Klenow Buffer, and 200μM dATP, dGTP, dTTP. The reaction was incubated at 37°C for 1 hour, and incorporation checked by TCA precipitation. If satisfactory incorporation (> 20% of free counts) was obtained, 100μl of denatured salmon sperm DNA (10 mg/ml) was added, the probe heated to 100°C for 8 minutes, and added to the prehybridising filter.

7.3 Oligonucleotide Labelling

Oligonucleotides were radiolabelled for use in sequencing or hybridisation experiments by T4 polynucleotide kinase. This enzyme transfers the $\gamma$-phosphate from $^{32}$P $\gamma$-ATP to the 5' OH group of a dephosphorylated oligonucleotide. For sequencing or probe reactions, 30ng of oligonucleotide in 4μl TE, 1μl 10X kinase buffer (0.5M Tris-Cl pH 7.5, 60mM MgCl$_2$, 50mM DTT), 3μl $^{32}$P $\gamma$-ATP, and 1μl T4 polynucleotide kinase was incubated at 37°C for 30 minutes. Incorporation was checked, if necessary, by chromatography on DEAE paper in 3M ammonium formate. 100μl of denatured salmon sperm was then added to probe reactions, and the mixture added to the prehybridisation. Oligonucleotides for sequencing proceeded into the sequencing protocol. Occasionally, oligonucleotides were used for Southern Blots. In these experiments, 80μCi rather than 30μCi of label was used to obtain a higher specific activity of probe.

7.4 Synthetic RNA for Hybridisation and Translation

RNA probes were produced in vitro using the SP6 and T7 phage polymerase systems (Kreig and Melton, 1984; Melton et al, 1984). Briefly, a plasmid containing the insert of interest and a flanking SP6, T3, or T7 promoter sequence (eg pBluescribe) was linearised with a restriction enzyme distal to the insert. After phenol and chloroform
extractions, DNA was incubated with appropriate polymerase, and ribonucleotides, at 37°C for an hour. Inclusion of a radio-labelled (\textsuperscript{32}P or \textsuperscript{35}S) nucleotide (minimum concentration 10\(\mu\)M to prevent truncated transcripts) produced a single stranded probe of high specific activity. (The template can then be removed by DNAse digestion if required, but this was not always necessary). Large amounts of unlabelled capped RNA can be produced by the inclusion of a m\textsuperscript{7}GpppG dinucleotide cap analogue in the reaction mixture (Contreras et al, 1982). The Transprobe T (Pharmacia) kit was used to produce mRNA for \textit{in vitro} translation (section 8.2) by this method, according to manufacturers instructions.

8. PROTEIN ANALYSIS

8.1 SDS Polyacrylamide Gel Electrophoresis

SDS PAGE gels were run using an LKB (Pharmacia) system (Laemmli, 1970). For optimal separation of proteins in the 40kD to 50kD size range, 11% gels were normally poured. Gels were composed of 13ml 30% acrylamide stock (58g acrylamide, 2g bis-acrylamide in 200ml H\textsubscript{2}O), 9mls of separating gel buffer (1.5M Tris-Cl pH 8.8, 0.5% SDS) and 14ml of ddH\textsubscript{2}O. 50\(\mu\)l of TEMED (N N' tetramethylethylene diamine) and 100\(\mu\)l of 10% ammonium persulphate were added just before pouring to polymerise the mix. Sufficient space for the stacking gel was left at the top of the gel, which was covered in water saturated isobutanol while it set. Stacking gel consisted of 1.5ml of acrylamide stock, 2.5ml of stacking gel buffer (0.5M Tris-Cl pH 6.8, 0.5% SDS) and 5.7ml ddH\textsubscript{2}O. 30\(\mu\)l ammonium persulphate and 10\(\mu\)l TEMED were added just before pouring over the separating gel which had been rinsed to remove unpolymerised acrylamide and isobutanol. After setting, wells were thoroughly rinsed before running to remove unpolymerised stacking gel acrylamide. Gels were run in electrophoresis buffer.
(10x EB = 0.5M Tris-Cl, 0.5M glycine, 10% SDS). Protein samples were added to 2 volumes of smash buffer (10% stacking gel buffer, 2% SDS, 5% β-mercaptoethanol, 10% glycerol, 0.005% bromophenol blue) and boiled for 5 minutes before loading. Gels were run at 6 to 10 mA overnight, until the dye-front reached the end of the gel. After soaking in destain (45% methanol, 45% H₂O, 10% acetic acid) gels were either stained with Coomassie Blue (Sigma) to visualise the protein bands, or soaked in Amplify (Amersham) and dried down for autoradiography.

8.2 Translation of mRNA with Reticulocyte Lysate

Translation of in vitro produced mRNA from full length cDNA sequences was performed using the Amersham Rabbit Reticulocyte Lysate kit N.90, according to manufacturers' instructions. In the presence of added mRNA, and radiolabelled L-[³⁵S] methionine, the lysate will synthesize specifically the mRNA-encoded proteins. A typical reaction included 40μl of lysate mix, 7μl (=70μCi) of radiolabelled methionine (Amersham SJ.204) and 3μl (=1μg) of capped synthesized mRNA. After incubation at 30°C for 70 minutes the incorporation of label was measured by TCA precipitation and the samples were subjected to SDS PAGE and autoradiography, with or without prior incubation with putative target proteinases.
CHAPTER THREE

The Evolution of Contrapsin in the Genus Mus

Introduction

Hill and Hastie (1987) demonstrated that accelerated evolution had occurred specifically within the reactive centre regions of four Spi-2 cDNAs. These were contrapsin, from the laboratory mouse strain Swiss, Spi-2.1 and Spi-2.2 from inbred R. norvegicus (Buffalo strain), and human α1-antichymotrypsin. The nucleotide sequences of rat Spi-2.1 and contrapsin were as divergent (33% identity) from one another in the reactive centre region as either sequence was from α1-antichymotrypsin (31% and 41% identity respectively). It was assumed that the contrapsin and rat Spi-2.1 sequences were orthologous as they shared P1 residue specificity (arginine), had a very similar mRNA size, and were transcribed in the liver at similarly high levels (~0.5% of mRNA). If a consensus divergence time of fifteen to twenty million years is accepted for rat and mouse (Figure 3.1), their rate of evolution during this period is calculated to be higher than that of neutral regions such as pseudogenes or 3' untranslated sequences. A computer analysis to calculate the rate of nonsynonymous (K_A) and synonymous (K_S) substitutions (Li et al., 1985) showed that the K_A values for the reactive centre, Region 2, were four-fold higher than the K_S values for the 5' coding sequence, Region 1. By several criteria, including K_S/K_A ratios in the remainder of the gene, expression levels of the Spi-2 transcripts in the liver, and probable inhibitory activity for both contrapsin (Takahori and Sinohara, 1982) and rat Spi-2.1 (Le Cam et al., 1987), the two genes appear to be active. It was suggested by Hill and Hastie that the genes are diverging in the reactive centre region at a rate which involves positive Darwinian selection for mutation.
The neutral theory of molecular evolution (Kimura, 1983) predicts that the great majority of observed substitutions during sequence divergence are due to random occurrence of mutations in a DNA sequence; that is unselected changes. Advantageous mutations are predicted to be comparatively rare, and disruptive mutations to be more abundant but subject to negative selection, and therefore ultimately lost from a population. The majority of observed mutations are considered to be neutral, due to codon degeneracy, conservative amino acid substitutions, or substitutions in domains which do not perturb protein function. Thus the majority of sequence polymorphism in a population will be due to the ongoing process of random fixation of alternative neutral alleles. The theory has become widely accepted, as a large body of sequence data indicates that a considerable amount of mutation is effectively neutral (Li et al, 1985).

However, there are examples of sequence analysis which are effectively incompatible with the neutral theory. Hudson et al (1987) have shown that the level of synonymous polymorphism in the Drosophila melanogaster alcohol dehydrogenase (Adh) gene is four fold higher in the coding region than in the 5' flanking region. This appears to be at odds with the neutral theory, which would predict that a coding region should be under greater evolutionary constraint than a non-coding flanking sequence, and should therefore contain less neutral variation. Comparison with a separate fruitfly species showed that the observed difference was not due to variation in mutation rate between coding and non-coding regions, but rather an "excess" of polymorphism in the Adh coding region. It would appear that there is positive selection for polymorphism in this region. In fact, it has long been held that there is selection for a balanced polymorphism for two electrophoretically distinct Adh alleles (Greenough and Harvey, 1987)

Given the substitution rates determined for the rodent Spi-2 genes, the neutral theory would predict that the reactive centre region of the Spi-2 genes should be highly
polymorphic, both within and between species. Either the reactive centre region is not under stringent compositional selection, or the genes lack an inhibitory function, and the reactive centre, as an external loop, would be free to evolve at almost pseudogenic rates. This neutralist explanation of the Spi-2 data has been advanced by Graur and Li (1988), as discussed in Chapter 1.

However, if positive Darwinian selection was acting upon specific nonsynonymous reactive centre substitutions, the reactive centre might not be expected to be highly polymorphic within a species, or to have diverged rapidly between closely related species. This would assume rapid fixation and subsequent stabilising selection for an unchanging inhibitory specificity. Conversely, if the selective agent were a co-evolving parasitic proteinase, frequency dependent selection might cause pronounced intra-species reactive centre polymorphism. This type of process has received a theoretical treatment (Clarke, 1976), and has been proposed as a possible cause of the pronounced polymorphism of certain MHC sequences (Bodmer, 1972). A recent analysis suggested that MHC polymorphism may instead be due to overdominant selection, ie heterozygote advantage (Hughes and Nei, 1988).

A frequency dependant selection scenario has been suggested as a possible source of the high synonymous substitution rate found for the Spi-2 reactive centres (J. Howard, personal communication), since fixation might be retarded by frequency dependent selection, allowing linked silent mutations to occur. An alternative explanation for the high synonymous substitution rate may be simply that the reactive centre sequences are too short and too divergent to provide suitable input for the computer program in its present form.
Figure 3.1 Consensus phylogeny for rodents used in this study. Earlier divergence times are particularly uncertain, especially the date of *Rattus/Mus* divergence. Areas of uncertainty about divergence times or branching order are shown as dashed lines, although the overall structure is broadly accepted. The timescale below is also a consensus, with contrasting estimates of up to two-fold higher or lower. The tree is based on data from mitochondrial DNA, ribosomal spacer DNA, immunoprecipitation, electrophoresis of soluble protein and restriction analysis of nuclear genes (reviewed in Sarich, 1985 and Bonhomme, 1987).
It was felt that an analysis of Spi-2 reactive centre sequence stability among species which had diverged very recently might reveal intermediates in the divergence process, and indicate possible evolutionary mechanisms for this phenomenon.

The genus *Mus* was a suitable choice for an analysis of this type, because of the existence of several murid species which diverged at intervals after mouse and rat divergence. A consensus molecular phylogeny of species for which liver RNA was available is shown in Figure 3.1. The actual divergence times for more distantly related *Mus* species are tentative, especially for the *Mus/Rattus* divergence which is thought to have occurred between 10 and 30 million years ago (Sarich, 1985; Bonhomme, 1987). The contrapsin gene was the most suitable Spi-2 sequence for examination, due to its high levels of expression in the liver, and the fact that it had previously been sequenced (Hill et al., 1984).

*Northern Blot Analysis of Contrapsin Evolution in Mus*

Experiments were carried out to examine the nature of the contrapsin reactive centre region in closely related species within the genus *Mus*. Initially this was carried out by oligonucleotide hybridisation to Northern blots of mouse total liver RNA. Oligonucleotides were used as probes since the variable reactive centre region comprises less than five percent of the total Spi-2 coding region, and cDNA probes are unable to distinguish between reactive centres within the conserved remainder of the gene. Short (< 20 nucleotide) oligonucleotides were used as these can detect single basepair mismatches in a complementary sequence under stringent conditions (Wallace et al., 1979). An 18mer antisense oligonucleotide (p54/RC-) was synthesised to complement the reactive centre region of contrapsin surrounding the P1 residue (Figure 3.2). This was used to hybridise to various inbred mouse and less closely related mouse RNAs (Figure 3.3B) and control rat RNA (not shown).
Figure 3.2 Sequence of the pLv54 contrapsin 3’ coding region showing the location of oligonucleotides used. Also shown are the predicted positions of intron-exon junctions in the rodent Spi-2 genes, based on the human α1-antichymotrypsin and α1-proteinase inhibitor genes. Numbering of sequence is as in Hill and Hastie (1987). Oligos are shown as asterisks, arrowed at their 3’ ends. Region 2 sequence is boxed and in lower case, except for the P1 residue. Region 1 is the entire coding sequence 5’ to Region 2, and Region 3 is defined as the remainder of the coding sequence 3’ to Region 2.
Figure 3.3 Northern blot of mouse liver RNAs. A: probed with oligo p54/1-; B: probed with oligo p54/RC-; C: probed with oligo p54/3'. Five μg of total liver RNA was loaded for each mouse strain. The position of ribosomal size markers is indicated.
The contrapsin signal was present in all the inbred mice examined, although at reduced levels in C57BL/6 mice. It was also present in wild *M. domesticus*, *M. spretus*, and surprisingly in *M. caroli*, a considerably more distant relative (6 to 8 million years diverged - Fig 3.1) of the laboratory mice.

The signal intensity for wild *M. musculus* (Danish - from G.Bulfield) was also reduced, to the same degree as that of C57BL/6. There was no detectable hybridisation of the contrapsin oligo to Sprague Dawley rat RNA (not shown). The blot was then probed with an antisense 16mer oligonucleotide (p54/1-) to a highly conserved region in the centre of the contrapsin sequence (Fig 3.3A); this hybridised to all the mouse RNAs more equally and indicated that the diminution of reactive centre signal from C57BL/6 and *M. musculus* might be due to a genuine sequence difference in the reactive centre of these mice. A third oligonucleotide, p54/3'-, hybridised strongly to all the mice, although less strongly to *Mus caroli*, indicating possible 3' divergence for this less related species (Figure 3.3C). Interestingly, two transcripts of similar abundance were observed in *M. caroli*; an 18S band (typical of other mice) and a 20S band, only seen faintly in some inbreds and in *Mus domesticus*, never in C57BL/6 or *Mus musculus*. The fact that *Mus spretus* (and *Mus caroli*) appeared to express a completely conserved reactive centre sequence was an indication that some degree of functional conservation had occurred in the contrapsin gene in the last 4 to 8 million years.

**Attempted PCR Analysis of Contrapsin Reactive Centre**

To confirm and extend this analysis, it was decided to attempt to obtain sequence data for the reactive centre region of the murine contrapsins. The initial efforts centred upon the use of the Polymerase Chain Reaction (PCR) technique which had been recently published (Saiki *et al*., 1985). This method of amplifying specific segments of polynucleotide sequence within a complex mixture employs two specific oligonucleotide
primers complementary to opposite strands of sequences flanking the desired segment. The PCR technique, and more recent associated strategies (Ehrlich, 1989) have since become of central importance to molecular biology.

It was decided to amplify the reactive centre region using a sense oligonucleotide (p54/5'-+) and an antisense oligonucleotide (p54/3'-) to span the reactive centre and carboxyl terminus (Fig 3.2). Successful PCR should yield a DNA fragment of 144bp which could then possibly be analysed by sequencing, fine restriction mapping, or an RNAse protection assay (Winter et al., 1985). The PCR was to have been carried out on total RNA samples from mouse liver, the first amplification step being carried out by reverse transcriptase to produce a cDNA template for subsequent amplification.

To test the system, initial experiments were carried out using the pLV54 plasmid which contained the original contrapsin cDNA clone. The plasmid was linearised by restriction with Eco RI, to allow efficient heat denaturation of the plasmid during the initial cycle. Two reactions containing 10ng and 1ng of plasmid in 50µl of 1X Klenow Buffer, and 2mM dNTPs were then performed, using Klenow enzyme, adding 5 units at every cycle. Initial PCR reactions were carried out for 20 cycles, which should have yielded clearly visible DNA fragments of 144bp by agarose gel electrophoresis. However neither reaction gave visible bands of the correct size, after ethidium bromide staining. The experiment was repeated using plasmid DNA cut with Msp I which cuts internally to the GC tails known to flank the contrapsin cDNA. It was thought that the GC tails might be inhibiting denaturation of the cDNA sequence, or enhancing rapid reanneallling. This experiment also gave negative results. PCR was abandoned at this stage in favor of other methods. It is worth noting that several improvements have been made to the technique since then, notably the use of heat stable DNA polymerase from *Thermus aquaticus* which does not have to be added at every cycle. PCR has since been frequently
used for the type of rapid RNA analysis hoped for above, and would certainly be attempted again if the project was still at this stage.

**Sequencing of Contrapsin Reactive Centres in Mus**

Since contrapsin is abundantly expressed in the rodent liver (0.5% of mRNA), the 3' oligonucleotide p54/3' - (Figure 3.2) was used to directly sequence total liver RNA using reverse transcriptase (RT) and dideoxynucleotides. Kaartinen *et al.*, (1984), reported the use of reverse transcriptase to sequence purified mRNA from hybridomas producing monoclonal antibodies of interest. The technique has since been applied to determine the evolution of the lysozyme signal peptide sequence in various fowl species (Weisman *et al.*, 1985). In this project initial attempts followed DNA sequencing procedures, substituting RT for Klenow enzyme (and the appropriate buffer) and incubating at 40°C during the termination reaction. However, Klenow enzyme and RT have dramatically different affinities for dideoxy, compared to deoxy nucleotides. Klenow enzyme requires ratios of about 25:1 dideoxy:deoxy for efficient (mis)incorporation of the termination analogue, whereas RT is less nucleotide specific (ratio of 1:1 dideoxy:deoxy). This became apparent when initial gels, run with reverse transcription reactions using Klenow dideoxy:deoxy mixes, only gave one to two bands in each lane at the bottom of the gel. This was readily corrected, and useful sequence data was obtained after a few further modifications.

Initially, three total liver RNAs (DBA, *M. domesticus*, *M. spretus*), and one polyA⁺ liver RNA (BALB/c) were sequenced. An early gel is shown in Figure 3.4. From this data, sequence for the total RNAs could be established for more than 90 basepairs, up to and including the P1 residue (Figure 3.5).
Figure 3.4 Example of early RNA sequence autoradiograph. A,B,C,D for each sample refers to the A°, C°, G°, and T° dideoxynucleotide reactions. The P1 residue sequence (AAG) can be seen at the very top of the photograph.
Figure 3.5 Sequence of contrapsin reactive centres and 3' coding region for various mice. The nucleotide and predicted protein sequences are shown. Dashes below the original pLv54 Swiss sequence indicate sequence identity. Region 2 (the reactive centre) sequence is boxed and shown in lower case, except for the P1 residue (CGT).

