The EMBO Journal vol.10 no.2 pp.255-261, 1991

The murine Spi-2 proteinase inhibitor locus: a multigene family with a hypervariable reactive site domain

John D.Inglis and Robert E.Hill

MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK

Communicated by N.D.Hastie

We have isolated 10 closely linked members of a proteinase inhibitor multigene family from the inbred mouse strain 129. These sequences, termed the Serine Proteinase Inhibitor 2 (*Spi-2*) genes, appear to have been derived from a common ancestor represented in man by the single copy α1-antichymotrypsin gene. The genes are clustered on two cloned genomic DNA segments spanning 220 kb, and have at least partially retained the intragenic structure of the ancestral *Spi-2* gene. Sequence analysis from the final coding exon indicates that most of the mouse genes may be competent to encode functional proteins, some with a predictable inhibitory spectrum, and several representing novel inhibitor types. An oligonucleotide probe designed to one reactive centre sequence enabled the isolation of the cognate expressed transcript from a liver cDNA library. However, whether expressed or not, the reactive centre regions of all the sequences have diverged at a rapid rate relative to structurally defined flanking sequences. The divergence is also appreciably greater than that occurring in an adjacent non-coding sequence. This phenomenon has established novel potential inhibitory specificities, while maintaining a functional inhibitor structure.

Key words: molecular evolution/reactive centre/serpin

Introduction

The serine proteinase inhibitors (serpins) are a protein superfamily which have diverged from a common ancestor about 600 million years ago (Hunt and Dayhoff, 1980; Carrell *et al.*, 1987). The group includes active serine proteinase inhibitors, as well as proteins with other biological roles but no known inhibitory activity such as angiotensinogen (Ohkhubo *et al.*, 1983) and thyroxine binding globulin (TBG; Flink *et al.*, 1986). Members of the family with inhibitory properties play an important role in the control of many mammalian proteolysis-dependent events, such as the blood coagulation cascade, the complement cascade and the inflammatory response (Travis and Salvesen, 1983).

The mechanism of proteinase inhibition has not been fully characterized, but involves proteolytic attack upon a defined region of the inhibitor, known as the reactive centre (Travis and Salvesen, 1983). A single amino acid, termed the P1 residue, within this region is primarily responsible for the proteinase specificity of an inhibitor. Naturally occurring or deliberate mutation of the P1 residue may cause dramatic changes in the spectrum of proteinases inhibited by a particular serpin (Owen *et al.*, 1983; Courtney *et al.*, 1985; Jallat *et al.*, 1986). However, it is clear that residues surrounding the P1 position are also involved in proteinase–inhibitor interactions, affecting both the specificity and efficiency of inhibition (Jallat *et al.*, 1986; Stephens *et al.*, 1988; Carrell *et al.*, 1989).

The mouse serpin contrapsin, which is abundantly expressed in the liver, shows 70% nucleotide, and 60% amino acid identity to its closest human counterpart, α1-antichymotrypsin (Hill *et al.*, 1984). However, in the reactive centre region the two sequences are highly divergent, with only one residue in the 10 amino acid region surrounding the P1 conserved between the two, and a leucine to lysine substitution at the P1 position in the mouse. This particular difference is consistent with the anti-tryptic specificity reported for the contrapsin protein (Takahori and Sinohara, 1982). Further investigation of contrapsin-related genes in rodents (Hill *et al.*, 1985) demonstrated that contrapsin was a member of a clustered multigene family, designated the *Spi-2* complex, which maps to chromosome 12 in the mouse.

Subsequent cloning and sequencing of two rat *Spi-2* cDNAs demonstrated that the rapid reactive centre divergence observed for the contrapsin and human α1-antichymotrypsin genes was a recurring feature of the gene family (Hill and Hastie, 1987). A comparative analysis of the four sequences demonstrated that the reactive centre regions of the encoded proteins have diverged at an accelerated rate, and that the rate of nucleotide substitution in the reactive centre region exceeds that of pseudogenes and introns from other genes (Hill and Hastie, 1987).

The *Spi-2* multigene family in mouse provides the potential to encode a range of different inhibitor proteins. Due to the relatively simple relationship between structure and inhibitory specificity, it is possible to analyse primary gene sequences to predict the range of possible activities within the *Spi-2* family. We have undertaken an analysis of the gene family, and isolated 10 *Spi-2*-related sequences as cosmid clones. We have found that all 10 genes differ from one another in the reactive centre region, despite strong similarities in flanking regions of the gene.

Results

Isolation of genomic Spi-2 sequences

Screening of a cosmid genomic library from the inbred mouse strain 129/St was carried out using a rat *Spi-2* cDNA, pSPFL2.1, as a probe for *Spi-2* specific sequences. A total of 5×10^5 colonies were screened, yielding 13 distinct *Spi-2* positive clones. Individual clones were mapped with the restriction enzymes *Bam*HI, *Eco*RI, *Hin*dIII, *Xho*I and *Sal*I, employing total and partal restriction digest strategies. The location and transcriptional orientation of *Spi-2* sequences within the cosmids was determined using a 5′ specific subclone of rat *Spi-2.1* and 3′ specific oligo-

nucleotide probes to analyse Southern blots of cosmid DNA restriction digests.

