Organization of Two Genes Encoding Cytotoxic T Lymphocyte-Specific Serine Proteases CCPI and CCPII[†,‡]

Corrinne G. Lobe,[§] Chris Upton, Brenda Duggan, Nancy Ehrman, Marc Letellier, John Bell, Grant McFadden,[∥] and R. Chris Bleackley*[,∥]

Departments of Biochemistry, Immunology, and Genetics, University of Alberta, Edmonton, Alberta, Canada T6G 2H7

Received February 29, 1988; Revised Manuscript Received April 25, 1988

ABSTRACT: The genes encoding two recently described cytotoxic T cell proteases, CCPI and CCPII, have been isolated and sequenced. The organizations of the coding and noncoding portions of the two genes are very similar to each other and also to the gene encoding rat mast cell protease type II. Similarly to other serine protease genes, each of the active-site residues is contained on a separate exon; however, two introns were found in particularly interesting positions. One occurs within the postulated activation dipeptide and the other in a position close to the active-site Asp residue. This latter intron interrupts the amino acid sequence in the invariant core region of the protein. We believe that these genes represent a new subfamily of serine protease genes.

Cytotoxic T lymphocytes (CTL)[1] bind to cells bearing foreign antigens and, by a mechanism which is still unclear, cause them to lyse. An understanding of how CTL function has important biological and clinical implications. To this end we recently isolated cDNA clones representing mRNAs specifically expressed in CTL in the belief that some would encode proteins which played important roles in CTL-mediated lysis (Lobe et al., 1986a).

Sequence analysis of two of these clones (C11 and B10) revealed that they were remarkably homologous to each other, with 80% identity at the nucleotide level. The full-length C11 clone encodes a protein of predicted M_r 25 319, designated cytotoxic cell protein I (CCPI). This protein appears to be a serine protease, since the amino acid residues that form the catalytic triad of the serine protease active site and the sequences neighboring these residues, which are highly conserved among proteases, are present in CCPI (Lobe et al., 1986b). From the partial sequence of the B10 clone, it also appears to encode a serine protease, referred to as CCPII.

The level of expression of the mRNAs corresponding to B10 and C11 correlated with cytotoxic activity (Lobe et al., 1986a). The maximum expression preceded the peak of cytotoxic activity in an in vitro allogeneic or mitogen-induced response by 12–24 h, thus suggesting that the protein products may well play an important role in mediating the killer cell function. Indeed, protease inhibitors have been shown to suppress CTL-induced killing (Redelman & Hudig, 1980; Simon et al., 1987). The time course of expression together with the sequence data raises the possibility of a protease cascade mechanism of activation, analogous to the activation of the complement components (Reid, 1986). Alternatively, they may themselves be toxic molecules that are directly involved in the destruction of the target cell.

The predicted C11 protein product has 12 residues at the N-terminus that are highly hydrophobic, suggesting that this is part of a signal sequence to direct secretion or intracellular organelle location (Lobe et al., 1986a). Indeed, by use of antibodies generated against a synthetic peptide derived from the CCPI sequence, the protein has been localized in the cytoplasmic granules (Redmond et al., 1987). Following the signal sequence is a dipeptide, Gly-Glu, believed to be the activation peptide. Such a short activation peptide, of only two amino acid residues, would be novel for a serine protease. By analogy to other serine proteases, the activation peptide would be cleaved from the proenzyme to convert the inactive zymogen to the active form of CCPI (Neurath & Walsh, 1976; Salveson et al., 1987).

CCPI is homologous with rat mast cell protease type II (RMCPII). This was particularly interesting, as RMCPII possesses a number of unusual structural features which indicate that it has a substrate specificity quite different from that of classical serine proteases (Woodbury et al., 1978; Woodbury & Neurath, 1980). CCPI shares several of these features and in addition has other unique changes that alter the environment in the active-site pocket (Lobe et al., 1986b), suggesting that it too has unusual substrate specificity.

The genes for a number of serine proteases have now been sequenced and compared for evidence of evolutionary relationships and correlations between protein domains and exons. Here we report the cloning of the genes encoding CCPI and CCPII and present a comparison of their genomic organizations with each other and with other serine proteases, notably rat mast cell protease type II (RCMPII) (Benfey et al., 1987).

Materials and Methods

DNA Preparation. Phage or plasmid DNA was purified by the plate lysis or rapid alkaline lysis method, respectively (Maniatis et al., 1982). To prepare genomic DNA, 10^7 cells/mL were lysed in 0.5% SDS, and the solution was adjusted to 100 μg/mL proteinase K and incubated at 37 °C overnight. Following phenol/chloroform extractions, the solution was adjusted to 0.3 M NaOAc and 67% ethanol. High molecular weight DNA was spooled out on a Pasteur pipet, air-dried, rinsed with 70% ethanol, and dissolved in 10 mM Tris-HCl, pH 8/1 mM EDTA at 4 °C.