<table>
<thead>
<tr>
<th>swiss</th>
<th>GTG</th>
<th>CTT</th>
<th>CTG</th>
<th>GGG</th>
<th>AGA</th>
<th>GCA</th>
<th>TGT</th>
<th>GTG</th>
<th>GCT</th>
<th>GGC</th>
<th>ACA</th>
<th>GCC</th>
<th>ACA</th>
<th>AAG</th>
<th>GCA</th>
<th>GCT</th>
<th>GCC</th>
<th>ACA</th>
<th>GGG</th>
</tr>
</thead>
<tbody>
<tr>
<td>C57BL/6</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DNA</th>
<th>GTG</th>
<th>CTT</th>
<th>CTG</th>
<th>GGG</th>
<th>AGA</th>
<th>GCA</th>
<th>TGT</th>
<th>GTG</th>
<th>GCT</th>
<th>GGC</th>
<th>ACA</th>
<th>GCC</th>
<th>ACA</th>
<th>AAG</th>
<th>GCA</th>
<th>GCT</th>
<th>GCC</th>
<th>ACA</th>
<th>GGG</th>
</tr>
</thead>
<tbody>
<tr>
<td>C57BL/6</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>swiss</th>
<th>GTG</th>
<th>CTT</th>
<th>CTG</th>
<th>GGG</th>
<th>AGA</th>
<th>GCA</th>
<th>TGT</th>
<th>GTG</th>
<th>GCT</th>
<th>GGC</th>
<th>ACA</th>
<th>GCC</th>
<th>ACA</th>
<th>AAG</th>
<th>GCA</th>
<th>GCT</th>
<th>GCC</th>
<th>ACA</th>
<th>GGG</th>
</tr>
</thead>
<tbody>
<tr>
<td>C57BL/6</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

| swiss | AGG  | CCA  | TTC  | CTG  | TTT  | GGG  | CTT  | ATC  | TAT  | CAC  | AGA  | AGT  | GCT  | CAG  | AGT  | ATC  | CTC  | TTT  |
|-------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| C57BL/6 | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  |

| DNA    | AGG  | CCA  | TTC  | CTG  | TTT  | GGG  | CTT  | ATC  | TAT  | CAC  | AGA  | AGT  | GCT  | CAG  | AGT  | ATC  | CTC  | TTT  |
|--------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| C57BL/6 | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  | ---  |
This confirmed the indications from the Northern data, described above, that the sequence surrounding the P1 residue appeared to have evolved at typical rates since the divergence of *Mus domesticus* and *Mus spretus* (2 to 4 mya). No differences were found between DBA/2 and wild *Mus domesticus* (Greek, from G. Bulfield).

Comparison of the pLv54 and *Mus spretus* contrapsin sequences revealed only one base change over the original region sequenced. This was an A to G transversion, causing a histidine to arginine substitution at the boundary of Region 2 (reactive centre) and Region 3 (as defined in Hill and Hastie (1987)). This is a conservative amino acid substitution, and would be unlikely to cause an alteration in contrapsin inhibitory specificity.

The BALB/c polyA* sample did not sequence as well as the total RNA samples. The reason for this was never clearly established, but was unexpected given a theoretical 20-fold enrichment of the sequencing template. Reaction conditions may have become limiting (e.g., nucleotide concentration) with higher concentrations of template, or degradation of template may have occurred during mRNA isolation. Experiments to optimise sequencing of purified polyA* never established a cause, or solved the problem. In any case, the partial BALB/c sequence which was obtained was identical to that of the DBA and *Mus domesticus* genes.

**Evolution of the Contrapsin Gene in More Divergent Mice**

The fact that *Mus spretus* contrapsin is so similar in sequence to the original (*Mus domesticus*) contrapsin, pLv54, indicates that the accelerated evolution (Hill and Hastie, 1987) appears to have slowed or halted, before the divergence of *Mus domesticus* and *Mus spretus*. These species are considered to have diverged two to four million years ago (Figure 3.1), about one fifth of the *M. domesticus*/rat divergence time. This raised a question: at what stage of murine radiation did fixation of the extant contrapsin actually occur?
Several murine species more distantly related to *Mus domesticus* (Figure 3.1) exist as laboratory stocks, and efforts were made to obtain some of these. *Mus caroli* was kindly donated by Dr. J. Ansell (Dept of Zoology, Edinburgh University), and total liver RNA was prepared from the liver of one male and one female individual. Sequencing of the RNA proved more difficult than for the inbred mice, possibly due to a mismatch between the primer, p54/3'-, and the *M. caroli* contrapsin (Figure 3.3C). Eventually, successful sequencing of the *M. caroli* RNA was achieved (Figure 3.5), which demonstrated the presence of a typical mouse contrapsin with twelve of thirteen amino acid residues in the reactive centre region conserved, a histidine to serine substitution at the Region 2/Region 3 boundary being the only change. This represented an independent mutation at the same residue which was altered between *Mus spretus* and *Mus domesticus*, indicating a lack of conservation at this position. There were several nucleotide substitutions in Region 3 of the *Mus caroli* contrapsin, reflecting the considerable evolutionary distance between *M. caroli* and *Mus domesticus* (consensus divergence time six to eight m.y.a.). Surprisingly, two thirds of these nucleotide substitutions are synonymous and would cause amino acid replacements. This could be due to relaxed selection at the carboxyl terminus of the protein due to lack of conformational constraint, but might be an indication that contrapsin is not under overall stringent selection in one or both of the species. Also, in the original Spi-2 study (Hill and Hastie, 1987), Region 3 was evolving appreciably more rapidly than Region 1, and it may be that this region of the inhibitor is also involved in proteinase interaction. To address the question of relaxed carboxyl terminal constraints, other liver-encoded RNAs for serum proteins were sequenced, as described below.

At this stage, C57BL/6 and *Mus musculus* contrapsins, which had also been refractory to sequencing, were also successfully analysed. The reactive centre probe, p54/RC-,
had never hybridised as efficiently to these mice as to the other inbreds, *Mus domesticus*, or *Mus spretus*. The reason for this became apparent when the sequences were compared. An A to G mutation causing an isoleucine to valine substitution was found in *Mus musculus* at the P2' position within the reactive centre. This might not be expected to have major functional effects, but would certainly destabilise the reactive centre oligo under the hybridisation conditions used (Figure 3.3B). The C57BL/6 mouse shared an identical contrapsin sequence to *Mus musculus*, which included a further two amino acid substitutions as compared to the pLv54 standard sequence. It is likely that a *Mus musculus* individual was the donor for the C57BL/6 contrapsin gene sequence during development of the inbred strain. Similarly the DBA/2 and feral *Mus domesticus* sequences were identical when compared, indicating that the DBA/2 strain may have derived its contrapsin sequence from a *Mus domesticus* individual prior to isolation of the inbred strain. To illustrate this further, an oligonucleotide, p54/BL6-, (Figure 3.2) complementary to the Region2/Region3 boundary sequence for the C57BL/6 contrapsin was used to probe a Northern blot of mouse liver RNAs (Figure 3.6). This demonstrated that the C57BL/6, *M. musculus*, and C3H contrapsins are identical in this region, and distinct from the pLv54 (Swiss)/DBA/*M. domesticus* contrapsin group. Presence of the *M. domesticus* contrapsin is associated with the constitutive expression of a 20S form of the contrapsin sequence, which is not observed in *M. musculus* or in certain strains of inbred mice (Hill *et al.*, 1985). The *Mus musculus* specific oligonucleotide p54/BL6- only hybridised to strains lacking a constitutive 20S transcript. Thus, specific oligonucleotides can be used to assign the contrapsin gene from inbred mouse strains to either a *M. musculus* or *M. domesticus* donor sequence.
Figure 3.6 Northern blots of rodent liver RNAs with oligo p54/BL6-. 18S specific hybridisation can be seen for *Mus musculus*, and the inbred strains C57BL/6 and C3H, indicating that these mice share the cysteine variant contrapsin.
Divergence of Contrapsin in Ancient Mice

The existence of a functionally similar contrapsin in both *Mus domesticus* and *Mus caroli* was unexpected. Stabilisation of the contrapsin reactive centre region appears to have occurred before the divergence of these mice, six to eight million years ago. This reduces the period during which contrapsin diverged from the rat *Spi-2.1* sequence (see figure 3.1), and consequently further increases the rate of reactive centre divergence, as calculated by Hill and Hastie (1987).

To further determine the point at which contrapsin became functionally stabilised in the ancestors of modern *Mus*, studies were carried out on murine species which diverged from the rat lineage before *Mus caroli*. The wood mouse, *Apodemus sylvaticus*, is a major non-commensal murine species in Britain. It is thought to have separated from the *Mus* lineage after mouse/rat divergence (Figure 3.1) although this is not thoroughly established. Specimens of *Apodemus sylvaticus* were trapped at the Bush Estate for analysis. Total liver RNAs from mouse, *A. sylvaticus* and rat were probed with the rat *Spi-2.1* cDNA pSPFL2.1 to determine the relative abundance and nature of *Spi-2* sequences in *Apodemus* (Figure 3.7). *Apodemus* proved to express both 18S and 20S *Spi-2* transcripts, but at far lower levels than either rat or mouse. This could be due to increased evolutionary distance between *Apodemus* and the other two rodents. However, analysis of other mRNA sequences, (see below), did not support this possibility. Alternatively, *Apodemus* may lack the contrapsin gene, at least as defined by level of expression in the liver. Several attempts to sequence *Apodemus* liver RNA with conserved *Spi-2* oligonucleotide primers failed. This could have been due to divergence in the rapidly evolving carboxyl terminus region, or to low abundance of a contrapsin transcript, but later cDNA cloning analysis confirmed the lack of a detectable expressed orthologue of the contrapsin sequence (Chapter 5).
Figure 3.7 Northern blot of rodent liver RNAs probed with rat Spi-2.1. The blot is heavily overexposed to reveal the presence of two *Apodemus Spi-2* transcripts (18S and 20S) at low abundance. Ten μg of total liver RNA was loaded for each species. 1. DBA, 2. C57BL/6, 3. and 4. Induced C57BL/6, 5. *Apodemus sylvaticus*, 6. Rat, 7. Induced rat. Sizes of ribosomal markers are indicated.
Since *Apodemus* appeared to lack the contrapsin characteristic of the modern mice, species more closely related to *Mus domesticus* were examined next. Liver RNA from *Mus pahari*, *Mus saxicola*, and *Mus minutoides* were obtained from F. Berger (University of S.Carolina). These mice diverged from *Mus domesticus* some 10 million years ago, although their branching order is not clear (Figure 3.1). The RNAs were electrophoresed and Northern blotted with the contrapsin cDNA subclone pBXC1.4 (Figure 3.8). This represents the 3' third of the coding sequence and is a more specific contrapsin probe than the rat cDNA pSPFL2.1 used in Figure 3.7. The probe hybridises strongly to mice as divergent as *Mus domesticus*, *Mus cervicolor* and *Mus caroli* all already known to have a conserved contrapsin sequence. Less intense hybridisation is observed to *Mus pahari* and *Mus saxicola* RNA. There is no obvious hybridisation to *Mus minutoides*, *Apodemus sylvaticus*, or rat liver RNA. This may indicate the presence of a contrapsin sequence in *M. pahari* and *M. saxicola* which arose before these species diverged from the more recently separated mice.

However, attempts to hybridise contrapsin-specific oligonucleotides to Northerns like the one shown in Figure 3.8 never revealed any hybridisation to the *M. pahari* or *M. saxicola* liver RNA (not shown), and attempts to sequence liver RNA directly were also unsuccessful. The contrapsin equivalent in these mice may have diverged from the pLv54 sequence in the sequences complementary to the oligonucleotide, or the hybridisation observed in figure 3.8 may be due to another Spi-2 sequence. This question will be discussed later (Chapter 7).

**RNA Sequencing of Albumin in Mus**

The increased rate of divergence observed in the carboxy terminus region of the otherwise conserved contrapsin gene raised the question of whether this was typical of serum proteins in relatively recently diverged species. Excluding the Spi-2 genes, the
most rapidly evolving coding sequences are those for the interferons and relaxins (Li et al, 1985). This may be due to the fact that specific regions of these proteins are required for their physiological function, rather than the entire three-dimensional structure of the molecule. Conversely, highly conserved structural proteins such as the histones or actins are coupled closely to other intracellular components in complex assemblies, and are presumably more evolutionarily restricted. Functionally monomeric serum proteins involved in transport or peptide recognition roles, may generally possess a high degree of sequence flexibility and display relatively rapid sequence divergence. This has been demonstrated for the albumin and alpha-fetoprotein genes (Minghetti et al, 1985), where the latter gene is diverging almost at the rate of a pseudogene.

The RNA sequencing technique was applied to the albumin transcript as this is one of the most abundantly expressed in the mammalian liver. An 18-mer antisense oligonucleotide was synthesised, complementary to the carboxy terminal coding sequence of the mouse gene (Figure 3.9). The sequence selected was highly conserved, though not identical, in the rat (Sargent et al, 1981), and permitted sequencing of both species as well as intermediate murine types. A Northern blot of the oligonucleotide against an array of murine liver RNAs is shown in Figure 3.11. The major albumin transcript is clearly visible in mice as divergent as Mus saxicola, and faint transcripts can be seen in the Apodemus sylvaticus sample. The rat albumin message was not detected under these stringent conditions, but was clearly visible when a cDNA probe was used. The data obtained from the RNA sequencing of the rodent albumins are shown in Figure 3.9.
Figure 3.9 Albumin 3’ coding sequences from RNA sequencing. The nucleotide sequence only is shown. Dashes represent identity with the DBA sequence. The antisense 18mer oligonucleotide starts from the albumin translation termination codon (TAA). X represents an undetermined nucleotide.
Albumin shows a greater degree of carboxyl terminus constraint than contrapsin for both synonymous and asynonymous base substitutions. Between *Mus caroli* albumin and inbred mouse albumin, only one synonymous, and one non-synonymous substitutions occur in fifty codons. The corresponding values for contrapsin are two synonymous and six non-synonymous substitutions in forty codons. Since albumin is considered to represent a comparatively rapidly diverging gene (Li *et al.*, 1985; Minghetti *et al.*, 1985), the rate of contrapsin Region 3 divergence is remarkable, especially in view of the reactive centre conservation.

*RNA Sequencing of α1-Proteinase Inhibitor in Mus*

RNA sequencing was also carried out on the α1-Proteinase Inhibitor (α1-antitrypsin, α1-PI) transcript as this is also highly expressed in the mammalian liver (> 0.5% mRNA). An antisense 18mer oligonucleotide was synthesised complementary to a conserved region close to the 3’ end of the coding region (Hill *et al.*, 1984). This was used to probe the rodent Northern blot shown in Figure 3.11. The oligonucleotide gave strong hybridisation in mice as evolutionarily distant as *M.domesticus* and *Mus minutoides*, and a transcript can also be seen in *Apodemus* and rat liver RNA. Sequencing with the oligonucleotide provided data for several rodents, and even human HepG2 hepatoma RNA was partially sequenced (Figure 3.10). The α1-PI sequence appears to be evolving at a similar rate to contrapsin in the murine species examined, and more rapidly than albumin. It appears that rapid sequence divergence at the carboxyl terminus may be a general feature of the serpin family in rodents.

Unexpectedly, the α1-PI RNA sequencing analysis revealed that more than one α1-PI related sequence is expressed at high levels in mouse liver. Extra bands were visible in the region of the P1 residue sequence on the autoradiograph, indicating that a mixed population of template was present.
Figure 3.10 α1-Proteinase Inhibitor sequences obtained by RNA sequencing. Nucleotide sequence only is shown. Dashes represent identity with the published BALB/c α1-PI sequence (Krauter et al., 1985). The P1 tyrosine variant, pL1796 (Hill et al., 1984), is shown below. mix represents the P1 position in DBA and *M. musculus* where both forms are present. X represents an uncharacterised nucleotide. HepG2 is a human hepatoma cell line. The P1 codon is the first in the sequence (ATG in BALB/c).

<table>
<thead>
<tr>
<th></th>
<th>ATG</th>
<th>TCT</th>
<th>ATG</th>
<th>CCC</th>
<th>CCT</th>
<th>ATC</th>
<th>CTG</th>
<th>CGC</th>
<th>TTC</th>
<th>GAC</th>
<th>CAC</th>
<th>CCT</th>
<th>TTC</th>
<th>CTT</th>
<th>TTC</th>
<th>ATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>BALB/c</td>
<td>ATG</td>
<td>TCT</td>
<td>ATG</td>
<td>CCC</td>
<td>CCT</td>
<td>ATC</td>
<td>CTG</td>
<td>CGC</td>
<td>TTC</td>
<td>GAC</td>
<td>CAC</td>
<td>CCT</td>
<td>TTC</td>
<td>CTT</td>
<td>TTC</td>
<td>ATA</td>
</tr>
<tr>
<td>pL1796</td>
<td>TAT</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>DBA</td>
<td>mix</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Danish</td>
<td>mix</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Caroli</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Rat</td>
<td>---</td>
<td>C-</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>HepG2</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>C-</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>ATG</th>
<th>TTT</th>
<th>GAA</th>
<th>GAA</th>
<th>CAC</th>
<th>ACT</th>
<th>CAG</th>
<th>AGC</th>
<th>CCC</th>
<th>ATC</th>
<th>TTC</th>
<th>GAGGAAAGTGGTAGATCCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>BALB/c</td>
<td>ATA</td>
<td>TTT</td>
<td>GAA</td>
<td>GAA</td>
<td>CAC</td>
<td>ACT</td>
<td>CAG</td>
<td>AGC</td>
<td>CCC</td>
<td>ATC</td>
<td>TTC</td>
<td>GAGGAAAGTGGTAGATCCC</td>
</tr>
<tr>
<td>pL1796</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td><em><strong>Primer site</strong></em></td>
</tr>
<tr>
<td>DBA</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>C</td>
</tr>
<tr>
<td>Danish</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Caroli</td>
<td>G-</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Apodemus</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>TAC</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Rat</td>
<td>---</td>
<td>G-</td>
<td>---</td>
<td>TAC</td>
<td>---</td>
<td>A-</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>X</td>
</tr>
<tr>
<td>HepG2</td>
<td>---</td>
<td>G-</td>
<td>---</td>
<td>TAC</td>
<td>---</td>
<td>A-</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>A-</td>
<td>---</td>
</tr>
</tbody>
</table>

Sizes of transcripts are 17S (α1-PI) and 20S (more abundant albumin transcript).
The existence of multiple α1-PI forms was supported by the previous isolation of two very similar but distinct α1-PI cDNAs from mouse liver libraries (Hill et al, 1984; Krauter et al, 1985). The main difference between the two sequences was the complete substitution of nucleotides in the P1 residue codon. In the Hill sequence (pLv1796) the codon is TAT (tyrosine), giving a predicted anti-chymotrypsin specificity. In the Krauter sequence (pliv3b) the codon is ATG (methionine), giving a predicted anti-elastase activity as in the mammalian sequence. RNA sequencing indicated that both messages are expressed in the mouse liver at similar levels. The tyrosine P1 variant was not detected in Apodemus, rat, or human liver RNA sequences. This gene may represent a very recent gene duplication event in view of its high overall similarity to the methionine P1 transcript, and its apparent absence from certain rodent species. If this is the case, the divergence at the critical P1 residue is remarkable and provides additional support for the concept of accelerated evolution to generate novel functional specificities in this gene family.