Restriction maps for each cosmid were aligned with one another by comparison of restriction site pattern, and the hybridization pattern of *Spi-2* sequences. Two contigs (contiguous genomic segments) were constructed in this way (Figure 1). Contig 1 comprises 138 kb of continuous DNA and contains seven separate *Spi-2* sequences, each up to 15 kb in length, with intergenic distances of 10–20 kb. Contig 2 contains 74 kb of DNA and contains three *Spi-2* sequences. The genes within the contig are oriented in both directions; no pattern of transcriptional orientation of genes is apparent. Comparison of the cosmid digests with Southern blots of inbred mouse DNAs indicates that the cloned *Spi-2* sequences represent ~80% of those in the mouse genome.

Cloned copies of the major characterized murine repeated sequences B1, B2 and L1 (Bennett *et al.*, 1984; Hastie, 1989) were also used to probe the cosmid digests to establish the nature and distribution of repeats within the contigs. Three B1 repeat elements, one B2 repeat and a single L1 sequence were mapped to restriction fragments within the contigs. The locations of these are shown in Figure 1. The repeat sequences did not reveal any specific pattern in their location, with respect to gene orientation. The abundance of these repeat sequences was somewhat lower than might have been expected from the average frequency of repeats in the mouse genome, which contains on average one B1 and one B2 repeat sequence per 10 kb, and one L1 sequence per 40 kb. (Bennett and Hastie, 1984).

A search was carried out for the presence of CpG islands (Bird, 1986) in the contigs, as indicated by clustering of sites for restriction enzymes cutting CpG containing sequences.

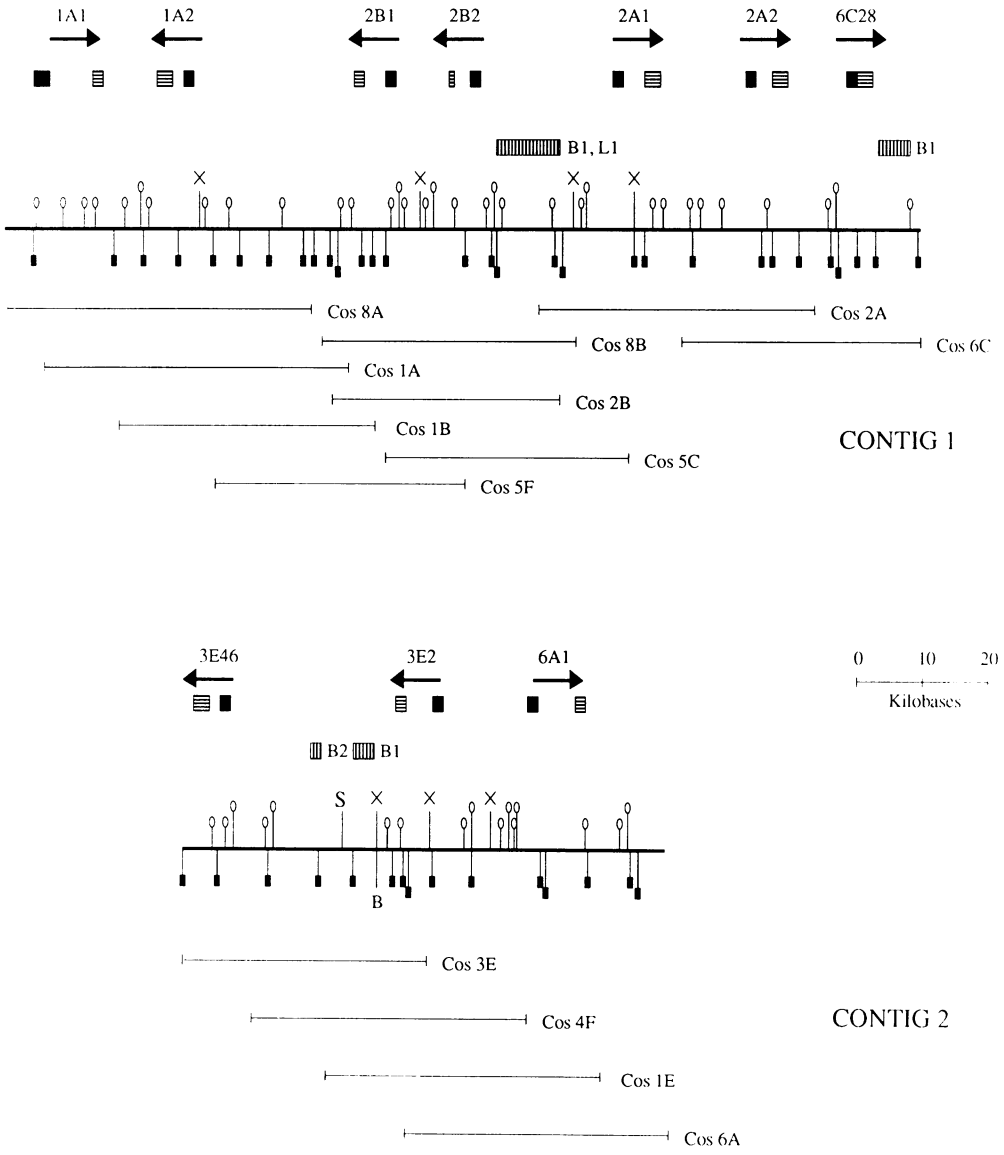

Fig. 1. Structure of contigs 1 and 2 from the mouse *Spi-2* locus. A restriction map for each contig is shown with sites for *Bam*HI, *Hin*dIII, *Xho*I and *Sal*I. The regions spanned by individual cosmids are shown below the map, with the cosmid title. The location of the *Spi-2* sequences is shown above the restriction map, with the relevant subclone title. Solid flags, *Bam*HI sites; open flags, *Hin*dIII sites; X-flags, *Xho*I sites; B-flag, *Bss*HII site; S-flag, *Sal*I site. The location of the 5′-hybridizing sequences are shown as solid bars, and 3′-hybridizing sequences as horizontally hatched bars. Transcriptional orientation is shown for each sequence by a solid arrow, 5′ to 3′. The location of repeated sequences is indicated by vertically hatched bars and the repeat type (see text).