Restriction enzyme digests were carried out in 1× Core Buffer (Bethesda Research Laboratories) at 37 °C either overnight (phage and genomic DNA) or from 1 to 4 h

[†] Supported by the National Cancer Institute of Canada

[‡] The nucleic acid sequences in this paper have been submitted to GenBank under Accession Number J02834.

* To whom all correspondence should be addressed.

[§] Present address: Max Planck Institute of Biophysical Chemistry, Göttingen, West Germany.

[∥] Scholars of the Alberta Heritage Foundation for Medical Research.

[1] Abbreviations: CTL, cytotoxic T lymphocyte; CCP, cytotoxic cell protease; RMCP, rat mast cell protease.

0006-2960/88/0427-6941$01.50/0

(plasmid DNA). Digested DNA was then electrophoresed on agarose gels and analyzed by Southern blotting (Southern, 1975). Alternatively, insert DNA was isolated by polyacrylamide gel electrophoresis and purified by the crush-soak method (Schleif & Wensink, 1981).

Molecular Probes and Hybridizations. The DNA probes used were inserts or subfragments from the CTL-specific clones B10 and C11 (Lobe et al., 1986b), purified by polyacrylamide gel electrophoresis and the crush-soak method (Schleif & Wensink, 1981). The DNA probes were nick-translated by using a nick-translation kit (Bethesda Research Laboratories) or oligo-labeled (Feinberg & Volgelstein, 1983) to 1–5 × 10^8 cpm/μg.

Blots were prehybridized in 50% formamide, 20 mM phosphate buffer pH 6.8, 2 mM pyrophosphate, 100 μm ATP, 5× Denhardt's, 5× SSC, 100 μg/mL salmon sperm DNA, 0.1% SDS, and 2.5 mM EDTA at 47 °C for 2–15 h. Hybridization was carried out in the same buffer, with 1–5 × 10^6 cpm/mL of DNA probe. After 15 h at 47 °C, blots were washed in 0.1× SET/0.1% SDS for 60′ at 65 °C, with three changes of buffer. The filters were then exposed to X-ray film using an intensifying screen.

Screening of Genomic Libraries. The genomic B10 clone was isolated from a partial *Eco*RI, size-selected genomic library prepared from CBA/J mouse liver DNA in the λ charon 4A vector. Two million recombinants, propagated in *Escherichia coli* NEM 259, were screened in duplicate at 2 × 10^5 plaques/plate. Another murine genomic DNA library, generously provided by Mark Davis (Stanford), was used to isolate the genomic C11 clone. The library contained partially *Mbo*I cut CBA/J mouse liver DNA in the vector λJ1 and was grown in *E. coli* DL191. A total of 2 × 10^6 plaques were screened in duplicate. Plaques were lifted onto nitrocellulose filters, denatured in 0.5 M NaOH/1.5 M NaCl, and neutralized in 1 M Tris-HCl, pH 7/1.5 M NaCl. The filters were allowed to air-dry and then baked at 80 °C for 1.5 h in a vacuum oven. They were then hybridized with insert DNA from B10 or C11, washed, and exposed to X-ray film as described above.

DNA Sequence Analysis. All DNA sequences were determined by primer extension on single-stranded template derived from M13 clones using the dideoxy method (Sanger et al., 1977). In the case of C11, recombinant phage DNA was sonicated to an average size of 200–500 base pairs and then repaired and ligated into mp18 (Deininger, 1983; Cool & MacGillivray, 1987). M13 clones containing exon sequences were detected by using a radioactive C11 cDNA probe (Lobe et al., 1986b) and sequenced. Further sequence information was gained by cloning subfragments of the genomic C11 gene into mp18. The sequence of the 3′ boundary and a significant portion of intron 1 was determined by using a primer oligonucleotide synthesized, on the basis of the C11 cDNA sequence, by the Regional DNA Synthesis Laboratory, Calgary, Alberta.

For the B10 gene, a series of overlapping deletions were generated from an internal *Pst*I site by using exonuclease III (Henikoff, 1984) and sequenced.

DNA sequence data were stored and analyzed by using the Microgenie programs.