The Rat 4.12B Gene - A Recent Spi-2.1 Duplication Event

During the RNA sequencing experiments, two papers were published describing a third Spi-2 gene in the rat (Tecce et al., 1986; Yoon et al., 1987; Le Cam et al., 1987). This sequence, first described as 4.12B, was distinct from the Spi-2.1 and Spi-2.2 sequences described by Hill and Hastie (1987). It is far more closely related to the Spi-2.1 sequence and appears to represent a duplication event after, rather than before, rodent divergence. It resembles a "typical" Spi-2.1 sequence in several ways. Comparison of transcript size, abundance and low inducibility by inflammatory stimuli showed that it was controlled in a very similar fashion to the rat Spi-2.1 gene (Figure 3.12). However the reactive centre regions of the two genes have diverged extensively, with 50% nucleotide conservation, but only three amino acid residues in common in a
fifteen residue polypeptide (Figure 3.13). Most of the substitutions which have occurred are non-synonymous, a feature shared by most of the substitutions which have occurred in Region 3. The carboxyl terminus divergence is reminiscent of other Spi-2 genes, but the reactive centre divergence is clearly demonstrated in this comparison. It seems possible that the 4.12B/Spi-2.1 gene pair might represent the type of rapid evolution following intraspecies gene duplication which was not found for the interspecies analysis of the contrapsin gene alone. This might be due to the fact that contrapsin, and presumably one of the 4.12B/Spi-2.1 pair, have become selectively constrained for a defined specificity unlike the other rat gene.
Figure 3.12 Northern blot of rat liver poly(A) RNA with specific $ Spi-2.1 $ and 4.12B reactive centre probes. Two $ \mu g $ of poly (A) RNA were loaded for each sample. Dex = Dexamethsone-treated, PB = Phenobarbital-treated, 3MC = 3-Methylcholanthrene-treated animals. The 18- mer antisense oligonucleotides were: $ Spi-2.1 $, AGGTTGTCTACGTATGAC; 4.12B, CGGTAACCTTTTAGGC. Blots were washed at high stringency (4X SSC, 50$^\circ$). The two probes show expression of the two genes are virtually identical in abundance, transcript size, and inducibility.
Figure 3.13 Comparison of 4.12B and Spi-2.1 sequences in the reactive centre region.

The nucleotide sequences for exon 5 of both cDNAs are shown. Nucleotide differences are indicated by asterisks. The complete amino acid sequence for the Spi-2.1 cDNA is shown above the DNA sequences, and that of 4.12B, where they differ, is shown below. Dashes represent deletions in the Spi-2.1 sequence. 2.1 = Spi-2.1, 2.3 = 4.12B. Region 2 (the reactive centre) is boxed.
Conclusions

This chapter examines the nature of contrapsin reactive centre stability in close relatives of *Mus domesticus* to determine whether the accelerated evolution found for *Spi-2* sequences by Hill and Hastie is a general phenomenon for these genes in rodents. The contrapsin reactive centre was found to be highly stable in sequence, over a six to eight million year period (since the divergence of *M. domesticus* and *M. caroli*). This may indicate that fixation of the contrapsin specificity for a particular inhibitory activity occurred in the rodent ancestor of these mice. This would reduce further the period during which the contrapsin gene and its rat orthologue(s) diverged, and hence increase the rate of evolution at this time. Sequence for the contrapsin gene from more distant relatives of *M. domesticus* was not obtained, and there is some evidence that these mice may not express a contrapsin gene, at least at the levels of the more closely related mice. The carboxyl termini of the albumin and α1-PI (*Spi-1*) genes were sequenced for these distantly related mice, and provide evidence that the carboxyl termini of both serpins are evolving more rapidly than that of albumin. The existence of a rat gene which has recently diverged from the rat *Spy-2.1* sequence appears to represent an intermediate in the divergence process (compared to mouse contrapsin and rat *Spy-2.1*). A very high proportion of asynonymous mutations in Region 2 and Region 3 was found for this gene pair. The data from this chapter suggests that the phenomenon of accelerated evolution in the *Spy-2* gene family occurs during the period following gene duplication, and stabilization occurs after a selectively useful specificity has been obtained.
CHAPTER FOUR

The Genomic Structure of the Spi-2 Locus

Introduction

Analysis of the Spi-2 locus in the inbred mouse was carried out for a number of reasons. Hill et al (1985) had shown previously that the Spi-2 locus contained several closely linked genes. The existence of several genes suggested the possibility that novel reactive centre specificities had evolved following gene duplication. Isolation and sequencing of genomic Spi-2 clones was required to indicate whether the genes contained different reactive centres to those characterised at the cDNA level, and whether any with novel specificities were expressed. The proportion of pseudogenes in the locus, and the true complexity of the locus, were of interest in terms of establishing the evolutionary mechanism of expansion of the locus. Finally, a detailed genomic structure for sequences within the locus might reveal possible physical mechanisms responsible for the accelerated evolution phenomenon.

Complexity of the Spi-2 Locus in Rodents

To establish the relative complexity and variability of the Spi-2 locus within the genus Mus, a Southern blot of rodent DNAs was carried out using the rat Spi-2.1 cDNA as a probe. This clone contains most of the coding sequence, including the relatively conserved 5' regions absent from the pLv54 contrapsin cDNA. The result is shown in Figure 4.1. It is apparent that the Spi-2 locus is more complex in Mus domesticus and its near relatives than in various rat strains. Since the probe employed was a rat sequence, this is unlikely to be due to sequence divergence at the rat Spi-2 locus.
Figure 4.1 Southern blot of rodent DNAs probed with the rat Spi-2.1 cDNA. Ten µg of each rodent genomic DNA was digested with Bam HI, electrophoresed, and transferred to a membrane. The Mus caroli DNA was used at an incorrect concentration, giving a poor signal. The smearing and the intense band in the C57BL/6 lane are due to plasmid contamination. The increased complexity of the mouse Spi-2 locus compared to the rat is apparent.
The complexity of the *Apodemus sylvaticus* hybridisation pattern approaches that of the mice, and would support the idea of a considerable amplification of this gene family in ancestral mice, after divergence from the ancestral rat line. Alternatively the *Apodemus* expansion may have occurred independently, or the rat locus may have contracted since rodent radiation. Since the cDNA probe may span several exons it cannot be used to quantify the exact number of *Spi-2* sequences in each species, but about twenty separate fragments are visible in the *Mus musculus* DNA compared to ten in the Wistar rat DNA for example (Figure 4.1).

**Analysis of Spi-2 DBA Genomic Clones**

Previously, eight mouse genomic clones had been isolated from a DBA spleen DNA bacteriophage library (C.Sime and R.Hill). The clones, in the phage vectorλEMBL3 (Frischauf *et al.*, 1983), had been selected by hybridisation with pLv54, the original contrapsin cDNA. The clones, which had not been further characterised at this stage, were numbered p54λ-1 to p54λ-9; there was no p54λ-6.

The clones were then amplified using the large scale plate lysate technique with the bacterial strain LE392. About 10μg of phage DNA was obtained for each clone. Each DNA was cut with the restriction enzyme Eco RI and analysed by electrophoresis. This indicated that p54λ-8 and p54λ-9 were probably clonal, but that all the other clones appeared to be independently derived (not shown). The DNA was then transferred to a Hybond-N nylon filter which was probed with the pLv54 cDNA, then with a specific oligonucleotide to a conserved sequence 5' to the reactive centre. The pLv54 probe hybridised to all the clones as expected, but only to a 1kb fragment in the case of p54λ-3 and only faintly in the case of p54λ-5 (Figure 4.2).
Figure 4.2 Southern blot of Eco RI digested p54\(\lambda\) clones probed with the contrapsin pLv54 clone. Lane 1: \(\lambda\)-Hind III size marker, 2: p54\(\lambda\)-1, 3: p54\(\lambda\)-2, 4: p54\(\lambda\)-3, 5: p54\(\lambda\)-4, 6: p54\(\lambda\)-5, 7: p54\(\lambda\)-7, 8: p54\(\lambda\)-8, 9: p54\(\lambda\)-9, 10: no DNA. The approximate positions of the \(\lambda\)-Hind III size markers are shown.
Figure 4.3 Southern blot of Pst I digested p54 clone probed with the oligonucleotide p54/25-. Lane 1: λ-Hind III size marker, 2: p54-1, 3: p54-2, 4: p54-3, 5: p54-4, 6: p54-5, 7: p54-7, 8: p54-8, 9: p54-9, 10: no DNA. The approximate positions of the λ-Hind III size markers are shown. The DNA in lane 3 (p54-2) ran abberantly, and was underloaded.
The reactive centre oligo, p54/RC5'+, hybridised to clones 8 and 9, indicating that these clones contained reactive centre sequences (not shown). At this stage, p54λ-8 was selected for further analysis, since it also hybridised to oligo p54/1-, a conserved sequence probe for a region 500bp 5' to the reactive centre (Figure 3.2). Direct dideoxy sequencing of the phage DNA was attempted using sense and anti-sense oligo p54/1 and also using oligo p54/RC5'+. The sequence data was not of very high quality, but indicated that the archetypal Spi-1 and Spi-2 genomic structure (as in α1- proteinase inhibitor and α1-antichymotrypsin - see Figure 3.2) had been conserved, at least in the sequenced regions. Oligo p54/1- hybridises to an exonic sequence just 5' to intron B in antichymotrypsin, and the sequence data showed clear homology to inhibitor sequences using the antisense, but not the sense, oligo p54/1+ primer. Similarly, the oligo p54/RC5'+ sequence data demonstrated that the reactive centre was not directly flanked by introns, but shared the antichymotrypsin structure. This also eliminated an obvious exon shuffling event as the mechanism for reactive centre divergence, for this gene at least.

### Subcloning and Sequencing of DBA Spi-2 Reactive Centres

The eight DBA genomic clones were digested with PstI and run on an 0.8% agarose gel. After Southern blotting, the filter was probed with oligo p54/25-, an 18-mer which hybridises to sequence encoding the highly conserved GTEAAA motif immediately upstream of the reactive centre in Spi-2 genes (Figure 3.2). The oligo hybridised to clones 1, 2, 4, 7, 8 and 9, indicating that all these inserts contained a reactive centre region (Fig 4.3). The PstI-cut hybridising fragments for clones 1 and 2 were about 1.2kb, for clone 4 greater than 7kb, and for clones 7, 8, and 9 about 0.9kb. It was noticeable that the oligo hybridised less efficiently to clone 7 than clones 8 and 9, indicating a sequence difference despite their similar fragment size. Clones p54λ-1, 2, 7 and
8 were then Pst I digested and shotgun cloned into the PstI site of Bluescribe. Recombinant clones containing the reactive centre fragment were isolated, by probing with oligo p54/25-, for each positive p54λ clone. The phage p54λ-4, which contained a >7kb oligo p54/25- hybridising PstI band, was further restriction mapped to determine a more clonable fragment. Eventually, a 3kb Eco RI/PstI fragment was selected and subcloned into Bluescribe.

The plasmid subclones were then all sequenced directly, using a variety of primers around the reactive centre region (Figure 3.2 shows the location and sequences of these). Oligo p54/XV+, a sense 18-mer, hybridised to the first six codons of exon 5 and was useful for most of the clones (but not p54λ-4-13). Oligo p54/51-, an anti-sense 17 mer, hybridised to the FNRPFL motif, just 3' to the reactive centre region. It was also useful for all DBA clones except p54λ-4-13. Oligo p54/25- (antisense, 18-mer) and its complementary oligo, p54/129+ (sense, 18-mer) were also used at this stage. The sequences obtained for exon 5 and the 5' flanking intron of the DBA Spi-2 genomic sequences are shown in Figure 4.4a, and the predicted exon 5 protein sequence in Figure 4.4b

The sequences obtained for p54λ-1-108 and p54λ-2-100 were identical and it appears that these clones contain overlapping, but not identical, inserts. The putative P1 residue for this inhibitor sequence is leucine, which immediately raised the question of whether it represented the functional murine equivalent of human α1-antichymotrypsin, which also contains a leucine P1 residue (see Chapter Five). The sequences of p54λ-7-45 and p54λ-8-1 were clearly closely related to one another, but showed unique cysteine-cysteine and histidine- cysteine residues, respectively, at the putative P1-P1' position. The partial sequence for p54λ-4-13 (due to lack of suitable sequencing primers)
Figure 4.4a Nucleotide sequences for exon 5 of the DBA Spi-2 genomic clones. Sequences obtained from the p54λ subclones, including all of exon 5 where possible, and a portion of the 5' flanking intron (shaded grey). Region 2 (the reactive centre) is boxed. The coding sequences for contrapsin are included for reference. Dashes represent identity with the top sequence, asterisks represent deletions, and full stops represent undetermined sequence. The putative P1 residue is marked.
Figures 4.4b and 4.4c 4.4b: Predicted amino acid sequence for exon 5, again including contrapsin. Conserved residues are boxed. This illustrates the rapid nature of divergence of the reactive centre, as the boxes almost exactly flank region 2. Asterisks represent deletions, and full stops represent undetermined sequence. The putative P1 residue is marked. 4.4c: Tables showing increased divergence of exonic sequences compared to introns for the data from Figure 4.4a. The figures in the tables represent the percentage nucleotide identity for each DBA gene pair. Table A: intron, Table B: exon 5, Table C: exon 5 without reactive centre.

![Figure 4.4b](image)

<table>
<thead>
<tr>
<th>Table A - intron 4</th>
<th>Table B - exon 5</th>
<th>Table C - exon 5 minus RC</th>
</tr>
</thead>
<tbody>
<tr>
<td>p7.45</td>
<td>87%</td>
<td></td>
</tr>
<tr>
<td>p1.108</td>
<td>81%</td>
<td>78%</td>
</tr>
<tr>
<td>p4.13</td>
<td>73%</td>
<td>72%</td>
</tr>
</tbody>
</table>
indicated that this sequence, at least, might be pseudogenic since many highly conserved residues in the reactive centre-flanking domains had been replaced in this gene. However, the putative P1-P1' residues are methionine-serine, identical to the highly active serpin α1-proteinase inhibitor.

Intergenic nucleotide sequence comparisons for the four DBA clones all showed greater conservation between introns than between exon sequences (Figure 4.4c). For the three genes which resemble functional Spi-2 sequences this may be due to accelerated divergence in the reactive centre, as all three are relatively conserved in Regions 1 and 3. Additionally, gene conversion events involved in homogenising Region 1 may also have homogenised upstream intron sequences and retarded the apparent rate of intron divergence. However, for comparisons involving p54λ-4-13, the intron showed considerably more conservation in terms of percent identity than the exon. Since this sequence appears to be pseudogenic in terms of loss of conserved residues in Region 3, intron and exon divergence values should be similar due to lack of selection on the coding sequences. It is possible that this sequence encodes a functional protein, or alternatively that gene conversion has homogenised intronic but not exonic sequences.

None of the DBA λ-EMBL3 clones analysed contained more than one detectable reactive centre region per insert, and none encoded contrapsin. It was clear that the genes isolated from the phage library were only a fraction of the ten to fifteen predicted by Southern analysis of mouse DNA. Unfortunately, the DBA phage library underwent a dramatic (total) loss of titre at this time, and further screening was not possible.
Genomic Analysis of DBA Spi-2 Sequences

To attempt to determine the distribution of the DBA Spi-2 sequences in various rodents, duplicate Southern blots of BamHI digested DNA from *Mus domesticus*, *Mus caroli*, *Apodemus sylvaticus*, and Wistar rat were analysed. Probes used were the reactive centre specific subclones from the DBA genomic clones p54λ-1-108, p54λ-7-45, p54λ-8-1, and the contrapsin cDNA clone pLv54. The results are shown in Figure 4.5.

As expected from their sequence similarity, the 7-45 (cys-cys) and 8-1 (his-cys) subclones give a very similar hybridisation pattern in the mouse DNA. These genes would appear to have diverged very recently, and to show little difference in their hybridisation specificity. The presence of several bands of similar intensity with both these genomic probes is an indication that a number of closely related genes exist in the mice and *Apodemus*. The 1-108 (leu-ser) subclone gives a related pattern, but with a difference in band intensities, indicating that it is less related to the 7-45 and 8-1 gene group. A single heavily hybridising band is apparent at about 4.5 kb, and presumably represents the corresponding genomic BamHI fragment for this gene. The absence of other heavy bands may indicate that this gene has not undergone recent duplication, or sequence conversion events. The contrapsin cDNA yields a slightly different pattern, recognisably related to that of the DBA genomic probes. Three intense bands can be seen, but it is not clear whether these represent separate genes, or are due to the fact that the probe probably spans several exons. The genomic probes did not hybridise strongly to the Wistar rat DNA, probably due to the fact that they consist largely of intron and 3' untranslated sequences. Even the contrapsin cDNA probe is a poor rat probe, due to its lack of conserved *Spi-2* sequences from the 5'coding region. This is in contrast to the full length rat cDNA pSPFL2.1, which is a good probe for both mouse and rat (Figure 4.1).
Figure 4.5 Southern blots of rodent DNAs probed with mouse Spi-2 sequences. Each lane contained 10 μg of Bam HI digested genomic DNA. DNAs: d = *Mus domesticus* (Greek), c = *Mus caroli*, a = *Apodemus sylvaticus*, and r = *Rattus Norvegicus* (Wistar). Probes: CON = pLv54 contrapsin cDNA probe, p108 = p54λ-1-108 subclone, p45 = p54λ-7-45 subclone, p8.1 = p54λ-8-1 subclone. Size markers are shown on the left in kilobasepairs.

<table>
<thead>
<tr>
<th>CON</th>
<th>p108</th>
<th>p45</th>
<th>p8.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>d</td>
<td>c</td>
<td>a</td>
<td>r</td>
</tr>
</tbody>
</table>

Bam Rodent Digests
Mapping of a DBA Spi-2 Gene to the spi-1/spi-2 Complex

The BXD Recombinant Inbred (RI) strain series of DNAs (Taylor, 1982) had previously been used to localise and closely link the Spi-1 and Spi-2 loci on mouse chromosome 12 (Hill et al., 1985). The two gene complexes were thought to be 1 (+/- 1) centimorgan apart, with one recombination event occurring between them within the BXD series. A number of sequences had been mapped to the Spi-2 locus using the contrapsin cDNA probe, pLv54. The DBA genomic probe p54λ-1-108 was now used to search for a polymorphism between DBA/2 and C57BL/6 mice which could be used to determine whether this sequence mapped to the Spi-2 locus. The p54λ-1-108 probe was a suitable DBA genomic sequence to use since it gave a single major band in Southern blots, and was considered a likely candidate for the DBA orthologue of the human α1-antichymotrypsin gene. The enzyme Hind III was found to give a DBA fragment of 8 kb, and a corresponding C57BL/6 fragment of 6 kb. The probe hybridised to other bands at low stringency (Figure 4.5) but repeated analysis indicated that the major fragments were equivalent. A Southern blot of HindIII-digested BXD Recombinant Inbred strain DNAs was analysed, and the result of probing with p54λ-1-108 is shown in Figure 4.6a and 4.6b. Surprisingly, the sequence mapped to the Spi-1 rather than the Spi-2 complex, probably indicating that the previously described recombination (Hill et al., 1985) had occurred within the Spi-2 complex rather than between the two families. This could indicate very close linkage of the two complexes, or dispersion of the contrapsin-related Spi-2 subfamily from its related genes. At this stage, no information was available on the physical linkage of the Spi-2 sequences to each other and mapping of the other DBA genomic clones was not carried out.
**Figure 4.6** Southern blot and analysis of Hind III digested BXD DNAs probed with p54λ-1-108. **4.6a:** Autoradiograph obtained by hybridisation with the p54λ-1-108 genomic subclone. The numbers at the top refer to the BXD recombinant inbred line DNA in that lane, Bl/6 = C57BL/6, DBA = DBA/2, deg = degraded DNA. The major C57BL/6 fragment was considered to be about 6 kilobasepairs, as the other hybridising bands were much less intense at higher stringencies. **4.6b:** Strain distribution pattern obtained from the autoradiograph compared to that obtained for the Spi-1, Spi-2 and Igh loci. B = C57BL/6 pattern, D = DBA/2 pattern, X = recombination between loci in that RI line.