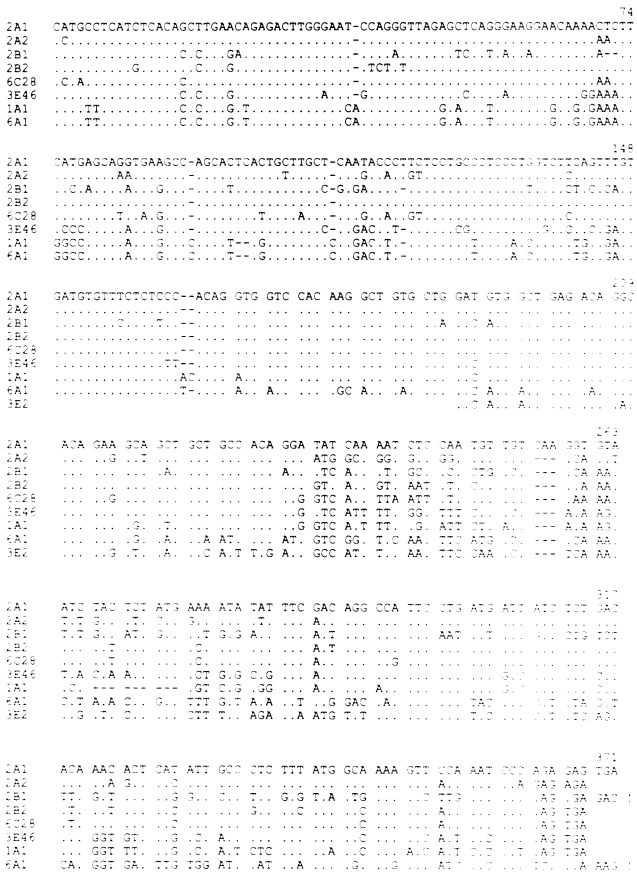

Fig. 2. Nucleotide sequences of exon 5 and part of intron 4 from the *Spi-2* locus genes. Sequences were aligned using the GAP program from the UWGCG package (Devereux *et al.*, 1984) and adjusted to give maximal identity in the reactive centre region. Sequence was not obtained for the contig *Spi-2* sequence 1A2, and a partial sequence only for clone 3E2. The exon 5 coding sequence is divided into coding triplets. Dots represent identity with the reference sequence (2A1), dashes represent deletions introduced to maximize sequence alignment, exclamation marks (!) represent termination codons for longer sequences. The intron/exon boundary is indicated by a double line (nucleotide 170). The reactive centre, defined as the region from nucleotide 237 to 284, is boxed. The P1 residue is encoded by nucleotides 250–252.

Using the enzymes *Bss*HII, *Mlu*I, *Not*I, *Pvu*I and *Sac*II, a single *Bss*HII site was found, 5 kb from the only *Sal*1 site in Contig 2 (shown in Figure 1). This indicates that none of the *Spi-2* sequences on the contigs are associated with highly CpG enriched upstream sequences and may imply that expressed genes within the complex are tissue specific, since many tissue specific genes are devoid of CpG islands.

Sequencing of reactive centres

In order to characterize the individual *Spi-2* contig sequences further, restriction fragments containing the 3′ coding regions of the genes were isolated and subcloned into the plasmid pBS for sequence analysis. Defined subclones containing reactive centre encoding regions were sequenced using primers designed from rodent *Spi-2* cDNA sequence data. (Hill and Hastie, 1987). The region sequenced includes the exon containing the reactive centre domain and part of the preceding intron. Nucleotide and predicted amino acid sequences derived from nine *Spi-2* contig clones are shown in Figures 2 and 3a. The sequences show a remarkable range of diversity in the reactive centre region, in contrast to the

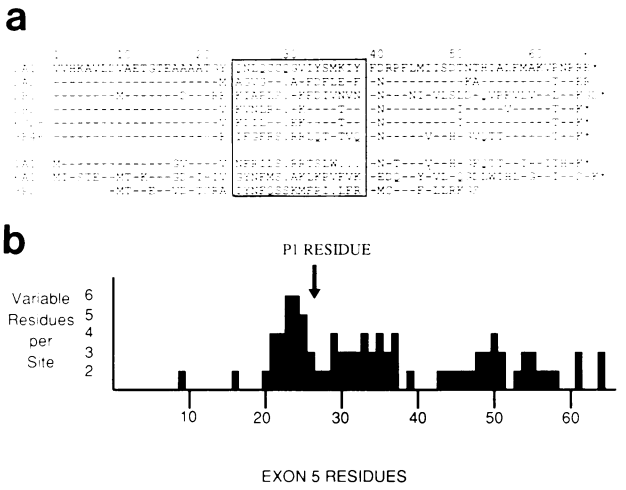

Fig. 3. Predicted peptide sequences and amino acid variability of exon 5. (**a**) The sequences are aligned as in Figure 2, and sequences are shown in the one letter amino acid code. The three sequences considered to represent possible pseudogenes are shown below the main cluster. The reactive centre (residues 23–28), is boxed. The putative P1 residue is at position 27. Dashes represent identity with the reference sequence (2A1), and asterisks represent termination codons. (**b**) The graph shows the number of different residues at each position of the exon 5 coding sequence, for the six apparently functional *Spi-2* sequences from 3a. The position of the P1 residues is indicated.