Results and Discussion

CCPI and CCPII Are Encoded by Separate Genes. The initial sequence comparison between the two genes which encoded CCPI (C11) and CCPII (B10) clearly demonstrated that they were very closely related (80% identity over 400 residues). One possible explanation for this was that the two

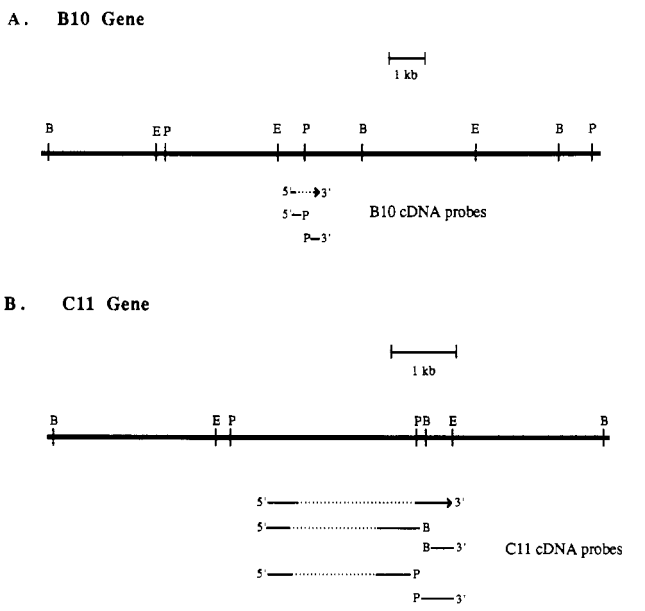

FIGURE 1: Partial restriction maps of the B10 and C11 genes. Genomic DNA was cut with restriction enzymes *Bam*HI (B), *Eco*RI (E) and *Pst*I (P), fractionated by electrophoresis in agarose, transferred to nitrocellulose, and hybridized with B10 and/or C11 cDNAs indicated. The direction of transcription is indicated by the arrow. The scale is indicated by a 1-kB bar and is different for the two genes.

were encoded by the same gene with different exon usage, as has been seen with calcitonin and calcitonin-related protein (Amara et al., 1982). Preliminary experiments revealed that Southern blots conducted under standard conditions of hybridization and washing gave very complicated patterns. This was due primarily to the relatedness of the B10 and C11 genes and, in retrospect, to the similarity of the whole family of granule proteases (Masson & Tschopp, 1987). However, conditions were found under which cross-hybridization was minimized. A series of genomic Southern blots with a variety of restriction enzymes using C11 and B10 cDNAs as probes revealed that the pattern of hybridizing fragments was different. On the basis of these data, preliminary restriction maps of the B10 and C11 were constructed as shown in Figure 1. Even though the maps are similar, the two related proteins are encoded by separate genes.