<table>
<thead>
<tr>
<th>Loci</th>
<th>BXD STRAIN NUMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>spi-1</td>
<td>B</td>
</tr>
<tr>
<td>spi-2</td>
<td>B</td>
</tr>
<tr>
<td>Igh-1</td>
<td>B</td>
</tr>
</tbody>
</table>

**Figure 4.6b**
Screening of 129 Cosmid Library

A cosmid library was obtained from Dr L. Stubbs (ICRF, London) to isolate further Spi-2 sequences. The library had been prepared from the inbred mouse strain 129/St which was expected to be similar to DBA at the Spi-2 locus (Hill et al., 1985). It had been derived from spleen DNA by Sau3A partial digestion and cloned into the cosmid vector pcosEMBL2 (Ehrich et al., 1987). Five independently packaged aliquots were supplied, each equivalent to about one genome of mouse DNA. Thus, there was a reasonable chance that any one aliquot would contain clones for a particular DNA sequence, and a better chance that at least one aliquot would contain any specific sequence.

Each aliquot of the library was screened in turn by plating 200,000 colony forming units onto a 20cm x 20cm Hybond-N filter. At pinprick size, (12 hours growth) two duplicate ("slave") filters were taken from the master plate and all three plates left to recover at 37°C for five hours. The slave filters were then processed and probed using the usual Grunstein/Hogness protocol. Duplicate positive colonies were subjected to secondary screening, and could usually be isolated following this stage. The first aliquot was screened with the full length rat 2.1 insert from pSPFL2.1 (see chapter 6). This contains the whole protein coding region and was used because it contained more conserved 5' sequences than any mouse probe at that time; it was thought that a broader range of sequences might be isolated in this way. The filters were washed to 2X SSC stringency. Ten positive colonies were selected for secondary screening, and eight were eventually isolated as single Spi-2 positive cosmid clones. These were amplified as cosmid DNA minipreps for further analysis. A rapid-boiling protocol was initially used, but gave a very low yield of DNA, probably because cosmid DNA is considerably more easily heat denaturable than plasmid DNA. Instead, a scaled down version of the alkaline lysis tech-
nique was found to give reasonable results. The clones were then digested with EcoRI and electrophoresed. It was clear that the clones fell into four groups:

clones 1 + 4 / clones 2 + 3 / clone 6 / clones 7, 8 + 10

Clones 1, 2, 6, and 8 were all given the suffix A, and grown up as large scale DNA preps. They were digested with BamHI and PstI and electrophoresed on an 0.8% agarose minigel. This was then blotted and probed sequentially with the 1.1kb PstI insert from p54\lambda-1-108, a 5'specific 200bp fragment from pDew7 (Figure 6.1), and oligo p54/25-, a conserved oligonucleotide flanking the reactive centre (Figure 4.7a - c).

This analysis showed that all the cosmids except 6A appeared to have two reactive centre regions, and all cosmids, including 6A, appeared to have two conserved 5' regions. One gene in clone 2A contained a complementary sequence to oligo p54/25-. This was the first indication of how closely linked members of the Spi-2 family might be. More accurate mapping was now required to establish the nature of the linkage.

**Cosmid Mapping Strategies**

Complete single and double digests were carried out for cosmids 2A and 6A, using the enzymes BamHI, EcoRI, SalI, XhoI and single digests with SmaI and SphI. The DNAs were run on 0.5% agarose gels overnight at low voltage. DNA was transferred to Hybond-N, and sequentially probed with a conserved exon 5 oligo, the p54\lambda-1-108 PstI fragment, and the 200bp 5' (exon 1-specific) insert, pDEW7 (see Fig 6.1). Fragments hybridising to the probes were then accurately sized, using the \lambda HindIII fragment standards in conjunction with the gelsize computer program (Schaffer and Sederoff, 1981). For later cosmid analysis, the enzymes used for complete single and double digests were BamHI, EcoRI, HindIII, SalI, and XhoI. SmaI was used if internal SalI sites were discovered in a cosmid. An example of a total-digest Southern blot for cosmids 1A, 6A
and 8A, using the 5'-specific 200bp probe pDEW7, is shown in Figure 4.8.

To exploit the total-digest data fully, it was necessary to obtain a restriction map for each cosmid. Because of cosmid size (43kb - 52kb), most restriction enzymes with a six basepair recognition site will cut several times within an insert making mapping by complete digestion extremely difficult. Mapping strategies employing partial digestion have been developed but these require a linear starting molecule (Rackwitz et al., 1984). The bacteriophage-encoded enzyme terminase will specifically cleave a cosmid at a single site within the vector, the \textit{cos} 12bp sequence formed by the circularisation of linear phage DNA, allowing partial digests to be performed (Rackwitz et al., 1985). Dr L.Stubbs kindly supplied a bacterial strain which had been engineered to overproduce terminase, but advised that the quality of extract activity was not entirely reproducible. Evidence for this was quickly produced. The crude extract produced by following the method of Rackwitz provided an activity which not only linearised cosmids, but extensively degraded them. Since no commercial terminase was available at this stage, an alternative strategy was developed.

Cosmids were first digested with the enzyme SalI, which has a single recognition site in pcosEMBL2, within the tetracycline resistance gene. Very few \textit{Spi}-2 positive cosmids contained additional SalI sites in their inserts. The enzyme contains a CpG dinucleotide in its recognition sequence and is found on average once every 100,000bp in CpG deficient mammalian genomes. As well as linear starting material, a second requirement for partial mapping of large DNAs is a specific DNA probe for each end of the molecule. These were obtained by isolating and subcloning the BamHI-SalI (BS probe) and SalI-AccI (SA probe) sequences flanking the tetracycline gene in pBR322 (the donor for the pCosEMBL2 gene).
Figure 4.7 Southern blot analysis of the initial Spi-2 positive cosmids. The filter was probed three times, from left to right with the p54λ-1-108 3'-specific subclone, the 5'-specific rat Spi-2.1 subclone pDEW7, and the exon 5 specific oligonucleotide p54/25-.

Lane 1: λ-Hind III size marker, 2: Bam HI digested Cos1A, 3: Bam HI digested Cos2A, 4: Bam HI digested Cos6A, 5: Bam HI digested Cos8A, 6: Pst I digested Cos1A, 7: Pst I digested Cos2A, 8: Pst I digested Cos6A, 9: Pst I digested Cos8A, 10: PhiX174 Hae III size marker. Approximate position of λ-Hind III size markers are shown at left.
Figure 4.8 Southern blot analysis of *Spi-2* positive cosmid total restriction digests. Probed with the 5'-specific subclone pDEW7. Approximately 0.5 μg cosmid DNA loaded in each lane after digestion. Bam = Bam HI, Eco = Eco RI, Hind = Hind III, Sma = Sma I, B/E = Bam HI and Eco RI double digest, B/H = Bam HI and Hind III double digest, E/H = Eco RI and Hind III double digest. Relevant size markers from λ-Hind III and PhiX174-Hae III digests are shown at left in kilobasepairs.
Figure 4.9 Southern blot analysis of Cos2A and Cos6A partial digests. Partial digests for each cosmid were electrophoresed in duplicate to give two panels for probing with the BS and SA end-specific fragments. Gels were 0.4% agarose to allow resolution of fragments over 30 kilobases. B = Bam HI, Bg = Bgl II, E = Eco RI, H = Hind III, Sp = Sph I. Size markers (λ-Hind III) are shown at left in kilobasepairs.
It was now possible to use these probes on SalI linearised cosmid DNA in an analogous fashion to the method of Rackwitz which employs cosR and cosL oligos to map terminase-linearised DNA.

Partial digests were carried out on SalI digested cosmids 2A and 6A. The same enzymes as for complete digests were normally used: BamHI, EcoRI, HindIII, and XhoI. Conditions for partial digestion were determined empirically. Dilution of enzyme to 1 unit/μl in 1X restriction buffer, and digestion of 1μg of linear cosmid with 1 unit of enzyme for three minutes at 37°C usually gave good results. Reactions were stopped with Orange G mix, divided into two, and run in parallel sets on 0.4% agarose gels overnight (generally until the λ/HindIII 2kb size markers had travelled at least 15 cm from the origin). The gels were then photographed with a ruler to give scale and southern blotted to filters. The filter was then divided into two parts, each with a full set of digests, and probed with either the BamHI-SalI (BS) or SalI-AccI (SA) probes. The resulting autoradiographs gave a pattern of the restriction site positions, with respect to the molecule ends, for each enzyme. These could then be accurately sized and located using the gelsize program and the data assembled into a precise restriction map of the cosmid. An early example for cos2A and cos6A is shown in Figure 4.9.

Assembly of Spi-2 Contigs

Combining the partial restriction maps with the complete digest gene hybridisation data gave an indication of the size and orientation of the Spi-2 sequences on each cosmid. The map for cos2A is shown in Figure 4.10. It contains two apparently complete Spi-2 sequences in a head to tail orientation about 10kb apart. Both genes appear similar in size; also about 10kb. No significant difference in intensity of probe hybridisation was seen for the 3’specific (p54λ1-108) and 5’specific probe (pDew7) between the two genes.
Figure 4.10 Restriction maps of Cos 2A, 6A, 1A and 8A. Maps were produced using the data from specific probes hybridised to total digests, and gelsize analysis of the partial restriction digests. B = Bam HI, E = Eco RI, H = Hind III, Sal = Sal I, Xho = Xho I. Fragments hybridising to the p54\lambda1-108 (3'specific) and pDew7 (5'specific) Spi-2 probes are shown, with arrows indicating the putative direction of transcription. The cosmids are shown linearised at the vector Sal I site. Vector sequences are shown as heavier lines.
Figure 4.10 (cont.)

[Diagram of restriction enzyme digest patterns showing restriction sites at XhoI, EcoRI, and BamHI.]
Figure 4.11 Restriction maps of remaining *SpI-2* cosmids. Maps were produced using the data from specific probes hybridised to total digests, and *gelsize* analysis of the partial restriction digests. B = Bam HI, E = Eco RI, H = Hind III, Sal = Sal I, Xho = Xho I. Fragments hybridising to 3' and 5'specific probes are shown, with arrows indicating the putative direction of transcription. The cosmids are shown linearised at the vector Sal I site. Vector sequences are shown as heavier lines.
Figure 4.11 (cont.)
Figure 4.11 (cont.)
Figure 4.11 (cont.)

[Diagram showing DNA restriction sites and genetic distances between Sal, Cos 3E, and Cos 4F]
Figure 4.11 (cont.)
This may indicate that they are both a similar evolutionary distance from the two probe sequences. Identical analysis of cos 6A, and cos1A and 8A, provided maps of similar resolution (Figure 4.10). The two clones cos1A and cos8A were found to overlap one another for most of their lengths, each cosmid having about 5kb of non-overlapping sequence at opposite ends. They did not overlap or "contig" with either cos2A or cos 6A which were also isolated from one another. Further aliquots of the library were screened at the same density to isolate more genes, and to produce more overlaps. The contrapsin cDNA subclone pBXC1.4 was used as a probe to increase the probability of isolating the contrapsin gene.

Screening of the second aliquot produced cosmids cos1B, cos2B and cos8B (Figure 4.11). As with cos1A and cos8A, cos2B and cos8B overlapped one another almost completely, and contained two genes. They both overlapped with cos1B by about 8kb, and cos1B had a substantial overlap with the cos1A/cos8A pair. Cos8B also overlapped with cos2A at its other end, thus linking all the cosmids except cos6A and providing a contig, a continuous cloned genomic region, of about 120kb, containing six closely linked Spi-2 sequences.

Subsequent screening of the remaining three aliquots of the library produced cos5C and 6C, cos1E and 3E, and cos4F and 5F respectively (Figure 4.11). Cos5C fell within, and cos6C extended to 138kb, the original contig (Contig 1), which now contained seven genes. Cos5F also fell within this region, but cos1E, cos3E and cos4F all overlapped with each other, or with the previously unlinked cos6A, to form a second contig (Contig 2) of 74kb containing three Spi-2 sequences. Partial and complete digests and probing of these clones were carried out as before to determine the gene orientations. The complete maps of both contigs are shown in Figure 4.12.
Figure 4.12 Restriction maps of Contigs 1 and 2 from the 129 mouse Spi-2 complex. These were constructed from overlaps of the cosmids in Figures 4.10 and 4.11 and show orientation of ten genes. B = Bam HI, H = Hind III, Sal = Sal I, Xho = Xho I. Fragments hybridising to the p54λ-1-108 (3’ specific) and pDew7 (5’specific) probes are shown, with arrows indicating the putative direction of transcription. The genes are named by the reactive centre subclone isolated from the relevant cosmid. The cosmids are shown in their overlapping configuration. Repeat sequences within the contigs are shown as checked boxes.
Contig 2 (74 kb)
Analysis of Spi-2 Contig Features

It is clear that the amplification of the Spi-2 sequences in the 129 mouse, relative to the human α1-antichymotrypsin locus, has not occurred as a simple expansion of a tandem array by unequal crossover events. The ten genes were all transcriptionally orientated by using the conserved 5'-specific and 3'-specific probes described above. This showed that there is no predictable pattern of gene orientation, or obvious substructure which has become further amplified. Comparisons of coding region restriction maps demonstrated no clear relationship between neighbouring genes in the contigs, in terms of gross physical structure. A series of amplifications and deletions may well have occurred, perhaps with obscuring gene conversion events, although no evidence was found for gene conversion at this stage.

A survey of repeated sequences within the contigs was carried out to determine whether these could have played a role in the amplification of the Spi-2 locus. The principal mouse repeat families are members of the SINE (Short INterspersed Element) and LINE (Long INterspersed Element) classes of repeated sequence (Hastie, 1989; Singer and Skowronski, 1985). In the mouse the B1 and B2 repeats are SINE family members, and the Mif (now termed L1Md) repeat is the major full-length representative of the LINE family (Bennett and Hastie, 1984). Probes for all three repeat types were hybridised to the cosmid digests used to construct the contigs, and the locations of repeats in the contigs are shown in Figure 4.12. The Spi-2 locus seems to be rather sparsely populated with repeats, as higher densities would be predicted in an average 220kb (Bennett et al., 1984). This may reflect recent amplification of the genes in the locus; repeats may simply not have had time to insert into this region. It is possible that repeat sequences were not detected by hybridisation, due to divergence. Several previously undetected divergent repeats were defined in the murine β-globin locus following com-
puter analysis of the complete locus sequence (Shehee et al., 1989). However, given the distribution of detectable B1, B2, and L1Md repeats in the Spi-2 locus, there is no evidence for their involvement in the gene amplification events. The presence of unrelated repeat classes, or smaller repeat motifs, responsible for unequal crossing-over events cannot be excluded.

A search for restriction sites of infrequently cutting restriction enzymes in the contigs was carried out to determine whether CpG enriched "island" sequences were associated with the 5' ends of any of the genes. CpG islands are commonly found associated with "housekeeping" or universally expressed genes, but more rarely with tissue specific sequences (Bird, 1986). The restriction enzymes BssH II, Mlu I, Not I, Pvu I, Sac II, and Sfi I were used to digest a set of cosmids completely spanning both contigs. Surprisingly, only one restriction site was found for one enzyme; a BssH II site between the genes pCos3E46 and pCos3E2. This lies within five kilobases of the only Sal I site in the contigs (Figure 4.12). This indicates that none of the cloned Spi-2 genes possess a 5' CpG enriched island.

Analysis of Reactive Centre Sequences in the Contigs

A total of ten distinct Spi-2 related sequences had now been isolated from the 129 genome, and these were further characterised by subcloning and sequencing. In almost all the cosmids, two Spi-2 sequences were present, and it was not possible to sequence the cosmid directly since conserved oligonucleotides would sequence both genes simultaneously. Subcloning was simplified by the existence of the detailed restriction maps used to construct the contigs, and it was often possible to "force-clone" selected gene fragments from double digests of cosmids without band isolation. In this way, all but one of the contig reactive centres was cloned into pBluescribe for sequence analysis. Sequencing was carried out directly on ds plasmid DNA using the conserved
oligonucleotides p54/XV+, p54/129+ and p54/51- where appropriate. The nucleotide and putative amino acid sequences obtained from the nine subcloned 129 Spi-2 genes are shown in Figure 4.13a and 4.13b.

Inspection of the reactive centre sequences revealed that the genomic contrapsin sequence had not been isolated on the contigs. However, probable orthologues of all four DBA reactive centres were found in the 129 genes. These were not always identical between strains, but always highly related (Figure 4.14). Whether this represents polymorphism in possible Mus domesticus progenitors of the DBA and 129 strains, or whether the strains have separate species components at the Spi-2 locus is not clear. None of the 129 sequences bore obvious similarities to Spi-2 sequences from species outside the genus Mus, apart from the orthologue of the DBA p54α-1-108 leu-ser α1-antichymotrypsin-like gene described above.

Translation of the reactive centre sequences showed that the cys-cys P1-P1' gene found in DBA, p54α7-45, was not unique since three 129 reactive centres shared this P1-P1' sequence. These genes were all clustered in one half of Contig 1 adjacent to the arg-cys 129 orthologue (pCos2B2) of the DBA his-cys gene (p54α8-1). Inspection of the cys-cys and arg-cys sequences showed that they were the most closely related, and presumably most recently diverged, sequences in the contigs. This was later confirmed using the PHYLIP computer programs (Felsenstein, 1985) to generate unrooted trees of the sequence phylogenies (see below). Three genes in the contigs look unlikely to encode active inhibitors, due to radical amino acid substitutions in the highly conserved FNRPFL and TEAAAAT domains which flank the reactive centre. No termination codons or frameshift mutations were found in any of the sequences, but the pCos3E2 gene in particular shows very little conservation of amino acid sequence, and resembles a typical "rotting hulk" pseudogene (Howard, 1987).
Figure 4.13a Nucleotide sequences of the nine Spi-2 genomic clones (exon 5). Clones are shown in alphabetical and numerical order in two groups; putative active genes above and putative pseudogenes below. Sequence is shown from codon 1 of exon 5 to the termination codon if complete. The reactive centre and putative P1 residue is boxed. Dashes represent identity with the top sequence, stars represent deletions relative to the top sequence.
Figure 4.13b Predicted amino acid sequences of the nine Spi-2 genomic clones (exon 5). Clones are shown in alphabetical and numerical order in two groups; putative active genes above and putative pseudogenes below. Sequence runs from residue 1 of exon 5 to the termination codon if complete. The putative P1 residue is starred. Dashes represent deletions relative to the top sequence. Region 2 (the reactive centre) is boxed. Below is a histogram showing the number of alternative residues found at each amino acid position in exon 5. Data was taken from the seven complete aminoacid sequences. The hypervariability of Region 2 (the reactive centre) is obvious. The position of the P1 residue is arrowed.
Figure 4.14 Comparison of DBA and 129 Spi-2 orthologues. The four sequences are aligned in pairs with DBA above and 129 sequence below. Region 2 (the reactive centre) is boxed. The P1 residue position is marked by a star. X indicates a difference between the DBA and 129 sequences. Numbers of synonymous and asynonymous substitutions in exon 5 are shown below each alignment.
Disregarding the DBA orthologues, recent Cys-Cys duplications, and probable pseudogenes, the remaining sequence, pCos3E46, appeared likely to encode an active inhibitor of predictable specificity. This gene had the P1-P1' residues Arg-Ser in common with several known human serpins. Attempts to determine if it was expressed in the mouse are described in Chapter 5.