coding sequences immediately 5′ and 3′ to this 45 bp domain. These flanking sequences exhibit high conservation throughout the *Spi-2* family, and the serpin superfamily (Carrell *et al.*, 1987; Huber and Carrell, 1989). They appear to be important in the maintenance of a functional tertiary structure, and probably in the actual mechanism of proteinase inhibition (Loebermann *et al.*, 1984; Stein *et al.*, 1990; Wright *et al.*, 1990).

Three of the genes (pCos6A-1, PCos3E-2 and pCos1A-1) exhibit substantial divergence in the highly conserved motifs flanking the reactive centre region, and may represent inactive genes. In the case of pCos3E-2, extensive sequence divergence throughout the final exon indicates that this gene may be an inactivated pseudogene. However, since no frameshift mutations or premature stop codons were observed in any of the genomic *Spi-2* coding sequences, none can be defined as pseudogenes at this stage.

Four of the genes (pCos2A-1, pCos2A-2, pCos2B-2 and pCos6C-28) encode a cysteine residue at the functionally important P1′ position; three of these in turn encode a cysteine residue at the neighbouring P1 position. The effect of a cysteine residue at the P1 and adjacent P1′ positions is not clear, as this amino acid does not appear in these positions in any naturally occurring serpin yet reported.

The remaining two genes appear competent to encode conventional serpins with predictable specificities. The sequence pCos2B-1 contains a leucine–serine pairing at the P1–P1′ position, and resembles the human α1-antichymotrypsin gene which has an identical P1 residue. The second gene, PCos3E-46, appears to encode an anti-tryptic inhibitor since its P1 residue is the trypsin substrate, arginine.

The hypervariability of the reactive centre region is demonstrated graphically in Figure 3b for six of the *Spi-2* sequences. The alignment of coding sequences demonstrates almost no similarity in the first half of the reactive centre, including the P1 residue. This corresponds precisely to the region thought to be responsible for the inhibitory specifi-

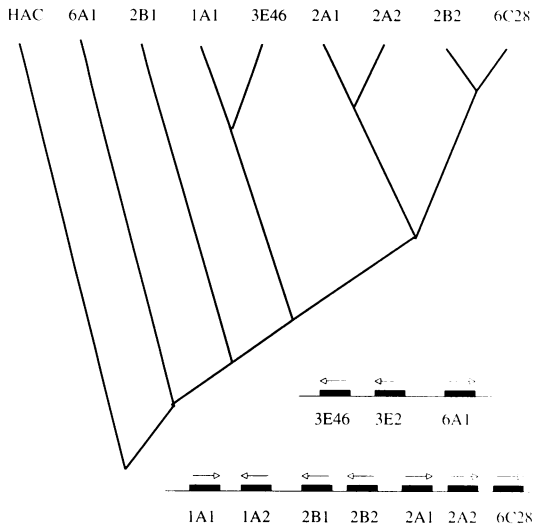

Fig. 4. Consensus phylogeny for the duplication of the *Spi-2* locus coding sequences. This figure represents the probable duplication order of the genes as indicated by the PHYLIP package. The final coding exon was used as input for the DNAPARS and PROTPARS progams. The sequences for 1A2 and 3E2 were excluded as a complete exon 5 sequence was not available. HAC, human α1-antichymotrypsin gene, shown as an outgroup. The maps below the tree represent the orientation of genes within the *Spi-2* locus, and demonstrate the lack of correlation with the consensus phylogeny.

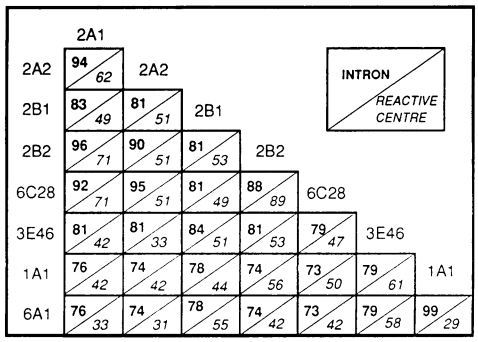

Fig. 5. Pairwise comparison of intron and reactive centre sequences for eight *Spi-2* sequences. For each pairwise comparison, the intron identity (%) is shown in bold type and the reactive centre identity (%) in italics.

city. The contrast between this region and its flanking sequences is apparent, with the putative pseudogenes contributing most of the observed variation in three conserved domains (see Figure 3a).

Analysis of gene branching order

Inspection of the cloned *Spi-2* sequences (Figures 2 and 3a) shows that the sequences display a hierarchy of relatedness; but that this is not clearly related to their transcriptional orientation, or their nearest neighbours on the contigs. The degree of relatedness of both the DNA and amino acid sequences were analysed using the PHYLIP package. The DNAPARS and PROTPARS programs were used to establish a probable phylogeny for the sequences. An equivalent tree was obtained for most variations of sequence input order, and is shown with human α1-antichymotrypsin as an outgroup in Figure 4. This tree is not quantitative as regards time of duplication, but gives an indication of possible branching order. More quantitative estimates are unlikely to be meaningful given the high probability of gene conversion events in a clustered gene family of this kind.