The nucleotide and predicted amino acid sequences from the 129 contigs were analysed for evolutionary relationships by the DNAPARS and PROTPARS programs in the PHYLIP 3.1 computer program package mounted at DLVH (Daresbury). Both programs infer unrooted phylogenies from biological sequences using a maximum parsimony method analogous to the traditionally employed Wagner tree algorithm (Felsenstein, 1985). A single tree was generated using either nucleotide and protein sequences (Figure 4.15). The relatedness of sequences did not correlate in an obvious fashion with their position or orientation on the contigs. Either a complex series of amplifications may have occurred, involving subsequent deletions or sequence inversions, or a considerable degree of gene conversion among the sequences has obscured the true phylogeny of the 129 Spi-2 genes. A combination of both processes is probably more likely. Interestingly, the sequence which most resembled human α1-antichymotrypsin in the reactive centre region is the most divergent mouse gene, and is closest to the α1-antichymotrypsin outgroup. This is the pcos2B1/p54λ-1-108 leu-ser genomic sequence. It does not appear to have undergone recent duplication (Figure 4.5), but may well have given rise to two expressed rodent-specific Spi-2 sequences prior to rodent radiation (discussed in Chapter Five).
Figure 4.15 Phylogeny of the 129 genomic Spi-2 sequences. The unrooted tree shown was produced by the PROTPARS program in the PHYLIP 3.1 Phylogeny inference package (Felsenstein, 1985). An identical topology was produced using nucleotide sequences for the same region. The length of each branch indicates the number of aminoacid substitutions from the previous branchpoint. This tree is un-rooted, but taking human α1-antichymotrypsin (Hachy) as an outgroup indicates that the root should probably be located between Hachy and 2B1. Below, a diagram of gene location and orientation on the contigs indicates the lack of an obvious relationship between map location and sequence.
Attempts to Isolate a Contrapsin-specific Genomic Sequence

The contrapsin genomic sequence, shown by Northern analysis to exist in the 129 mouse, had not been isolated at this stage, despite its use as a library probe. Specific library screens were undertaken to isolate cosmids containing the contrapsin gene using highly specific cDNA fragments, and oligonucleotides specific for the contrapsin gene. Two separate attempts using an exon 5 (reactive centre exon)/3' untranslated region probe at high stringency reisolated Spi-2 cosmids which had already been cloned, presumably due to partial similarities in the exon 5 coding sequences. A third screen using the contrapsin specific oligonucleotides p54/RC- and p54/3'- produced a single cosmid clone which hybridised strongly to the oligonucleotide mixture. This cosmid, appropriately named Coscon, was restriction mapped by partial and total restriction digests and a map is shown in Figure 4.11. The oligo mixture hybridised to a specific region of the insert, but repeated probings with the p54\lambda-1-108 reactive centre exon probe, the pDEW7 conserved 5' specific probe, and even the contrapsin cDNA probe pBXC1.4 did not reveal any homologous sequences in the cosmid DNA. Either the contrapsin gene possessed a unique genomic structure with the reactive centre flanked by introns of at least 7 and 35 kilobases, or the oligonucleotide positive hybridisation had occurred to a random genomic sequence. Unfortunately, the sequencing to establish the nature of the oligonucleotide homology was not carried out at this stage.
Conclusions

Genomic $Spi$-2 clones were isolated from two inbred mouse genomic libraries in an effort to establish the number, variability, and genomic organisation of these genes in *Mus domesticus*. The DBA/2 library yielded four independent reactive centre sequences, none of which encoded contrapsin. One of the sequences resembled the human $\alpha_1$-antichymotrypsin sequence in the reactive centre region, but all efforts to detect expression in the mouse liver failed. This sequence, p54$\lambda$.1-108 was mapped to the $Spi$-1 complex in the BXD recombinant inbred strain panel, although it is clearly an $Spi$-2 gene. This may indicate that the $Spi$-1 and $Spi$-2 complexes are virtually contiguous, with the BXD recombination occurring within the $Spi$-2 complex. The intragenic structure of the $Spi$-2 sequences seems to have been completely conserved since mammalian radiation, since all sequences examined display the human $\alpha_1$-antichymotrypsin intron/exon structure.

The second genomic library yielded ten distinct $Spi$-2 sequences contained in two contiguous genomic segments of 138 kilobases and 74 kilobases. Analysis revealed that the genes were closely clustered with an average spacing of 10 to 15 kilobases. Their orientation, and that of repeated sequences within the contigs, did not reveal an obvious mechanism for the amplification of the locus. Sequencing of the reactive centres from nine of the 129/St $Spi$-2 genes revealed orthologues of all four DBA genes, two probable pseudogenes, two further genes with Cys-Cys reactive centres, and a gene which resembled the human PCI serpin in the reactive centre region. Analysis of the sequences with phylogeny inference programs confirmed the lack of relationship between location on the contigs and evolutionary descent, and indicated that pcos2B1/p54$\lambda$.1-108 sequence may indeed represent the primordial $Spi$-2 gene.
CHAPTER FIVE

Characterisation and Distribution of Expressed Spi-2 Sequences

Introduction

The liver is the major organ known to express Spi-2 transcripts in mammals, including rodents. The original pLv54 cDNA clone which encoded contrapsin (Hill et al 1984) had been isolated from a Swiss mouse liver cDNA library. The rat cDNA clones pDex2 and pInf5, encoding the rat 2.1 (arg-arg) and 2.2 (val-ser) inhibitors, were also cloned from a liver cDNA library (Hill and Hastie 1987).

A Northern blot to examine the tissue distribution of Spi-2 transcripts, carried out by R.Hill before this work, indicated that low levels (about 1/1000 of liver level) of a 20S transcript could be seen in BALB/c spleen and testes, and a 16S transcript could be found in the submaxillary gland at a similar abundance. However, the teratocarcinoma cell line EB22, derived from the 129 strain of inbred laboratory mouse by Dr Martin Hooper, expressed a 20S transcript (or transcripts) at levels 10 to 20 fold higher than the tissues mentioned above; perhaps as high as 1/100 of the hepatic level. The exact nature of the EB22 cell line is not certain, but appears to have differentiated towards a chondrocytic phenotype (M.Hooper, personal communication). No other rodent tissues were known to express detectable levels of Spi-2 sequence.

A series of cDNA library screens were carried out to determine the nature of novel expressed Spi-2 sequences: a) in the liver, using probes derived from novel genomic sequences (described in Chapter 4) and b) by screening species or tissues which had not been previously characterised for Spi-2 sequences. Initially, attempts were made to isolate cDNAs corresponding to cloned genomic sequences.
Screening Genomic Clones for Expressed Sequences

Following isolation of the mouse genomic Spi-2 sequences from the DBA and 129 strain genomic libraries (Chapter 4), the question arose as to which, if any of the genes were transcribed, and in what tissues.

DBA genomic clones

The original DBA-derived genomic clone p54-λ-8.1 (cys-cys) did not resemble the expected anti-chymotryptic reactive centre, leu-ser. An oligonucleotide (oligo p54/62-), complementary to the putative mRNA, did not reveal any signal in various tissues examined. This oligo was also designed to hybridise to the rat Spi-2.2 sequence, and did give a 20S signal on Northern blots of rat liver RNA (not shown).

The next DBA genomic clone to be isolated, p54-λ-1.108, had the P1-P1' residues leu-ser and looked an ideal anti-chymotrypsin candidate. An antisense 18mer oligonucleotide (p54/LS-) was made, complementary to the DNA encoding the reactive centre amino acid residues APLSAK, and this was used to probe the DBA cDNA pBR322 plasmid library. One apparent positive clone was observed on the primary screen of 40000 colonies, but proved to be artefactual after repeated library screens. Several Northern blots using this oligo as a p54-λ-1.108-specific probe never gave a detectable signal. It was concluded that this gene was unlikely to be the functional equivalent of human α1-antichymotrypsin. The two remaining DBA genomic sequences, p54-λ-7.45 (cys-cys), and p54-λ-4.13 (met-ser) were not investigated in terms of tissue transcription at this stage. The use of the full length genomic fragments (1 to 3 kb) as Northern blot probes was not informative since all the sequences cross-hybridise to the contrapnsin sequence with which they were initially isolated.
Figure 5.1 Alignment of cos3E-46 and human PCI reactive centre sequences. The protein and DNA sequences of the two genes are shown, with stars representing identical residues or nucleotides. Interestingly, the protein similarity in Region 2 (7/15 residues) is higher than the nucleotide identity over the same sequence (18/45 nucleotides). Region 2 is boxed for both alignments. The location of the 18mer antisense oligonucleotide p54/PCI- is shown above the nucleotide sequence for pcos3E-46. HuPCI = human PCI (Suzuki et al., 1987), 3E46 = pcos3E-46 (Figure 4.13)
Figure 5.2 Nucleotide and protein sequence of pMPCI-2 (exon 5 only). This sequence is identical to the 129-derived reactive centre sequence from pcos3E-46, despite its isolation from a DBA/2 liver cDNA library (cf Figure 4.14). The amino acid sequence is repeated below, aligned with that of contrapsin. The P1 residue is indicated by an asterisk.

V V H K A V L D V A E T G T E A A
GTG GTC CAC AAG GCT GTG CTG GAC GTG GCT GAG ACA GGC ACA GAA GCA GCT

A A T G F I F G F R S R R L Q T M
GCT GCC ACA GGG TTC ATT TTT GGC TTT CGT TCT AGA AGA TTA CAA ACT ATG

T V Q F N R P F F L L M V I S H T G V
ACT GTG CAG TTC AAC AGG CCA TTC CTG ATG GTC ATC TCT CAC ACA GGT GTT

Q T T L F M A K V T N P K !
CAG ACT ACC CTC TTT ATG GCC AAA GTC ACT AAC CCC AAG TGA

* MPCI-2 VVHKAVLDVAETGTEAAAATGFIFGFRSRLQTMVQFRNPFLMVISHTGVQTTLFMAKVTNPK!
129 strain genomic clones

The characterisation of the Spi-2 genomic sequences from the 129/St mouse strain again raised the question of which novel genes were expressed, and where. The sequence of pCos3E-46 was interesting in that it resembled the reactive centre of the human serpin PCI, (Protein C Inhibitor (Suzuki et al, 1987)) in seven of ten residues in the amino terminal region of its reactive centre (Figure 5.1).

An antisense 18mer oligonucleotide (p54/PCI-) was synthesised, corresponding to the sequence GFRSRR, and used to screen the DBA mouse liver cDNA library (50000 colonies). Five hybridising signals were observed on duplicate filters, and positives were obtained for each primary on secondary and tertiary screens. Mini-preps of the plasmids were grown up, and the longest PstI insert, from clone 2 (1.4kb), was selected for further analysis. The insert was excised from the pBR322 vector with PstI and sub-cloned into pBluescribe. This plasmid was named pMPCI-2, and sequencing of CsCl purified dsDNA using flanking and internal oligonucleotides confirmed that it represented a transcript from the pCos3E-46 genomic sequence (Figure 5.2).

This was the first Spi-2 sequence for which both a genomic and a cDNA sequence was available. Northern blots of mouse tissues, and of various murine species liver RNA were carried out using the oligonucleotide, but unfortunately it hybridised fairly strongly to 18S and 28S ribosomal bands. This was not a problem when the cDNA library was screened since enrichment for mRNA sequences had occurred during library construction. However the pMPCI-2 transcript would be expected to be of low abundance (1 in 10000 cDNA clones), and possibly about the same size as the 18S rRNA transcript. Even when poly (A)+RNA was used, no specific clean signal was observed for the pMPCI-2 oligonucleotide.
Figure 5.4 Sequence comparison of the MPCI gene to other inhibitors (exon 2). 5.4a: Alignment of the pMPCI-2 exon 2 nucleotide and predicted amino acid sequences with the published rat 4.12B (=Spi-2), human α1-antichymotrypsin and human protein C inhibitor amino acid sequences. The MPCI protein sequence is shown in upper case. MPC = pMPCI-2, 4.12 = rat 4.12B, HAC = human α1-antichymotrypsin, HPCI = human protein C inhibitor. 5.4b: Table showing the proportion of identical residues for all four protein sequences shown in 5.4a. The figures support the idea that the MPCI sequence is Spi-2-like rather than PCI-like.

Figure 5.4a

<table>
<thead>
<tr>
<th></th>
<th>MPC</th>
<th>4.12</th>
<th>HAC</th>
<th>HPCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPC</td>
<td>GAAGGCAAGGCCTCTCTGACCAGACTGAGCCCTTACAGCGACTTTACAGAGGCCTACTGA</td>
<td>ktraly qaeafvadfkqce</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPC</td>
<td>KARALYQTEAFTADLKPTTE</td>
<td>4.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HAC</td>
<td>dakrlygsafatdfqdsaa</td>
<td>HAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPCI</td>
<td>amktlyladtfptnfrdsag</td>
<td>HPCI</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.4b

<table>
<thead>
<tr>
<th></th>
<th>MPC</th>
<th>4.12</th>
<th>HAC</th>
<th>HPCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPC</td>
<td></td>
<td>27/40</td>
<td>21/40</td>
<td>13/40</td>
</tr>
<tr>
<td>4.12</td>
<td></td>
<td></td>
<td>22/40</td>
<td>18/40</td>
</tr>
<tr>
<td>HAC</td>
<td></td>
<td></td>
<td></td>
<td>22/40</td>
</tr>
</tbody>
</table>
The pMPCI-2 insert was also used to probe Northern blots of various murine liver RNAs ("evolutionary blots"), and part of an evolutionary blot is shown (Figure 5.3a). The cDNA gives a similar hybridisation pattern to contrapsin and is presumably crosshybridising to the abundant contrapsin message as well as its own less abundant 18S transcript.

It appears from additional sequence analysis (Figure 5.4) that the pMPCI-2 transcript may not be a strict orthologue of the human PCI gene, which diverged from antichymotrypsin about the time of mammalian radiation (Suzuki et al 1987). The mouse PCI-like gene appears to have diverged from the antichymotrypsin gene considerably later, perhaps around the time of rodent radiation (~15 million years ago). This will be further discussed after consideration of the other rodent results (Chapter 7).

No other reactive centre sequences from the 129 cosmid contigs looked as likely as pCos3E-46 to encode functional inhibitors, either because of unusual P1 residues (eg cysteine), or because of potential disruption of the highly conserved TEAAAAT or FNRPFL boxes. Alternative approaches to defining expressed genes were attempted.

**Expressed Spi-2 Sequences in Apodemus sylvaticus Liver**

To establish the nature of expressed Spi-2 sequences in *Apodemus sylvaticus*, a λgt10 cDNA liver library (from R.Hill) was screened. The library was constructed from poly(A)+ RNA prepared from the male specimen trapped at the Bush estate for the original contrapsin RNA sequencing study (Chapter 3). It had been shown by that study that *A. sylvaticus* did not express high levels of Spi-2 message in the liver, but that an 18S and a 20S transcript were visible in approximately equal amounts at low abundance (Figure 3.7). No evidence for a contrapsin-like sequence had been obtained.

The library was plated out at a density of $1 \times 10^5$ plaques per 20x20cm plate and
screened with a 5'fragment of pSPFL2.1 (encoding the rat Spi-2.1 product) and a conserved oligonucleotide (oligo p54/129+) on duplicate filters. A total of 12 positives were observed at the primary stage, these appeared to fall into two groups on the basis of hybridisation intensity. Four clones remained positive during secondary and tertiary screens using the 5' cDNA fragment only. Restriction analysis of the individual clones confirmed the presence of two classes of cDNA clone in equal abundance. These were sequenced, using oligonucleotide primers (oligos p54/51- and p54/129+) to conserved sequences which flank the reactive centre. The sequences obtained are shown (Figure 5.5).

Clone 12 (Figure 5.5a) showed considerable similarity to the rat Spi-2.2 (val-ser) sequence described by Hill and Hastie (1987). This sequence is considered to represent the rat equivalent of human α1-antichymotrypsin on the basis of message size, chemical nature of P1 residue and inducibility during the acute phase response. The Apodemus clone appeared to be the direct equivalent of rat Spi2.2, with the interesting difference of a valine to leucine substitution at the P1 residue. This would be unlikely to alter the inhibitory specificity of the molecule since the putative human orthologue also has a leucine residue in the P1 position.

Clone 1 (Figure 5.5b) was also very similar to a previously cloned sequence, the 129 and DBA-derived gene pCos3E-46 (genomic) or pMPCI-2 (cDNA). Like its mouse counterpart, Apodemus clone 1 has an arg-ser reactive centre and shows a similar degree of Apodemus/mouse divergence to the α1-PI and albumin transcripts in the RNA sequencing experiments (Chapter 3). The existence of such similar genes in both Apodemus and mouse echoed the similarity of the Apodemus Clone 12 (α1-antichymotrypsin like) and the rat spi-2.2 sequence discussed above.
Both *Apodemus* sequences, Clone 1 and Clone 12 were subcloned into pBluescript for further analysis, yielding plasmids pAP-5 and pAAC-8 respectively. Both cDNAs were used as probes to the murine "evolutionary blot" (Figure 5.3b and 5.3c) previously described in Chapter 3. This contains a series of liver RNAs from mice which diverged from *M. domesticus* after the rat.

The pAP-5 clone (arg-ser) gives a very similar 18S mRNA hybridising pattern to that of contrapsin and of pMPCI-2, the DBA arg-ser cDNA. This would be expected as these sequences are closely related. The only obvious difference between the *Apodemus* and DBA probes is that the *Apodemus* probe hybridises more efficiently to the low abundance *Apodemus* PCI-like transcript than the mouse probe. The equivalent effect for the DBA cDNA probe and DBA arg-ser transcript is obscured by the presence of the contrapsin sequence in mouse, to which both *Apodemus* and DBA probes hybridise with similar efficiencies.

The pAAC-8 (*Apodemus* antichymotrypsin-like) probe preferentially hybridises to a 20S mRNA transcript in the evolutionary blot. An 18S transcript is detected in recently diverged mice but this is probably due to high levels of contrapsin mRNA in these murids. The presence of contrapsin mRNA at levels perhaps one hundred-fold higher than the 20S transcript would compensate for decreased homology between the pAAC-8 probe and contrapsin transcript. The pAAC-8 probe hybridises efficiently to a 20S *Apodemus* transcript as expected, but less efficiently to the rat 20S, to which it appears orthologous by sequence. A 20S transcript is not visible in C57/BL6 or *Mus musculus* liver, confirming previous experiments (Hill *et al*, 1985) which demonstrated that the 20S transcript was not constitutive at detectable levels in these mice.
Figure 5.5 Expressed *Apodemus sylvaticus* spi-2 Sequences. 5.5a: Sequence of *Apodemus* Clone 12 (exon 5). The DNA and protein sequences are shown. Below the protein sequences of Clone 12 and the rat orthologue *Spi-2.2* are shown aligned. Apo 12 = *Apodemus* Clone 12, rat 2.2 = rat *Spi-2.2*. 5.5b: Sequence of *Apodemus* Clone 1 (exon 5). The DNA and protein sequences are shown. Below the protein sequences of Clone 1 and the mouse orthologue MPCI-2 are shown aligned. Apo 1 = *Apodemus* Clone 1, mpci = MPCI-2. Dashes represent sequence identity. The position of the P1 residue is indicated by an asterisk.