The pair of sequences apparently most closely related, pCos2B-2 and pCos6C-28, show almost complete identity outside the reactive centre region with five synonymous and one asynonymous (Leu for Val) changes in 147 nucleotides. Strikingly, within the reactive centre region there are two synonymous and four asynonymous changes in 48 nucleotides (Val for Leu, Asn for Ile, Arg for Cys and Gly for Glu). This concentration of divergent nucleotides within the reactive centre even in the two most recently separated sequences, further emphasizes the unusual nature of sequence divergence at this locus.

Evolution in flanking introns

The intron flanking the 5′ end of the reactive centre exon was partially sequenced for several of the cosmid subclones, to determine whether the reactive centre divergence was appreciably greater than that of a nearby non-coding region. The oligonucleotide p54/25− was used to generate ~180 bp of intron sequence for eight genes. These are shown in Figure 2. These data provide an indication that gene conversion events have occurred during the evolution of the *Spi-2* locus. The gene pairs 2A1/2B2, 2A2/6C28 and 1A1/6A1 all show extensive stretches of total sequence identity within their intron, although exon sequence data suggest that none of the pairs are most closely related phylogenetically. All of the conversion events terminate 5′ of the reactive centre region. In the case of the 2A1/2B2 conversion, where the 5′ boundary of the conversion event can be tentatively identified, a short region of nucleotide mismatch occurs. As shown in Figure 5, the introns exhibit far less divergence than the reactive centre regions, confirming that the reactive centres tend to be hypervariable following duplication, and also that homogenizing gene conversion events have tended to exclude the reactive centre. Nucleotide divergence values in the remainder of the exon are much more similar to the intron values (data not shown) as would be expected for recently separated genes.

Isolation of a novel Spi-2 transcript

To determine whether any of the cloned genomic *Spi-2* sequences were expressed, a mouse liver cDNA library was screened with probes derived from the sequences in Figure 2. The use of genomic fragments containing reactive centre sequences to isolate cDNAs was complicated by the considerable sequence similarity of family members in regions flanking the reactive centre, and also by high levels of contrapsin expression in the liver (Hill *et al.*, 1984). To circumvent this problem, gene specific oligonucleotide probes were synthesized for the two sequences which conform most closely to other characterized serpins, pCos2B1 and pCos3E46. Expression of the 2B1 sequence was never detected using the 2B1 oligonucleotide to probe Northern blots of mouse RNAs, and no cDNA clones corresponding to this sequence were isolated. Screening of the cDNA library with the 3E46 specific oligonucleotide resulted in the isolation of four cDNA clones from a primary screen of 10^5 colonies. The largest of these clones was a 1.3 kb partial *Spi-2* cDNA sequence identical to the pCos3E46 sequence in the reactive centre region (data not shown). Northern analysis confirms the expression of this sequence as a ~1.9 kb transcript in the mouse liver.

Discussion

The work presents the genomic structure of most of a tightly clustered mouse multigene family which has arisen from a comparatively recent series of gene duplications. Comparison of genomic complexity at the *Spi-2* locus between the laboratory mouse and rat reveals that the murine genes may have effectively tripled in number since the last common ancestor shared by these two species (Hill *et al.*, 1985). Most of the genes resulting from the duplication process have retained conserved motifs which flank a hypervariable domain which encompasses the reactive centre. No data on the structure of an intact serpin inhibitor are available, but the recently published crystallographic structure of intact ovalbumin, a serpin with no known inhibitor activity, may provide an insight into the general tertiary structure of intact serpins (Stein *et al.*, 1990). The ovalbumin domain corresponding to the reactive centre of inhibitory serpins is isolated from the remainder of the protein and is readily accessible as a bait-type structure to act as a substrate for a specific serine proteinase. The domain to the amino-terminal side of the reactive centre is a nine residue peptide stalk which holds the reactive centre domain away from the protein. The structures of proteinase-cleaved α1-antitrypsin and cleaved ovalbumin (plakalbumin) have been determined and compared (Loebermann *et al.*, 1984; Wright *et al.*, 1990). They indicate that a major difference between inhibitory and non-inhibitory serpins is the behaviour of the stalk region following scission at the P1 site. In the inhibitor the strand becomes inserted into the major β-sheet of the inhibitor, while in plakalbumin the reactive centre domain remains detached from the core structure. The domain which comprises the peptide stalk (residues 13−20 in Figure 3a) is remarkably conserved in the *Spi-2* gene family, and in fact throughout the serpin superfamily (Carrell *et al.*, 1987; Huber and Carrell, 1989). In active inhibitors, this region is rich in residues with small side chains such as glycine and alanine, and replacements may lead to serpin inactivation (Perry and Carrell, 1989; Levy *et al.*, 1990). Thus, functional inhibitors are under constraints to conserve this stretch of amino acid residues. Taking into account the number of replacements which have occurred elsewhere in this exon, we predict that the majority of the *Spi-2* sequences here are capable of encoding active inhibitors.

No crystallographic data are available on the structure of a serpin−proteinase complex to date. However, numerous structural studies of the interaction of smaller unrelated inhibitors with serine proteinases have been performed. These indicate that the reactive centre conformation of entirely different inhibitor types may be highly similar when interacting with any one proteinase. Thus, tentative inferences can be made about general reactive centre properties in different systems. Recent analysis of complexes between proteinases and non-serpin inhibitors has shown that up to nine amino acid residues flanking the P1 residue may directly contact the proteinase (Bode *et al.*, 1986a,b). This region appears to correspond to residues 23−31 in Figure 3a, which is the most variable region in the reactive centre. Thus, the *Spi-2* hypervariable domain may be ideally placed to provide the potential for the gene family to encode a wide range of proteinase inhibitor specificities.

Within this exon of the *Spi-2* genes there may be several molecular evolutionary mechanisms at work. In an interspecies comparison of *Spi-2* genes in mouse, rat and human, we have previously demonstrated that the rate of nucleotide substitution in the reactive centre exceeds that expected for neural regions of DNA, such as pseudogenes, introns and untranslated regions (Hill and Hastie, 1987). The intraspecific reactive centre hypervariability may also be a result of this accelerated rate of change. However, analysis of the interspecific sequences is complicated by the possibility that horizontal sequence exchange may have occurred within the gene family following gene duplication. The occurrence of gene conversion events in multigene family evolution is now thoroughly established (Slightom *et al.*, 1980; Bentley and Rabbits, 1983; Weiss *et al.*, 1983; Clark *et al.*, 1985). Analysis of the *Spi-2* sequences indicates that three probable conversion events have homogenized the structurally important upstream motifs and 5′ flanking intron. We have not detected any gene conversion events encompassing the reactive centre domain. This is reminiscent of the process observed in the major histocompatibility complex (MHC) genes, where localized conversion events preserve structurally important framework regions while allowing nearby specificity-determining domains to diverge and remain polymorphic (Lawlor *et al.*, 1990). It should be noted that the observed conversions have been detected by virtue of their lengths, and that microconversion events comprising only a few basepairs may have been overlooked. It is tempting to speculate that the actual process of gene conversion may cause sequence disruption at the conversion event boundaries, as can be seen in the 2A1/2B2 event. This might represent one plausible mechanism for the generation of novel reactive centre sequences which could then undergo selection for novel specificities.

It is apparent that the predicted gene duplication branching order does not directly relate to the position or orientation of *Spi-2* sequences within the contigs. This increases the difficulty of constructing an evolutionary history for the formation of the complex. At the kallikrein locus, a clustered multigene complex on mouse chromosome 7, the genes are all orientated in a head to tail fashion and have clearly arisen by a series of recent tandem duplication events (Evans *et al.*, 1987). For the *Spi-2* locus, no such process can readily be identified, since the transcriptional orientation of the genes does not reveal a defined unit of amplification. Either a large part of the contig which might reveal an overall pattern is missing, or the complex has evolved by amplification of a complex subunit, followed by a deletion of individual sequences within a cluster, as has been postulated for the MHC class II family (Figueroa and Klein, 1988). Since we feel that we have isolated the majority of mouse *Spi-2* sequences in this work, the latter explanation is more likely.

The paucity of repeats in the *Spi-2* contigs may exclude another potential mechanism for the amplification process which has occurred. Unequal crossing over events due to mispairing of repeated sequences on homologous or non-homologous chromosomes are a well documented source of intragenic and intergenic duplication events (reviewed in Maeda and Smithies, 1986). No pattern of repeat distribution in the contigs appears able to explain the gene duplication events which have occurred. However, the possibility that an undetected class of repeated sequence, a simple polynucleotide repeat for example, has mediated gene amplification cannot be excluded. In addition, the comparative rarity of repeated sequences may reflect recent amplification of the gene family, as the insertion of repeat elements may accumulate over time (Rodgers, 1985).

An important point that the *Spi-2* reactive centre data raises is the nature of the selective forces causing hypervariability in this region of the gene. Selection for resistance to parasitic infection has been proposed as a potential mechanism, since many parasites secrete proteinases during host invasion (Hill and Hastie, 1987). We suggest that, following gene amplification, a period of co-evolution occurs between proteinase and novel inhibitor. If the proteinase is parasitic in origin, this may be a continuous evolutionary process as the parasite evolves to avoid the host response. Alternatively, amplification of the *Spi-2* genes may have occurred as a counterbalance to the amplification and divergence of a family of serine proteinases. The striking amplification, mentioned above, of the kallikrein serine proteinase family in mouse is of particular interest, since a rat *Spi-2* gene has recently been shown to encode a major kallikrein binding protein (Chao *et al.*, 1990). The kallikrein family has also been shown to display hypervariability in its active site domain (Creighton and Darby, 1989).

The discovery of *Spi-2* sequences with a putative Cys−Cys pairing at the P1−P1′ position in the reactive centre was unexpected. The effect of cysteine on putative inhibitory specificity and activity is unclear, as no published serpin sequences share these residues. Site directed mutagenesis on the P1′ residue of antithrombin III indicated that the replacement of serine by cysteine caused a considerable reduction in, but did not abolish, inhibitory activity (Stephens *et al.*, 1988). However, the inhibitory activity of such a protein may be compensated for by other changes in the reactive centre region. The Cys−Cys dipeptide at the P1−P1′ position may indicate the existence of novel serine proteinases to which the mouse is exposed. It is also possible that the Cys−Cys genes may encode proteins with an alternative function. Thyroxine binding globulin and cortisol binding globulin are recent examples of serpins with a defined biological activity but no obvious inhibitory activity (Flink *et al.*, 1986; Hammond *et al.*, 1987). We are exploring the possibility that these unusual *Spi-2* sequences are expressed, with good evidence that at least one is transcribed.