Fig 5.5a

Apo 12 V V H K A V L D V A E K G T E A A
exon 5 GTG GTC CAC AAG GCT GTG CTG GAC GTG GCT GAG AAA GCC ACA GAA GCA GCT

A A T G V K I V P L S A K L D P L
GCT GCC ACG GGA GTC AAA ATT GTC CCA CTG TCT GCA AAG CTG GAC CCT TTG

T I Y F N R P F L V A I Y D E T T
ACT ATA TAT TTC AAC CGG CCT TTC CTG GTG GCT ATC TAT GAC ACA GAA ACT

E I I Y F L G K I S N P K !
GAA ATT ATC TAC TTT TTG GCC AAA ATA TCC AAC CCC AAA

Apo 12 VVHKAVLDVAEGTEAAATGV KIVPSAKLDPITIY FNRPFLVAIYDTETIIFYFLGKISNP!
rat 2.2 ----------- T ---------- ----V ------- I-A ----MI-S----A-AP---A---F-----

Fig 5.5b

Apo 1 V V H K A M L E V A E K G T E A A
exon 5 GTG GTC CAC AAG GCT ATG CTG GAA GTG GCT GAG AAA GCC ACA GAA GCA GCT

A A T G V K F V F R S G R V P T M
GCT GCC ACA GGA GTC AAG TTT GTC TTT GTC TCT GGT GCC AGA GTA CCA ATG

T V R F D R P F L M V V S H T G V
ACT GTG AGG TTT GAC AGG CCA TTC CTC ATG GTT GCT TCT CAC ACA GCT GTT

E S I L F F G K V T N P N !
GAG TCT ATC CTC TTT TTG GCC AAA GTC ACC AAC CCC AAT

Apo 1 VVHKAMLEVAEGTEAAATGV KIVPSAKLDPITIY FDRPFLVVSHTGVEISIFGKVVTNPN!
mpci V-D- ----T--------F I-G----R-LQ-----Q --N------I------QTT--MA--I--K!
The conservation of both the Spi-2.2-like and PCI-like genes between rodent species might be due to selection for internal physiological function. If this were the case, inbred mice would be expected to express an antichymotrypsin-like sequence, orthologous to rat Spi-2.2 and Apodemus Clone 12, and the rat would be expected to express a PCI-like sequence orthologous to mouse pCos3E-46 and Apodemus Clone 1.

**Screening for Mouse Anti-Chymotrypsin**

An attempt was made to isolate the putative mouse anti-chymotrypsin. Screening of the DBA cDNA liver library which yielded the pMPCI2 clone was carried out using an oligonucleotide complementary to the SAKLPD codons of the Apodemus Clone 12 and Rat Spi-2.2 gene pair. This did not give any positive signals in a screen of about 5 x 10^4 colonies. There was a high chance that the oligo would not be exactly complementary to the putative mouse antichymotryptic sequence even if the amino acid residues were conserved, due to codon degeneracy. This later proved to be the case.

**Screening of EB22 cDNA Library for Spi-2 Sequences**

The rodent liver could not readily be searched further for Spi-2 sequences which could be specifically screened for (excluding randomly selected reactive centre sequences from the genomic contig). The next tissue to be examined was the mouse cell line EB22. This is considered to represent a chondrocytic-derivative cell type from a teratocarcinoma, and had been shown previously to express a 20S Spi-2 mRNA. Total EB22 RNA, isolated by R.Meehan, was used to produce poly(A)+ enriched RNA and a cDNA library was then constructed in λgt10 by R.Hill. Screening of 10^5 plaques was carried out with pSPFL2.1, the full length rat Spi-2.1 cDNA. Five positive plaques were obtained, and isolated clones examined by restriction analysis. These fell into two groups; clones 1, 4 and 5, and clones 2 and 3. Sequencing was carried out on clones EB22/3 and EB22/5,
using internal primers (oligos p54/129+ and p54/51-) to reactive centre flanking regions.

The sequence from clone EB22/3 (Figure 5.6a) proved to be identical to that of the 129 genomic sequence pCos2A2, which has a putative cysteine-cysteine reactive centre bond. This was unexpected as a cysteine-cysteine pairing had not previously been observed and it had been thought that the genomic cysteine-cysteine sequences discovered in the contigs (Chapter 4) might represent inactive descendants of duplication events. The fact that this gene was at least transcribed, and at equivalent levels to the EB22/5 met-ser gene, was an indication that other unusual sequences in the contig might also be active. The effect upon inhibitor specificity of a cysteine in the P1 position is unclear.

The sequence obtained from clone EB22/5 is shown (Figure 5.6b) aligned with the Apodemus Clone 12 and Rat 2.2 reactive centres. It appears likely to encode the mouse orthologue of the α1-antichymotrypsin gene from the other two rodents. A major difference however is the presence of a methionine residue in the P1 position, compared to a valine in Rat and a leucine in Apodemus. This might be expected to alter the target specificity of the inhibitor to some degree, as methionine is a preferred residue for hydrolysis by elastase type proteinases, rather than chymotrypsins.

*Analysis of Putative Murine α1-antichymotrypsin*

If the EB22/5 clone actually represents the mouse orthologue of *Apodemus* clone 12 and rat *Spi-2.2*, it should meet certain functional criteria. The other rodent sequences are expressed in the liver as a 20S transcript, which, at least in the case of the rat, are induced as part of the inflammatory response. The mouse also expresses an inducible 20S transcript (Hill *et al*, 1985), which had not been characterised. EB22/5 represented a candidate sequence for this mRNA.
Figure 5.6 Expressed Spi-2 sequences from EB22 cells. 5.6a: Nucleotide and predicted amino acid sequence of exon 5 for clone EB22/3 (cys-cys). The cDNA sequence is identical to that of the 129 genomic Spi-2 subclone pcos2A2 over the region sequenced. 5.6b: Nucleotide and predicted amino acid sequence of exon 5 for clone EB22/5 (met-ser). The location of the antisense 18mer oligonucleotide p54/226- is shown. Below, the protein sequences for EB22/5, Apodemus Clone 12 (Apo 12) and the rat Spi-2.2 (rat 2.2) gene are aligned. Dashes represent residue identity. The P1 residue is indicated by an asterisk.

Figure 5.6a

<table>
<thead>
<tr>
<th>EB22/3</th>
<th>GTG GTC CAC AAG CTT GCX GCT GTG CTG GAT GCT GAG ACA GCC ACA GAG GCT GCT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A A T G M A G V G C C A V F D F L</td>
</tr>
<tr>
<td></td>
<td>GCT GCC ACA GCA ATG GCA GCA TGT TGT GCA GTC TTT GAC TTT CTG</td>
</tr>
<tr>
<td></td>
<td>E F F N R P F L M I I S D T K A</td>
</tr>
<tr>
<td></td>
<td>GAA ATA TTT TTT AAG CCA TAC ATG ATG ATG ATG ATG ATG ATG ATG ATG ATG ATG</td>
</tr>
<tr>
<td></td>
<td>H I A L F M A K V T N P K</td>
</tr>
<tr>
<td></td>
<td>CAC ATT GCC CTC TTT ATG GCA AAA GTT ACA AAA CAA AAG GCA</td>
</tr>
</tbody>
</table>

Figure 5.6b

<table>
<thead>
<tr>
<th>EB22/5</th>
<th>GTG GTC CAC AAG CTT GCX GCT GTG CTG GAT GCT GAG ACA GCC ACA GAG GCT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A A T G V K F V P M S A K L Y P L</td>
</tr>
<tr>
<td></td>
<td>GCT GCC ACT GGA GTC AAA TTT GTT CCA ATG TCT GCG AAA CTG TAC CCT CTG</td>
</tr>
<tr>
<td></td>
<td>&lt;*** *** *** *** *** ***</td>
</tr>
<tr>
<td></td>
<td>oligo p54/226-</td>
</tr>
<tr>
<td></td>
<td>ACT GTA TAT TTC AAG CCT TTC CTG ATA ATG ATG ATG ATG ATG ATG ATG ATG</td>
</tr>
<tr>
<td></td>
<td>E I A P F I A K I A A N P K</td>
</tr>
<tr>
<td></td>
<td>GAA ATT GCC CCC TTT ATA GCC AAG ATA GCC AAG CCC AAA</td>
</tr>
</tbody>
</table>

* 

EB22/5  VVHKAVLDAETGTEAAAATGV KFVPMSAKLYPLTVY FNRPLIMIFDTETEIAPPIAKIANPK!
Apo 12  ----------- K ----------- -I-L-D-I- V-Y- IY-LG-S---
rat 2.2  ---------------------- V-D-IIA -D-MI-S-A-L-F---
Figure 5.7 Induction of Mouse met-ser Spi-2 gene by acute phase reactants. A: Northern blot of liver and EB22 RNAs probed with the mouse met-ser specific oligonucleotide pS4/226-. Lane 1: 2µg DBA poly(A)+ liver RNA, 2: 10 µg C57BL/6 total liver RNA (uninduced control), 3: 10 µg C57BL/6 total liver RNA (induced by 50 µg LPS), 4: 10 µg EB22 total RNA (uninduced control), 5: 10 µg EB22 total RNA (induced by Dexamethasone). B: Northern blot of liver and EB22 RNAs probed with the mouse met-ser specific oligonucleotide pS4/226-. Lane 1: 10 µg DBA poly(A)+ liver RNA, 2: 1 µg DBA poly(A)+ liver RNA, 3: 2 µg C57BL/6 poly(A)+ liver RNA (uninduced control), 4: 2 µg C57BL/6 poly(A)+ liver RNA (induced by 50 µg LPS), 5: 10 µg EB22 total RNA (induced by Dexamethasone). Sizes of 18S, 20S and 28S transcripts are indicated.
An oligonucleotide (p54/226-) was synthesised complementary to the reactive centre region and was used to probe a Northern blot of uninduced and induced mouse liver and EB22 cell RNA (Figure 5.7). These showed that EB22/5 did express a 20S inducible transcript in both liver and EB22 cells. This clone is highly likely to represent the strict murine orthologue of Apodemus clone 12 and the rat Spi-2.2 by sequence and function, and the functional equivalent of human α1-antichymotrypsin.

Oligo p54/226- was also used to probe the murine Northern "evolutionary blot" (Figure 5.3d). This showed that the reactive sequence was conserved and expressed in the livers of mice as divergent from Mus domesticus as Mus cervicolor and Mus caroli. This degree of conservation is similar to that of contrapsin and may represent fixation of the reactive centre sequence following the acquisition of the methionine P1 residue.

Analysis of Spi-2 expression in the 16-day mouse

Previous analysis (Barth et al., 1982; Meehan et al., 1984) of contrapsin gene expression in the developing mouse embryo had shown that contrapsin is not detectable in the liver before seventeen days of development, when it rapidly rises to the high level of transcription found in the adult. Apart from the adult tissues mentioned at the beginning of the chapter (submaxilliary gland, spleen, and testis) no other mouse tissues were known to express Spi-2 sequences. In situ hybridisations to mounted sections of mouse embryo were carried out to establish further patterns of expression. The stage chosen was the sixteen day embryo as it was considered that the high levels of hepatic contrapsin expression from more advanced stages might obscure other areas of Spi-2 expression.

Sections of sixteen day old mouse embryo (cut and prepared by D.Davidson) were probed with a $^{35}$S-UTP labelled RNA probe, transcribed using phage T7 polymerase, from the contrapsin subclone pBXC1.4. This probe is relatively specific for the Spi-2.1
family members, but can hybridise to most Spi-2 sequences at normal stringency (65°C, 2X SSC). The in situ hybridisation procedure was supervised by E.Graham and followed established protocols (Davidson et al., 1988). Slides were developed after a three week autoradiography period and examined, using dark-field microscopy, for areas of positive hybridisation.

There was no detectable probe hybridisation to the liver, confirming the absence of hepatic contrapsin transcription before seventeen days of development. Areas of positive hybridisation were found in several sections, and upon examination appeared to be localised to areas where cartilage-like tissue had been deposited. Three examples are shown in Figures 5.8 to 5.10. Figure 5.8 shows an area from a transverse section through the head of the embryo. The dark-field photograph shows increased hybridisation to an umbrella-shaped structure, which can be localised to a particular cell-type on the light-field print. This structure is the nasal septum, separating the developing nostril airways, and is composed of cartilage. Figure 5.9 shows a detail from a transverse section through the mouse in the pectoral girdle region. The elbow joint is clearly visible in the upper centre of the light-field photograph, and cells of a similar appearance to the nasal septum cartilage can be seen on either side of the joint. In the dark-field photograph these can be seen to be positive for an Spi-2 transcript. The hybridisation does not extend to the boundaries of the bones at the joint; it appears that the transcripts are not found in fully differentiated bone. Finally, Figure 5.10 shows part of a sagittal section through the mouse, in the area of the lower limb. The internal structure of the foot is obvious in the light-field illustration, and heavy hybridisation to inner areas of the skeletal structure can be seen in both light and dark-fields. Again, little signal is obtained from the boundaries of the skeletal structures.
Figure 5.8 Light and dark-field photomicrographs of a transverse section through 16 day mouse (head region). 56X magnification. The light-field print (top) shows the cartilaginous tissue forming the nasal septum as an umbrella shaped structure. In the dark-field print below, the clustering of silver grains (light coloured) can be seen overlying this structure which is composed of a distinctive loosely-clustered cell type.
Figure 5.9 Light and dark-field photomicrographs of a transverse section through 16 day mouse embryo (pectoral girdle region). 56X magnification. The light-field print shows an upper limb joint in the centre of the photograph. The dark-field print shows that grain clustering is restricted to areas of tissue close to, but distinct from, the joint-forming surfaces of the developing bones. Expressing cells have the same appearance as those in Figure 5.8.
Figure 5.10 Light and dark-field photomicrographs of a sagittal section through a 16 day mouse embryo (hind limb region). 56X magnification. The light-field print shows the lower limb (foot) and silver grains are visible in this photograph, over the developing tarsals and metatarsals. The dark-field print (below) shows that grain clustering is restricted to localised areas within the developing bones. Expressing cells have the same appearance as those in Figures 5.8 and 5.9.
Other regions of the embryo which gave a positive signal with the contrapsin probe always contained cells of cartilaginous phenotype, and were located in areas where cartilage is normally present. No hybridisation was observed in regions lacking a defined skeletal component.

In the light of the EB22 cDNA library data, it seems likely that the transcript(s) detected by in situ hybridisation in the cartilage-like cells is one, or both, of the Spi-2 sequences found in the EB22 library. A proteinase inhibitor activity has previously been reported for cartilage (Kuettner et al., 1977). A role for proteinase inhibitors in controlling matrix degradation during tissue re-modelling can readily be envisaged, and this could represent either a second function for α1-antichymotrypsin in the mouse, or a function for the novel cys-cys serpin sequence, EB22-3. The latter is perhaps less likely since this gene is thought to have arisen after mouse/rat divergence and no other mammalian orthologues are known.

Screening for the Rat Arg-Ser Orthologue.

As discussed above, the presence of two very similar arg-ser inhibitor transcripts in mouse and Apodemus, and the existence of a conserved α1-antichymotrypsin equivalent in mouse, Apodemus, and rat, suggested that an arg-ser inhibitor sequence might be expressed in rat liver. Searching for this sequence in a rat liver cDNA library was complicated by the fact that the Spi-2.1 and 4.12B transcripts are expressed at high levels in this tissue, and while distinct from the putative rat arg-ser gene, are sufficiently similar to hybridise strongly to any cDNA probe available. The obscuring transcripts might be expressed fifty to one hundred-fold more abundantly than the target arg-ser sequence, as in mouse liver. This made the characterisation of randomly isolated Spi-2 sequences from rat liver appear laborious, and instead a differential screen was carried out.
Triplicate Benton-Davis filter replicas were made from a λ gt10 rat liver cDNA library plating of $5 \times 10^4$ plaques. Two filters were probed with the DBA arg-ser insert pMPCI-2 at a stringency of 2X SSC, allowing equivalent hybridisation to all contrapsin-like Spi-2 sequences. The third filter was probed with a mixture of two oligonucleotides specific for the reactive centre regions of the Spi-2.1 and 4.12B sequences (see Figure 3.12), and washed at stringency close to the predicted probe $T_m$. The majority (> 95%) of duplicate insert-positive plaques were also oligonucleotide positive, confirming the relative abundances of the three sequences. Six clones which were insert positive, but oligonucleotide negative were isolated. Restriction mapping to eliminate partial clones of the Spi-2.1 and 4.12B transcripts, yielded one clone which could not be either, and was not the Spi-2.2 rat transcript. This insert was 1.7kb long and was subcloned to pBluescribe as pRPCI-5. Sequencing using the conserved oligonucleotide p54/129+ revealed that the clone was likely to encode the direct rat orthologue of the DBA and Apodemus arg-ser sequences (Figure 5.11). This gene class appears to be part of the Spi-2.1 family, based on transcript size (18S), reactive centre specificity (anti-tryptic), and sequence similarity to the contrapsin and rat Spi-2.1 genes.

Two Conserved Spi-2 Functions in the Family Muridae.

Two distinct expressed Spi-2 subfamilies had now been isolated from mouse, Apodemus, and rat. The putative inhibitory specificity of the genes appears to have been conserved within each class, although in the Spi-2.2 subfamily the P1 residue is variable. This apparent functional stability may indicate that each class has a defined physiological role in vivo against a specific rodent proteinase. This is in contrast to the role envisaged for the contrapsin and rat Spi-2.1 proteins where selection by external (pathogenic) proteinases was postulated to be responsible for their accelerated evolution.
Figure 5.11 Sequence of rat PCI-like cDNA (exon 5). The nucleotide and predicted amino acid sequences are shown for exon 5. Below, the protein sequences for the rat PCI-like cDNA (rpci-5), the Apodemus Clone 1 cDNA (Apo 1), and the mouse pMPCI-2 (mpci-2) cDNA. Dashes represent identical residues. The position of the P1 residue is marked by an asterisk.
However comparative analysis of the sequences revealed several unusual features of the 18S Arg-Ser gene evolution. Examination of the reactive centre P1' serine in the three rodent species showed that the serine was encoded by a different codon (TCT) in mouse than the corresponding codon in Apodemus and rat (AGT). To evolve from one codon to the other inevitably involves transient loss of the serine residue in this position, if mutations occur as single base changes. The loss of serine at the P1' position would be expected to have serious effects on inhibitor function, although other residues can be tolerated in certain circumstances (Stephens et al., 1988). However, mutation to the alternative serine codon could theoretically take place in one step by a gene conversion event. Further analysis was carried out on the rodent cDNAs to try to establish an evolutionary descent for the 18S genes.

Using a sense oligonucleotide primer to exon 4 (p54/X4+, see Figure 3.2), sequencing was carried out on the mouse and Apodemus 20S cDNAs, pEB22/5 and pAAC-8, and on the mouse, Apodemus and rat 18S PCI-like cDNAs, pMPCI-2, pAP-5 and pRPCI-5. This provided sequence for 70% of exon 4 for these cDNAs, which are shown in figure 5.12 compared to the previously sequenced rat Spi-2.2 20S gene. An immediately striking observation was that a conserved serine codon in exon 4 showed the same codon type as the P1' residue in exon 5. For the rat PCI-like and Apodemus PCI-like sequences the codon was AGT, and for the other cDNAs, including mouse PCI, it was TCT.