The pCos3E-46 (Arg−Ser) gene, which is expressed in the liver, resembles a more orthodox inhibitory serpin. Sequence analysis indicates that this sequence is a typical *Spi-2* gene, whose closest human equivalent is the α1-antichymotrypsin gene. However, the reactive centre sequence of the pCos3E46 gene is more similar to that of another less related human serpin, Protein C Inhibitor (PCI) (Suzuki *et al.*, 1987). The PCI and pCos3E46 genes share seven common residues in the first half of the reactive centre, including the P1 residue. Whether this represents a case of convergent evolution on the part of the mouse gene remains to be determined.

Materials and methods

Materials

All restriction enzymes, T4 DNA ligase and *Escherichia coli* DNA polymerase 1 (Klenow fragment) were obtained from Boehringer Mannheim. T4 polynucleotide kinase, Hybond-N, [α^{32}P]dCTP (>3000 Ci/mmol and [γ-^{32}P]ATP were from Amersham. Oligonucleotides were synthesized on an Applied Biosystems 381A machine. Sequencing was performed with the Sequenase 2 kit from United States Biochemicals. A 129/St mouse cosmid library in pCosEMBL2 (Ehrich *et al.*, 1987) was supplied by Dr L.Stubbs (ICRF, London). The mouse pBR322 liver cDNA library was purchased from Clonetech.

The *Spi-2* specific probes used to isolate and characterize genomic clones were pLv54 (Hill *et al.*, 1984) and pSpi2.1 (Hill and Hastie, 1987), and sublones of these sequences. The mouse repeat specific probes pMR225 (B1 specific), pMR142 (B2 specific) and pMIF (LI-Md specific) were obtained from N.Hastie (Bennett *et al.*, 1984).

Oligonucleotides employed as probes and sequencing primers were as follows: p54/XV+CAGGTGGTCCACAAGGCT sense 18 mer sequencing primer; p54/129+ GGCACAGAAGCAGCTGCT sense 18 mer sequencing primer; p54/51− AG(AGCT)GG (AG)AACCT(AG)TT(AG)AA antisense 17 mer sequencing primer; p54/RS− TCTTCTAGAACGAAAGAC antisense 18 mer probe; p54/LS− TTTGCAGACAGTGGGGCA antisense 18 mer probe.

Isolation of Spi-2 positive sequences

Isolation and analysis of cosmid clones was carried out using standard techniques (Maniatis *et al.*, 1982). Partial digestions to facilitate cosmid mapping were carried out by first linearizing cosmid DNA with *Sal*I, followed by limited endonuclease digestion with the appropriate enzyme (1 U/μg, 3 min, 37°C). Following electrophoresis and Southern transfer, partial digests were hybridized with end specific probes subcloned from pCos2EMBL to establish restriction site order. Subcloning and sequencing of reactive centre fragments were carried out using standard methods.

Screening of plasmid cDNA libraries, and subsequent analysis of isolated clones was also performed according to standard methods (Maniatis *et al.*, 1982). Southern transfer and isolation of DNA and RNA were carried out as described in Hill *et al.* (1985).

Sequencing analysis

The GELSIZE program (Schaffer and Sederoff, 1981) was used to establish accurately the lengths of partial digest restriction fragments. Alignment and manipulation of sequences was performed with the UWGCG package (Devereux *et al.*, 1984). For sequence phylogeny analysis, the PHYLIP programs (Felsenstein, 1985) maintained at SEQNET, on the VAX3600 at Daresbury, were employed.

Acknowledgements

We would like to thank Chris Sime for excellent technical advice, and Nick Hastie for stimulating discussion. We are grateful to Dr Lisa Stubbs for the 129/St cosmid library.

References

Bennett,K.L. and Hastie,N.D. (1984) *EMBO J.*, **3**, 467−472.
Bennett,K.L., Hill,R.E., Pietras,D.F., Woodworth-Guati,M., Kane-Haas,C., Houston,J.M., Heath,J.K. and Hastie,N.D. (1984) *Mol. Cell. Biol.*, **4**, 1561−1571.
Bentley,D.L. and Rabbitts,T.H. (1983) *Cell*, **32**, 181−189.
Bird,A.P. (1986) *Nature*, **321**, 209−212.
Bode,W., Papamokos,E., Musil,D., Seemueller,U. and Fritz,H. (1986a) *EMBO J.*, **5**, 813−818.
Bode,W., Wei,A.-Z., Huber,R., Meyer,E., Travis,J. and Neumann,S. (1986b) *EMBO J.*, **5**, 2453−2458.
Carrell,R.W., Pemberton,P.A. and Boswell,D.R. (1987) *Cold Spring Harbor Symp. Quant. Biol.*, **LII**, 527−535.
Carrell, R.W., Aulak,K.S. and Owen,M.W. (1989) *Mol. Biol. Med.*, **6**, 35−42.
Chao,J., Chai,K.X., Chen,L.-M., Xiong,W., Chao,S., Woodley-Miller,C., Wang,L., Lu,H.S. and Chao,L. (1990) *J. Biol. Chem.*, **265**, 16394−16401.
Clark,A.J., Chave-Cox,A., Ma,X. and Bishop,J.O. (1985) *EMBO J.*, **4**, 3167−3171.
Courtney,M., Jallat,S., Tessier,L.H., Benavente,A., Crystal,R.G. and Lecocq,J.P. (1985) *Nature*, **313**, 149−151.
Creighton,T.E. and Darby,N.J. (1989) *Trends Biochem. Sci.*, **14**, 319−324.
Devereux,J., Haeberli,F. and Smithies,O. (1984) *Nucleic Acids Res.*, **12**, 387−395.
Ehrich,E., Craig, Poustka,A., Frischauf,A.M. and Lehrach,H. (1987) *Gene*, **57**, 229−237.
Evans,E., Drinkwater,C. and Richards,R. (1987) *J. Biol. Chem.*, **262**, 8027−8034.
Felsenstein,J. (1985) *Evolution*, **39**, 783−791.
Figueroa,F. and Klein,J. (1986) *Immunol. Today*, **5**, 78−81.
Flink,I.L., Bailey,T.J., Gustafson,T.A., Markham,B.E. and Morkin,E. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 7708−7712.
Hammond,G.L., Smith,C.L., Goping,I.S., Underhill,I.A. Harley,M.J.,

Reventos,J., Musto,N.A., Gunsalus,G.L. and Bardin,C.W. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 5153–5157.
Hastie,N.D. (1989) In Lyon,M.F. and Searle,A.G. (eds), *Genetic Variants and Strains of the Laboratory Mouse*. Oxford University Press, Oxford, UK, pp. 559–573.
Hill,R.E. and Hastie,N.D. (1987) *Nature*, **326**, 96–99.
Hill,R.E., Shaw,P.H., Boyd,P.A., Baumann,H. and Hastie,N.D. (1984) *Nature*, **311**, 175–177.
Hill,R.E., Shaw,P.H., Barth,R.K. and Hastie,N.D. (1985) *Mol. Cell. Biol.*, **5**, 2114–2122.
Holmes,W.E., Lijnen,H.R. and Collen,D. (1987) *Biochemistry*, **26**, 5133–5140.
Huber,R. and Carrell,R.W. (1989) *Biochemistry*, **28**, 8951–8966.
Hunt,L.T. and Dayhoff,M.O. (1980) *Biochem. Biophys. Res. Commun.*, **95**, 864–871.
Jallat,S., Carvallo,D., Tessier,L.H., Roecklin,D., Roitsch,C., Ogushi,F., Crystal,R.G. and Courtney,M. (1986) *Protein Eng.*, **1**, 29–35.
Laskowski Jr,M. and Kato,I. (1980) *Annu. Rev. Biochem.*, **49**, 593–626.
Lawlor,D.A., Zemmour,J., Ennis,P.D. and Parham,P. (1990) *Annu. Rev. Immunol.*, **8**, 23–63.
Levy,N.J., Ramesh,N., Cicardi,M., Harrison,R.A. and Davis,A.E. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 265–268.
Loebermann,H., Tokuoka,R., Deisenhofer,J. and Huber,R. (1984) *J. Mol. Biol.*, **177**, 531–556.
Maeda,S. and Smithies,O. (1986) *Annu. Rev. Genet.*, **20**, 81–108.
Maniatis,T., Fritsch,E. and Sambrook,J. (1982) *Molecular Cloning: a Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
Ohkubo,H., Kageyama,R., Ujihara,M., Hirose,T., Inayama,S. and Nakanishi,S. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 2196–2200.
Owen,M.C., Brennan,S.O., Lewis,J.H. and Carrell,R.W. (1983) *New Engl. J. Med.*, **309**, 694–698.
Perry,D.J. and Carrell,R.W. (1989) *Mol. Biol. Med.*, **6**, 239–243.
Rodgers,J.H. (1985) *Int. Rev. Cytol.*, **93**, 187–279.
Schaffer,H.E. and Sederoff,R.R. (1981) *Anal. Biochem.*, **115**, 113–122.
Slightom,J.L. Blechl,A.E. and Smithies,O. (1980) *Cell*, **21**, 627–638.
Stein,P.E., Leslie,A.G.W., Finch,J.T., Turnell,W.G., McLaughlin,P.J. and Carrell,R.W. (1990) *Nature*, **347**, 99–102.
Stephens,A.W., Siddiqui,A. and Hirs,C.H.W. (1988) *J. Biol. Chem.*, **263**, 15489–15852.
Suzuki,K., Deyashiki,Y., Nishioka,J. Kurachi,K., Akira,M., Yamamoto,S. and Hashimoto,S. (1987) *J. Biol. Chem.*, **262**, 611–616.
Takahori,H. and Sinohara,H. (1982) *J. Biol. Chem.*, **257**, 2438–2446.
Travis,J. and Salvesen,G.S. (1983 *Annu. Rev. Biochem.*, **52**, 655–709.
Weiss,E., Golden,L., Zakut,R., Mellor,A. and Fahrner,K. (1983) *EMBO J.*, **2**, 453–462.
Wright,H.T., Qian,H.X. and Huber,R. (1990) *J. Mol. Biol.*, **213**, 513–528.

Received on October 8, 1990; revised on November 13, 1990