The exon 4 and exon 5 sequences of the rodent cDNAs including the rat Spi-2.1, rat 4.12B and mouse contrapsin were used to generate phylogenetic trees using the PHYLIP computer program package. A number of equally parsimonious trees (not shown) were generated for the exon 4 sequences and for the exon 4 and 5 combined sequences, but only one for the exon 5 sequences alone. The trees showed clearly that the rodent
Spi-2.2 genes (mouse EB22/5, Apodemus clone 12 and rat Spi-2.2) form a clear subfamily, and also that the rat PCI-like sequence, the rat Spi-2.1 and 4.12B sequences, and the Apodemus PCI-like sequence form a tightly clustered family. The position of the mouse PCI-like and contrapsin genes was more variable and their evolutionary relationships are therefore more ambiguous, although closely related to one another. Summarising all the tree data, together with the serine codon distribution, a model was developed for the evolution of the 18S (Spi-2.1) and 20S (Spi-2.2) genes during rodent radiation (Figure 5.13.). It appears likely that the mouse PCI-like gene may have undergone gene conversion to the TCT serine codons after divergence from Apodemus. Contrapsin probably arose from the PCI-like gene after this conversion event, or may have been itself converted. Similarly, the rat Spi-2.1 and 4.12B genes appear to have arisen from the rat PCI like sequence after rat and mouse divergence. These two genes have retained the AGT serine codons of their putative parental sequence. The model is undoubtedly simplistic, as short tracts of some cDNAs show similarities to sequences from separate subfamilies (see Figure 5.12). Evidently some sequence conversion has occurred in the evolution of the present day genes. The question of how the primordial, pre-rodent PCI-like sequence developed two AGT serines in separate codons is difficult to answer, but the timescale for this process may have been considerably longer than the twenty million years since rat and mouse diverged.

The phylogenetic relationship of the genomic clone pcos2B1/p54λ-1-108 (Figure 4.15) to the other Spi-2 sequences indicates that it may represent the primordial Spi-gene (Figure 5.13). This gene may have surrendered its function following duplication, since the functional equivalent in rodents is probably the Spi-2.2 subfamily.
Figure 5.12 Nucleotide sequence of Exon 4 for 18S and 20S Rodent Spi-2 Genes. The sequences for six cDNAs are shown aligned below that of EB22/5, the mouse 20S α1-antichymotrypsin-like sequence. Dashes represent identity with the EB22/5 sequence.

Mac = EB22/5, Aac = Apodemus clone 12, Rac = rat Spi-2.2, Mpc = pMPCI-2, Apc = Apodemus clone 1, Rpc = pRPCI-5. The conserved serine codon mentioned in the text is boxed.
Figure 5.13 Proposed Evolution of the 18S and 20S Spi-2 Genes in Rodents. Exons 4 and 5 are shown for each extant gene, or putative precursor together with the actual or predicted serine codon state for each exon. Stages at which mammalian and rodent radiation occurred are shown by a heavy dashed line. The putative gene conversion event is shown by an arrowed narrow dashed line. PCI refers to an 18S anti-tryptic inhibitor, and 2.2 to a 20S antichymotryptic inhibitor.
Conclusions

Analysis of expression of cloned genomic Spi-2 sequences was undertaken to establish which, if any, encoded actively transcribed (functional?) genes. Attempts to show expression of the DBA gene p54λ-1-108 by Northern analysis and by screening of DBA liver cDNA libraries was unsuccessful. Expression of the 129 sequence pCos3E46 was demonstrated by cloning of an identical sequence from a DBA liver cDNA library. This sequence, pMPCI-2, resembles the related human PCI gene in the reactive centre region, but is clearly an Spi-2 gene. Northern analysis confirmed the Spi-2.1-like nature of the sequence, but gene-specific oligonucleotide analysis was unsuccessful.

An Apodemus liver cDNA library was screened for Spi-2 sequences, yielding an orthologue of the rat valine-serine inhibitor Spi-2.2, and an orthologue of the mouse PCI-like sequence pMPCI-2. These were expressed at similar, low, levels. No evidence for an Apodemus contrapsin was obtained.

A mouse EB22 (chondrocytic teratocarcinoma) cell line was screened nonspecifically for Spi-2 sequences. Two low abundance sequences were obtained. The transcript from the 129 genomic clone pCos2A2, a cys-cys reactive centre sequence, was cloned as λEB22-3. This is the first cys-cys serpin shown to be expressed in any tissue, and may challenge their apparent pseudogene status. The second EB22 gene was a direct orthologue of the rat and Apodemus Spi-2.2 anti-chymotrypsin like sequences, with the interesting difference of a third P1 residue, leucine, in the reactive centre. This gene was shown to be expressed in the mouse liver, and to be inducible by Dexamethasone or bacterial lipopolysaccharide. It is almost certainly the functional rodent equivalent of human anti-chymotrypsin.

In situ hybridisation of contrapsin antisense RNA to the 16-day mouse embryo supports the EB22 data, since expression was confined to chondrocytic tissue in the
developing animal. This may indicate an interesting and novel role for Spi-2 proteinase inhibitor genes in the control of skeletal tissue deposition.

The existence of a rat PCI-like orthologue, deduced from the mouse and Apodemus PCI-like sequences and the general rodent Spi-2.2 orthologue, was demonstrated by the use of differential probing to isolate the rat sequence pRPCI-5. Analysis of this sequence indicates that it may represent an example of the primordial rodent Spi-2.1 sequence from which the mouse contrapsin and rat Spi-2.1 sequences arose independently, following mouse/rat radiation.

Finally, a model was developed for the evolution of expressed Spi-2 sequences in rodents (Figure 5.13). Following rodent/primate divergence, the primordial α1-antichymotrypsin-like Spi-2 gene (perhaps pcosB2?) duplicated to give the Spi-2.1 and Spi-2.2 subfamilies in the rodent lineage. The Spi-2.1 subfamily acquired an anti-tryptic PCI-like specificity before rodent radiation, and the Spi-2.2 sequences retained their antichymotryptic specificity. Following rodent divergence, the subfamilies have undergone further duplications, and exhibited accelerated reactive centre evolution during the acquisition of novel inhibitory specificities.
CHAPTER SIX

Expression of Cloned Spi-2 cDNAs

Introduction

A considerable amount of time during the project was spent attempting to produce functional inhibitor protein from cloned Spi-2 sequences. This was because arguments for accelerated reactive centre evolution in the gene family require an anti-proteinase activity for positive selection to act upon. The existence of several serpins whose physiological role does not depend upon inhibition leaves open the formal possibility that the rodent Spi-2 sequences are essentially neutral in their reactive centre regions. An antitryptic activity has been reported for the contrapsin protein (Takahori and Sinohara, 1982) and possibly for the rat Spi-2.1-related sequence 4.12B (Le Cam et al., 1987). However it seems likely that the latter results were obtained with a mixture of the Spi-2.1 and 4.12B proteins (Tecce et al., 1986; Yoon et al., 1987), and that the activity may have been derived from either or both genes. No inhibitory activity had been reported for the only other rodent Spi-2 protein characterised at this time, rat Spi-2.2.

Several other serpin proteins have been successfully expressed recently in various prokaryotic (E. coli) and eukaryotic systems (yeast and mammalian cells). These include α1-Proteinase Inhibitor (Courtney et al., 1984; Jallat et al., 1986; Johansen et al., 1987), α2-antiplasmin (Holmes et al., 1987), Plasminogen activator inhibitor (Pannekoek et al., 1986), and human C1 inhibitor (Elder et al., 1988). These studies have shown that while many serpins are extensively glycosylated in vivo, unglycosylated forms from prokaryotes show essentially no decrease in inhibitory activity. Expression of cloned genes in E. coli often requires less sequence manipulation, and this was therefore
attempted first. Later efforts involved transient expression in mammalian COS cells, and *in vitro* translation of RNA produced from cloned cDNAs.

**Construction of a Complete Spi-2 Coding Sequence**

Initially, no full length cDNAs were available for use in expression studies, and various synthetically spliced constructs were produced from cDNA fragments. Efforts were concentrated on the rat *Spi-2.1* gene partly because fragments covering most of the coding region were available, and because demonstration of contrapsin-like antitryptic activity would strengthen the argument for evolutionary orthology of the two genes. The strategy used to produce a full-length coding region from the fragments Dex1, Dex2, and a pair of synthetic oligonucleotides is shown in Figure 6.1. The final product, a DNA of 1.2 kilobases, was cloned into the plasmid pSP64 and renamed pSPFL2.1

**Attempts to Express Inhibitor Protein in E. coli**

The cDNA from pSPFL2.1 was inserted into two types of bacterial expression vector; ptac85 which utilises the *tac* fusion promoter for gene expression (Marsh, 1986) and the pJLA plasmid series which employ a tandem $P_L/P_R$ promoter system (Schauder *et al*., 1987). The *tac* is a hybrid of the *trp* and *lac* bacterial promoters, and more closely approaches the consensus promoter sequence than either donor sequence (de Boer *et al*., 1983). As a result, it is a more efficient promoter than either *trp* or *lac*, but is efficiently suppressed by the *lacI* repressor protein in the absence of an inducing molecule such as $\beta$-galactosidase or the synthetic analogue IPTG. This promoter repression prevents expression of foreign sequences in the ptac85 plasmid which carries the *lacI* gene to ensure that no "leaky" expression occurs, even with a high plasmid copy number.
Figure 6.1 Strategy for construction of full-length Spi-2.1 coding sequence. The extent of the pDex1 and pDex2 clones are shown below the predicted complete Spi-2.1 transcript, expected to be extremely similar to the 4.12B cDNA (Tecce et al., 1986). Coding regions of the transcript are stippled rectangles, untranslated regions are open rectangles, and the reactive centre region (RC) is shown as a dark rectangle. Inserts are shown as dark bars in the circular plasmids and the orientation of promoters by thin, right angled arrows.
The pJLA expression vector series is an extremely efficient system for producing high (up to 30%) levels of heterologous protein in *E. coli* (Schauder *et al.*, 1987). The pJLA plasmid contains the bacteriophage $\lambda P_L$ and $P_R$ promoters in tandem, giving enormous levels of transcription, and also expresses the heat-sensitive C1257 mutant of the $\lambda$ repressor protein ensuring that bacterial growth at 30°C occurs without gene expression. Raising the culture temperature to 42°C destroys the mutant repressor and permits the expression of cloned sequences, at high levels.

Unfortunately, neither vector system ever produced detectable amounts of recombinant protein. Detection was hampered by a number of technical problems. No antibody was available to detect *Spi-2.1* expression and, in the absence of a established inhibitory assay, Coomassie staining of cell lysates to detect *de novo* expression will not reveal protein levels below about 1% of total cell protein. Since *E. coli* possesses major protein bands at the same mobility (about 43 kD) as the expected inhibitor, this made detection by eye virtually impossible. Pulse labelling of heat-induced pJLA/Spi-2.1 containing clones with $^{35}$S-methionine for two minutes showed no evidence of increased expression in proteins of the expected size range (Figure 6.2). Attempts to detect inhibitory activity in bacterial lysates using standard methods gave negative results. Either the rat *Spi-2.1* gene lacks inhibitory activity for the proteases tested (trypsin, thrombin, plasmin, chymotrypsin and elastase), or the constructs were deficient, perhaps in transcription, or post-translational processing of inhibitor protein. The protease-deficient *E. coli* strain CAG 629 was used to counteract possible inhibitor degradation, but did not produce detectable levels of inhibitor.
Figure 6.2 Pulsed-labelling of *E.coli* proteins with $^{35}$S-Methionine. Lane 1: pJLF20 uninduced (growth at 30°C), 2: pJLF20 induced (growth at 42°C for 10 minutes), 3: pJLF19 uninduced, 4: pJLF19 induced, 5: pJLF21 uninduced, 6: pJLF21 induced. The three plasmids represent independent isolates of the putative full-length *Spi-2.1* spliced construct. No induction of protein in the 45kD range is observed for any construct. The position of molecular weight markers in kiloDaltons are shown on the left.
The vectors used have been shown to yield high levels of expression (5% - 30% of total cell protein) under optimal conditions, with appropriate eukaryotic sequences (Schauder et al., 1987). However it is difficult to predict in advance the suitability of a particular sequence and dramatic increases in expression can be obtained by deletion or mutation of short (15-30 base pairs) regions at the 5' end of the inserted DNA (Johansen et al., 1987). This is not always due to a simple disruption of a stable secondary mRNA structure.

**Inhibitor Expression by in vitro Translation**

To determine whether the pSPFL2.1 gene contained an open reading frame capable of encoding an inhibitor protein, the pSPFL2.1 plasmid was linearised with Bam HI to produce a template for SP6 RNA polymerase (Krieg and Melton, 1984). RNA was transcribed from the 5' SP6 promoter using the di-nucleotide analogue pGpppGTP to generate a "cap" structure at the 5' end of the in vitro message (Contreras et al., 1982). The RNA was then translated using a rabbit reticulocyte lysate system and 35S-methionine to produce radiolabelled inhibitor protein. This was electrophoresed on an SDS/Polyacrylamide gel, and proteins visualised by autoradiography. An example of a gel is shown in Figure 6.3. This shows that a single polypeptide of approximately 43 kD was translated from the pSPFL2.1 encoded RNA. This was of the size expected for a full-length inhibitor, and indicated that the Spi-2.1 construct contained an open reading frame of the correct size. Incubation of radiolabelled protein with the proteinases trypsin, thrombin, plasmin, chymotrypsin and elastase failed to show interaction of inhibitor and proteinase. This may have been due to "swamping" of the proteinase by reticulocyte lysate proteins, and poor interaction between proteinase and inhibitor. Even at high proteinase concentrations, the inhibitor band was visible, although reduced in concentration (not shown).
Figure 6.3 Expression of \textit{Spi-2.1} protein in a reticulocyte lysate system. Lane 1: No mRNA (control), 2: \textit{Spi-2.1} mRNA - degraded pGpppG, 3: \textit{Spi-2.1} mRNA + degraded pGpppG, 4: \textit{Spi-2.1} mRNA - undegraded pGpppG, 5: \textit{Spi-2.1} mRNA + undegraded pGpppG, 6: degraded \textit{Spi-2.1} mRNA + undegraded pGpppG. \textit{Spi-2.1} protein can be seen in lanes 2, 4 and 5. The position of molecular weight standards in kiloDaltons are shown on the left. The inhibitory effect of degraded cap analogue upon yield of active RNA transcripts is obvious.
A second inhibitor construct was produced at this stage. The 3' coding sequence of the human α1-antichymotrypsin gene (Hill et al., 1984) was fused to the 5' pSPFL2.1 coding sequence at the unique Dra I restriction site common to both sequences. This chimaera would be expected to encode an inhibitor with the anti-chymotryptic specificities of the human gene if it was active. To establish that this construct, pCHEX, contained an open reading frame, it was translated in the reticulocyte lysate system used for pSPFL-2.1. This also produced a single protein band of similar mobility to the Spis-2.1 protein (not shown). Again, interaction with the proteinases mentioned above could not be demonstrated.

A final attempt to produce inhibitor protein from the reticulocyte lysate system was carried out using the Apodemus full-length cDNA clone pAP3. This clone encodes the Apodemus PCI-like protein described in chapter 5, but in the opposite orientation, to enable T7 RNA polymerase transcription of insert sequences. It was considered that construction of the pSPFL-2.1 and pCHEX clones might have introduced structural errors into the sequences, and that these might be avoided by using an intact unmodified inhibitor cDNA. Protein of the correct size was produced from in vitro translation of pAP3-encoded RNA, but incubation with the serine proteinase bovine Kallikrein did not show appreciable modification of inhibitor, even at high proteinase concentrations (not shown). The in vivo target for this inhibitor is unknown, but it was felt that kallikrein might represent a suitable target proteinase in terms of specificity. Unfortunately, trypsin, thrombin and plasmin which also demonstrate specificity for an arginine P1 residue were not tested with this inhibitor.

Attempts to Express Inhibitors in Mammalian Cells

The cDNA inserts from pSPFL2.1 and pCHEX were cloned into the eukaryotic expression vector pSVL (Pharmacia) and transfected into a transformed African Green
Monkey kidney cell line (COS cells). This cell line constitutively expresses large amounts of SV40 T antigen, and will express heterologous sequences linked to the SV40 late promoter at high levels. It was felt that the proteins expressed by reticulocyte lysate might lack a functionally important post-translational modification, although successful expression of active serpins in *E. coli* would argue against this. COS cells were transfected with pSVL/inhibitor constructs by the DEAE-Dextran and calcium phosphate techniques (McCutchan and Pagano, 1968; Graham and van der Eb, 1983) and allowed to recover in serum-free supplemented medium for 72 hours. Aliquots of cell extract, and of serum-free medium were then tested for the presence of inhibitory activity for the proteinases trypsin and chymotrypsin. All proved negative compared to control vector-only extracts. These experiments were not repeated, and were unlikely to have been optimal for inhibitor expression. Antisera to human α1-antichymotrypsin (from SAPU) were used to probe a Western blot of pCHEX COS cell extracts and medium (assistance from P. Budd), since this sequence contains the 3' coding sequences from the human gene. No specific signal was obtained however. This may have been due to the fact that the construct was not expressed in COS cells, or because the rat/human gene fusion has destroyed the specific epitopes recognised by the human specific antisera.

**Conclusions**

Attempts were made to express cloned Spi-2 inhibitors in a variety of systems, and to test the inhibitory specificity of proteins produced. Initially, no full-length clones were available and extensive fragment manipulation was required to produce suitable sequences. These were tested in several *E.coli* systems but expression was never detected. Translation of RNA produced *in vitro* by a reticulocyte lysate system did produce polypeptides of the expected size, but these did not exhibit interactions with the common proteinase types tested. Transfection of constructs into mammalian COS cells
did not give expression detectable by inhibitory assay, or by Western blotting with antibody.
CHAPTER SEVEN

Discussion

This thesis describes the analysis of the evolution of an Spi-2 gene, contrapsin, in the genus Mus and the structure and expression of the Spi-2 locus in the inbred laboratory mouse. The major findings can be summarised as follows. The accelerated evolution in the reactive centre region of the contrapsin gene appears to have occurred over a comparatively short period of time, followed by stasis when the inhibitor acquired a useful specificity. Consideration of closely related rat sequences indicate that the mode of divergence may be via point mutation, or very restricted gene conversion events, in the reactive centre region. Analysis of genomic and additional expressed sequences indicates that more extensive sequence conversion may have occurred in some genes, obscuring phylogeny to some extent. Certain inhibitory specificities appear to have been retained since before rodent radiation, while other novel species-specific reactive centre specificities have arisen following recent gene duplications.

The evolution of contrapsin

The starting point for the project was the observation by Hill and Hastie (1987) that the reactive centre regions of cloned rodent Spi-2 sequences were behaving in a highly unusual fashion. The precise relationship of the genes to each other was not fundamentally important in this analysis, since the sequence divergence observed was so clearly localised to Region 2 (reactive centre) and to a lesser extent Region 3. However several lines of evidence indicated that the mouse contrapsin and rat Spi-2.1 were directly orthologous, including shared transcript size, putative inhibitory specificity, and similar levels of abundance in the rodent liver. In spite of this, their reactive centre
sequences were only aligned with difficulty and showed an identity of 33% for Region 2 (compared to 84% for Region 1 and 75% for Region 3). The analysis of the contrapsin reactive centre sequence in the mice species closely related to the *M. domesticus/M. musculus* derived inbred mouse strains was expected to show a reduced but still detectable degree of divergence. The virtually complete fixation of the sequence (Figure 3.5) in the ancestor of *Mus domesticus* and *Mus caroli* was not expected, especially as Region 3 is continuing to evolve at a comparatively rapid rate relative to albumin. Contrapsin appears to have acquired a function soon after its appearance as a highly expressed liver gene (Figure 3.8), and a specificity which has remained unaltered over the last six to eight million years.

The above arguments have assumed that contrapsin actually has an inhibitory role, but serpins can adopt other functions such as hormone transport (Hammond *et al.*, 1987) or undergo proteinase cleavage to yield vaso-active or chemoattractant peptides (Banda *et al.*, 1988; Hoffman *et al.*, 1989). A non-inhibitory function might be less likely to select for an unchanging reactive centre sequence, and would not explain the stabilisation of the contrapsin reactive centre relative to Region 3. Conceivably, a non-inhibitory but anti-parasitic function for contrapsin could still depend upon a specific "inhibitor"-proteinase interaction, in which case arguments for accelerated evolution would remain valid. There is indirect evidence that contrapsin may have an involvement in the mouse host response to parasitic infection, since schistosomes isolated from infected mice are associated with antigen which crossreacts with contrapsin-specific antibodies (Mohda *et al.*, 1988). In addition, contrapsin with altered electrophoretic properties (presumably cleaved) is found in the serum of infected mice. This may well indicate a host-inhibitor/parasitic-proteinase interaction, but this remains unproven since schistosome flukes are known to incorporate host proteins into their membranes, presumably to evade
the hosts immune response (McClaren and Terry, 1982; Simpson et al., 1983). It is possible, although highly unlikely, that contrapsin actually facilitates parasitic infection in this manner.

The discovery of a well conserved Spi-2 function, the 18S PCI-like gene, in mouse, Apodemus and rat during the project raises the question of whether contrapsin and rat Spi-2.1 are true orthologues, or whether both sequences arose independently from the common parental PCI-like sequence after rat/mouse divergence. Certainly the intra-species similarities are greater for the two gene pairs, but this might be due to sequence conversion events rather than a reflection of divergence time. The fact that Apodemus does not express an abundant 18S anti-tryptic type transcript also supports the idea of independent acquisition of the contrapsin/Spi-2.1 function. (This raises the additional question of why Apodemus does not require the contrapsin function). If it is true that the abundant rat and mouse liver Spi-2 sequences represent independently derived functions from the PCI-like gene, then the accelerated reactive centre divergence noted by Hill and Hastie actually occurred in a considerably shorter period than they assumed in their original calculation. Taking into account the stasis of the contrapsin reactive centre for the last six to eight million years, this considerably increases the already rapid rate of contrapsin reactive centre centre divergence

The contrapsin sequence analysis failed to indicate the nature of mutations in the reactive centre region since so few were discovered. Comparison of the Mus domesticus and Mus caroli contrapsin sequences did reveal a high ratio (4:1) of nonsynonymous to synonymous substitutions in Region 3. As observed in Chapter 3, rapid divergence of the carboxyl terminus (Region 3) seems to be a general feature of serum serpins even in the absence of reactive centre divergence. The nature of reactive centre divergence is shown more clearly by comparison of the rat Spi-2.1 and 4.12B genes (Figure 3.13).
These sequences can be aligned with more confidence than contrapsin and rat Spi-2.1 and show that the mechanism of divergence appears to involve multiple nonsynonymous point mutations, although some clustering of substitutions is apparent. Although the two genes are very closely related in Region 1, and would certainly seem to have arisen from a gene duplication event after mouse/rat divergence, Region 2 and Region 3 display a large number of amino acid substitutions, and a substantial (9 basepair) deletion. It is possible that one of the genes has become inactivated, although at least one of the two encodes an anti-tryptic inhibitor (Le Cam et al., 1987). This activity was reported for the 4.12B encoded protein, but the purification strategy probably co-isolated both rat proteins. Certainly the 4.12B sequence resembles more closely an Spi-2 consensus with regard to conserved residues and reactive centre length (Figure 3.13).

The Spi-2 locus in the laboratory mouse

The isolation and physical linking of ten Spi-2 sequences, and sequencing of nine, from the mouse genome would be expected to reveal much of the locus complexity and evolution of this gene family. This is true to an extent, but the analysis leaves several questions about the precise mode of amplification unanswered. Two genomic sequences were later found to have expressed transcripts, isolated from cDNA libraries, but again two cDNAs exist for which no genomic counterpart was cloned (contrapsin and EB22/5). The evolutionary range of Spi-2 sequences cloned was wide enough to include either sequence, and their absence may be due to under-representation in the libraries analysed rather than radically different genomic structure or isolation from the rest of the family.

The amplification of the Spi-2 locus certainly seems to be of recent origin in the mouse. The relatedness of the "cys-cys cluster" of genes at one end of contig 1 is pronounced, but is not due to clear-cut gene conversion events. The distinct lack of repeated
sequences detectable by hybridisation may also indicate recent amplification. Unfortunately, no simple mechanism for the amplification which can account for gene orientation, sequence relationships and pattern of gene expression was established. A definitive description of the amplification process will require the isolation of the remaining Spi-2 genomic sequences (at least two), further sequencing analysis of different gene regions, and probably analysis of the locus in related species such as Mus spretus, Mus caroli and Apodemus sylvaticus.

Examination of reactive centre sequences from the genomic clones indicates that divergence is occurring primarily in the reactive centre region and, to a lesser extent, Region 3 (Figure 4.13b). This process is probably not a result of the intron/exon structure of the genes, and does not appear to be due to whole-exon sequence conversion events, given the diversity of the reactive centres. The main conclusion from the genomic analysis is that the putative functional domain in this multigene family is hypervariable (Figure 4.13), and that evolution has generated a range of expressed and potentially expressed novel inhibitory specificities.

One possible model is that all of the genes are expressed, and are under selection for retention of inhibitor structure motifs, as well as for diverse reactive centre specificities. Expression was only established for two genomic sequences (pcos3E46 and pcos2A2), and in the case of the latter sequence was so low that the presence of other Spi-2 sequences at similar levels might go undetected. However, exhaustive attempts to isolate the p54±-1-108/pcos2B1 leu-ser sequence from the liver failed. Three of the 129/St strain genomic sequences (pcos1A1, pcos6A1, and especially pcos3E2) seem highly likely to represent pseudogenes since they have unfavorable mutations in the conserved reactive centre-flanking domains. However, no termination codons, or frameshift mutations were detected in either exon four or exon five of any Spi-2 sequence during
Perhaps a more likely scenario is that only a subset of the genes are active (or at least under selection). Their presence in the Spi-2 cluster, and selection for retention of the critical conserved structural motifs flanking their reactive centres could in turn select for homogenisation of the non-expressed sequences to a functional structure sequence in these domains. A possible example of this type of localised conversion is the presence of a lysine P16 codon in Region 1 of both Apodemus cDNAs (Figure 5.5), as this residue is threonine in all other Spi-2 sequences bar one. Reactive centre regions do not appear to be homogenised in the same fashion. This suggests a potential analogy with the storage of the chicken λ-light chain immunoglobulin repertoire in a pseudogene pool (Reynaud et al., 1989), although Spi-2 diversification would occur at a far lower frequency, and in the germline rather than as a somatic process. A combination of point mutations and restricted directional sequence conversion events could develop a pool of novel reactive centre specificities in silent genes. These could then transfer into expressed genes at low frequencies by conversion events, to await selection by proteinases with novel specificities. The episodic nature of parasitic invasion of a species might even favour low-level expression of "silent" genes as they accumulated novel reactive centre mutations. There would be some degree of endogenous selection against unbridled development of new inhibitory specificities in expressed genes, as these could interfere with important endogenous proteinase-mediated cascades. The α1-antitrypsin Pittsburgh mutation described in Chapter 1 is an example of this type of selection.

Expression of Spi-2 genes in the laboratory mouse

It seems clear that the two Spi-2 transcript types isolated from mouse, Apodemus and rat represent genes which were present before rodent radiation, and which have relatively stable endogenous functions. These are the putative α1-antichymotrypsin (or Spi-
2.2) orthologue with a 20S transcript, and the PCI-like (or Spi-2.1) anti-tryptic 18S transcript. Both classes have undergone different evolutionary events since rodent radiation, despite probable retention of their original function(s).

The Spi-2.2 genes are good candidates for the role of functional human α1-antichymotrypsin orthologue for a number of reasons. In the mouse and rat at least, they are inducible by inflammatory stimuli such as bacterial lipopolysaccharide treatment, and also by dexamethasone. They are expressed in the liver at fairly low constitutive levels in all three rodents examined. The reactive centre of the mouse gene, EB22/5, is conserved in Mus cervicolor which diverged six to eight million years ago (Figure 5.4), a similar degree of stabilisation to that of contrapsin. The inhibitory specificity appears to have been relatively conserved throughout murid radiation, although the P1 is different in mouse, Apodemus and rat. The Apodemus (leucine) and rat (valine) P1 residues are conservative differences and should cause very little difference in specificity, as the human gene also has a leucine P1 residue. The mouse P1 residue, methionine, is also compatible for a chymotrypsin-type inhibitor (Potempa et al., 1988), and could have arisen from the primordial (leucine?) codon by a single point mutation. It was an unexpected P1 residue, since the considerably more abundant α1-PI serum inhibitor (P1 methionine) might be expected to perturb the physiological function of the EB22/5 encoded protein. However, inhibitory specificity is not entirely dependent upon the identity of the P1 residue, and the two sequences display little identity in flanking reactive centre sequences. The remainder of the exon 5 amino acid sequence for the Spi-2.2 gene in the three rodent species shows a high degree of conservation (Figure 5.6b).

The role of the Spi-2.2 anti-chymotryptic inhibitors is unknown, as is that of their proposed functional human orthologue α1-antichymotrypsin. The target proteinases for this inhibitor are thought to be "mast cell chymase" and neutrophil cathepsin G, both
proteinases with a nebulous physiological role at present (Wintroub et al., 1981; Reilly et al., 1982). They are likely to have a role in response to parasitic infection or inflammation, given their sites of expression. A function in the control of the inflammatory response is also likely for the inhibitor since its levels rise rapidly in response to physiological insult (Fey and Fuller, 1987). Cathepsin G may have a role in tissue remodelling since it can degrade connective tissue proteins such as collagen and proteoglycans (Roughley and Barrett, 1977). This could account for the originally discovered site of EB22/5 expression, in chondrocytic-type tissue in the 16 day mouse foetus (Figures 5.8 to 5.10). Deposition of skeletal tissue during development might depend upon inactivation of proteinase activity while ossification occurs at bone and joint margins. This is an attractive, if extremely simplistic, scenario.

The second Spi-2 sequence isolated from the EB22 cell line, EB22/3, had previously been observed as a genomic sequence in both DBA and 129 inbred mice. It contains the unusual cysteine-cysteine P1-P1' sequence, and had been considered likely to be a pseudogene until its isolation as a full-length cDNA. The EB22/3 gene clearly falls into the Spi-2.2 subfamily in terms of transcript size (20S), and sequence relatedness to the (primordial?) Spi-2.2 anti-chymotryptic sequence, EB22/5. It is one of the most recently duplicated Spi-2 sequences, in the "cys-cys cluster" at the end of Contig 1 (Figure 4.12 and 4.15). The level of expression in the EB22 cell line is similar to that of the EB22/5 gene, at about 0.002% of independent cDNA clones. The nature of the activity encoded by the EB22/3 sequence is not known. Considerable selective pressure seems to exist for a serine residue at the P1' position in the majority of inhibitory serpins, although α2-antiplasmin contains a methionine P1' residue (Holmes et al, 1987). A deleterious condition involving substitution of leucine for serine at the P1 residue has been described, for antithrombin Denver (Stephens et al, 1987). A study employing
site-directed mutagenesis to exchange the antithrombin P1' serine residue for alternative amino acids indicated that a cysteine in this position reduced inhibitory activity by two orders of magnitude (Stephens et al., 1988). This might indicate that the EB22/3 sequence is not a functional inhibitor. However the same study demonstrated complete abolition of antithrombin activity when methionine was substituted for serine, yet methionine occurs in the α2-antiplasmin P1' position. Definite evidence of the inhibitory capacity of the cys-cys EB22/3 sequence will require expression and testing of the protein, perhaps in vitro.

The second Spi-2 gene found in all three rodent species was the PCI-like 18S anti-tryptic transcript. This sequence has undergone an entirely different type of evolution following rodent divergence. In contrast to the Spi-2.2 antichymotryptic class, the P1 residue (arginine) has remained constant, but more extensive differential sequence conversion appears to have occurred in the reactive centre and Region 1 (Figures 5.12 and 5.13). It seems likely that this gene gave rise to the abundantly expressed rat Spi-2.1 and 4.12B genes and the mouse contrapsin gene independently after the two species diverged. As in the case of the Spi-2.2 genes, the function of the gene is unknown. The similarity of the mouse reactive centre sequence to human protein C inhibitor (Figure 5.1) may be coincidence, or indicate the actual function of this subfamily. Sequencing data indicates that the rodent genes were derived from the primordial α1-antichymotrypsin rather than PCI lineage and, unless an unusual interfamily sequence conversion has occurred, this could conceivably represent a case of convergent evolution.

Future directions

This project has unavoidably left several questions about the evolution and expression of the Spi-2 locus unanwered. Approaches to answering some of these will be briefly discussed below.
Determination of the degree of polymorphism of the contrapsin reactive centre within a species is a problem that might be relatively simply addressed by RNA sequencing or a PCR sequencing approach. This was not undertaken due to lack of time, and the absence of a suitable mouse population. The demonstration of polymorphism within a species could indicate that a number of alleles of any one inhibitor gene existed in a species to provide a range of specificities against parasite proteinases. Frequency dependent selection would ensure that none of these alleles were likely to be fixed, and their continuing presence in the gene pool would permit reactive centre targeted gene conversion events to generate novel specificities. This scenario has been proposed for the MHC (Howard, 1988). However the remarkable similarity of all contrapsin reactive centres sequenced, in inbred and wild mice as divergent as *Mus domesticus* and *Mus caroli*, may indicate that the sequence would show very little intraspecific polymorphism.

If the contrapsin gene is truly restricted to mice which diverged from a common post-Apodemus ancestor around eight million years ago, analysis of a second expressed Spi-2 gene might extend the analysis throughout Mus. Either the Spi-2.2 or PCI-like sequences should be suitable for a cDNA/PCR based approach. The uncertain nature of the PCI-like gene evolution since rodent divergence would make this a doubly interesting analysis.

The precise association between the Spi-1 and Spi-2 loci is of interest given their close linkage in man (P. Kearney, personal communication) and the fact that at least some Spi-2 sequences map to the Spi-1 rather than the Spi-2 locus in the BXD recombinant inbred strain series (Figure 4.6). This might be most conveniently addressed by pulsed field analysis, using locus-specific probes. Tight linkage of the complexes might increase the likelihood that Spi-1/Spi-2 hybrids could arise at the interface, allowing novel patterns of expression and novel specificities to flow between the two gene fami-
lies. The internal structure of the two complexes appears to differ, since the Spi-1 locus appears to contain fewer distinct sequences (Hill et al., 1985). Analysis of cosmids containing Spi-1 sequences during the present work indicated that the mouse Spi-1 genes are more widely separated than their Spi-2 counterparts, with a maximum of one per cosmid.

The relation of the human protein C inhibitor (PCI) to its putative rodent orthologue would also be of interest in determining the history of the Spi-2 locus. This sequence is more closely related to α1-antichymotrypsin (Spi-2) than α1-proteinase inhibitor (Spi-1), but probably diverged from α1-antichymotrypsin before rodent radiation (Suzuki et al., 1987). Whether it remains as a distinct lineage in rodents is unknown. The mouse Spi-2 sequence pcos3E-46 shows striking similarities to human PCI in the reactive centre region (Figure 5.1). The question of whether the human PCI sequence has become subsumed into the Spi-2 lineage, donating its reactive centre but losing its separate identity, might be answered by analysis of the mouse genomic clones with the human PCI probe. If the rodent equivalents of the human PCI and α1-antichymotrypsin genes have intermingled to form a hybrid Spi-2 locus, unravelling the evolutionary origins of any one gene may prove impossible.

To determine the recent evolutionary history of the Spi-2 locus, an analysis of the genomic structure and arrangement of Spi-2 sequences in species such as Mus spretus and Mus caroli could be carried out. This might establish the relative importance of gene conversion in the evolution of the locus, and could shed light on possible intermediates in the amplification process.

Determining the anti-proteinase specificities of the characterised inhibitors will be an essential next step in the analysis of the Spi-2 locus evolution, since the reactive centre evolution arguments depend on inhibitor function. Chapter six described attempts to express and characterise active inhibitors, largely in vain. Determining the in vivo pro-
teinase target for the cloned inhibitors may not be possible until more is known about the range of serine proteinases in the mouse. The existence of the kallikrein multigene family on mouse chromosome seven is interesting in this respect. The kallikreins are serine proteinases with highly specific cleavage sites, thought to be involved in the processing of bioactive peptides from protein precursors (Mason et al., 1983). The family has become amplified in the mouse by a series of tandem duplications and about twelve (of twenty four) sequences in the locus appear capable of encoding active proteins (Evans et al., 1987; Drinkwater et al., 1987). They show hypervariability in residues flanking the catalytic cleft, and it is tempting to speculate that they might be co-evolving with the Spi-2 locus (Creighton and Darby, 1989).

A collaboration with H. Miller (Moredun Institute, Edinburgh) has been established to determine whether any cloned rodent Spi-2 sequences represent functional inhibitors of the rat mast cell chymase II. This proteinase is induced to extremely high levels during infection by Nippostrongylus sp. nematodes and has been shown to form an SDS-resistant complex with a component of rat or mouse serum (H. Miller, personal communication). Given the cleavage specificity of the chymase (hydrophobic P1 residue), and the suggested association of human α1-antichymotrypsin with human mast cell chymase, the conserved rodent Spi-2.2 protein is an attractive candidate inhibitor.

Finally, the use of in situ hybridisation to locate expression of specific Spi-2 transcripts in different tissues at different stages of development would be a powerful method for continued analysis of the family. The preliminary work described in Chapter Five revealed the unexpected expression of inhibitors in foetal cartilage, using a non-specific probe. It is possible to use sequence-specific oligonucleotide probes to detect transcripts at low abundance with the in situ hybridisation technique. This could be undertaken with a range of adult tissue, as well as stages of foetal development, and
could be used to detect changes of *Spi-2* expression in rodents exposed to parasitic and inflammatory stimuli.
ABBREVIATIONS

bp......basepairs

cDNA......complementary deoxyribonucleic acid

ddH$_2$O......double-distilled deionised water

dNTP.....deoxyribonucleotide triphosphate

ds.......double stranded

EDTA......ethylenediaminetetra-acetic acid

EtOH.......ethanol

kb.........kilobasepair

ml.........millilitre

µg.........microgram

µl.........microlitre

mRNA......messenger ribonucleic acid

MW.........molecular weight

mya, m.y.a.....million years ago

ng.........nanogram

oligo......oligodeoxynucleotide

RM.........reaction mixture

RT.........reverse transcriptase

SDS.........sodium dodecyl sulphate
REFERENCES


Bennett K. L., and Hastie N. D. (1984) Looking for relationships between the most repeated dispersed DNA sequences in the mouse: small R elements are found associated
consistently with long MIF repeats. *EMBO J.* 3, 467-472.


Hedin S. G. (1903) Investigations on the proteolytic enzymes of the spleen of the ox. J. Physiol. 30, 155-175.


terminase. *Gene* 40, 259-266.


Simpson A. J. G., Singer D., McCutchan T. F., Sacks D. L., and Sher A. (1983) Evidence that schistosome MHC antigens are not synthesized by the parasite but are acquired from the host as intact glycoproteins. *J. Immunol.* **131**, 962-.


Wright H. T. (1984) Ovalbumin is an elastase substrate. *J. Biol. Chem.* 259, 14335-