Isolation and Sequencing of the C11 Gene. A murine genomic library (kindly provided by Mark Davis) was screened with radioactively labeled C11 cDNA probe. A number of positive plaques were identified, but the restriction map of only one corresponded to that of C11, as defined by the high-stringency genomic blots described above. Fragments corresponding to exon–intron boundaries were identified by sonicating the cloned genomic DNA to an average size of 200–400 base pairs, subcloning into M13, and screening by hybridization with a full-length C11 cDNA (Deininger, 1983; Cool & MacGillivray, 1987). Their sequences were determined by the dideoxy method described by Sanger (Sanger et al., 1977). The sequence of the gene encompassing the translated portion of C11 and the intronic regions is presented in Figure 2. The lower line of amino acids corresponds to the protein sequence predicted from the cDNA analysis (Lobe et al., 1986b). The coding portion of C11 is interrupted by four introns. The first exon encodes the putative signal peptide, a common feature among genes encoding proteins that are destined for secretion or intracellular localization. However, the site of the first intron is particularly interesting, as it corresponds to an interruption of the putative activation dipeptide postulated for CCPI (Lobe et al., 1986b), human cathepsin G (Salveson,

```
gc11fromATG

ATGAAGATCCTCCTGCTACTGCTGACCTTGTCTCTGGCCTCCAGGACAAAGGCAGGTGAGTAAGCAACACTTCCTTTAGT    80
MetLysIleLeuLeuLeuLeuThrLeuSerLeuAlaSerArgThrLysAlaG

GGTCTAACCCCGCTGTCACCAGTAGCAGGAGGGCAGAGGTCTTTCTCAGGAAGGGCCATGAATGCTCTCCGCTCCTGAGC   160
CATTCCTTATCATCTGCAAAGTGGATCTTCCTAGAAGTACATGATCAAGGGTAGGAACCTGCAAATCAGAAAACAATATG   240
TTCTCCTGGAGCAGATAGAATGGGCCTGGTTTCTGCAACTATGATGATACAGGCTTTCCCCTCACCCCAAACTCAGTCAT   320
CAAAGAAGCCTAGGTGTTTATGACCCCCGAGTAGATGGAATTGCTGTCCTCAGAACAAGTGATTTTGCTTGTCCCTAGCT   400
ACTTGGTCTCTTCCTCTAACTCAGTGCTTCACAGGGCTCTAGTGGTTCCCCAGTGTTCAAGGATGAGCCACTGGGCACTC   480
TGCTTTCTGGAAGGCCAGCACCTACATCTAGGAATGAAAGCCAGTGGGAGGCCTGAAGAAATAAGATGCTGTGAAGTGTG   560
AACTTTGGCCCCTAGAAAGCAGCAGTATTCCAGGCATCACAGGGCTGGGAGCAAGGGAGAGAAGGTCTTAGAAACTCTGG   640
GACCAGAAGTCAGAAGACAGACAAGCTGAATCCTTAGGATTTCCTTGTCCAGAGAATTATGTTTGCCTCCAACTGTGTGG   720
AAGGAAGATTGATAAGCCAAGACCTATATTTCTCTCCTTTTAAAGAAGGGGAAAATGAGGGCACTCCTCCCAACCTCCTT   800
CCAACTCCAGGACTCCCTGGCTCCCAAATGCCTAGCAAGTCTGGAGCTACCAGACAAAACCCATCTGGTGATCTAGCAGG   880

TGCCCACAAAATGATTGGCTTTCCTTCAGGGGAGATCATCGGGGGACATGAAGTCAAGCCCCACTCTCGACCCTACATGG   960
                              lyGluIleIleGlyGlyHisGluValLysProHisSerArgProTyrMetA

CCTTACTTTCGATCAAGGATCAGCAGCCTGAGGCGATATGTGGGGGCTTCCTTATTCGAGAGGACTTTGTGCTGACTGCT  1040
laLeuLeuSerIleLysAspGlnGlnProGluAlaIleCysGlyGlyPheLeuIleArgGluAspPheValLeuThrAla

GCTCACTGTGAAGGAAGGTGAGGAGCAAAAAAACCAGTCCATACCTGCCTAGAAAGATTCCATGGAGGCTCCGCTTCATC  1120
AlaHisCysGluGlySe

CTAAGGTGCTTGGGAGGAGAGAGGGTCTAGCAAGTCTCATGAGGGGGAAACAGACTGTCACCAGAGTCACTGATAAGGAG  1200
TCAAGCACATTTCAACCAGGGTTAGCATTGCAGTATGAGCTGAATAACTCAGTTTGCCTTCTTAGAGCTGATGGGCTCCT  1280
GAGCACCTCTAATGCAGCAACTGGTGCTTCAGATTATCCCTTTACTTCCCTCAGCTGGAGCCCCACCTGCGCCCTGCCTG  1360

TCCTTCACACACCTCCTGGGCAGCATCCTCTCTGACTCCACGTCTCTTACGTCTGCTCTCTTCACAGTATAATAAATGTC  1440
                                                                rIleIleAsnVal

ACTTTGGGGGCCCACAACATCAAAGAACAGGAGAAGACCCAGCAAGTCATCCCTATGGTAAAATGCATTCCCCACCCAGA  1520
ThrLeuGlyAlaHisAsnIleLysGluGlnGluLysThrGlnGlnValIleProMetValLysCysIleProHisProAs

CTATAATCCTAAGACATTCTCCAATGACATCATGCTGCTAAAGGTGAGACCTGTCCCTGTCTTGGTCCACGGACTGCCCT  1600
pTyrAsnProLysThrPheSerAsnAspIleMetLeuLeuLys

CTTGCTCTCATCCTCTCAATTTCCTCCCTTTCATCCTGGCCTTCATGTGTTTCAGACCAGCAGGCCATGAGCTGGACTCT  1680

TGCTTCTTCCCCAACAGCTGAAGAGTAAGGCCAAGAGGACTAGAGCTGTGAGGCCCCTCAACCTGCCCAGGCGCAATGTC  1760
                    LeuLysSerLysAlaLysArgThrArgAlaValArgProLeuAsnLeuProArgArgAsnVal

AATGTGAAGCCAGGAGATGTGTGCTATGTGGCTGGTTGGGGAAGGATGGCCCCAATGGGCAAATACTCAAACACGCTACA  1840
AsnValLysProGlyAspValCysTyrValAlaGlyTrpGlyArgMetAlaProMetGlyLysTyrSerAsnThrLeuGl

AGAGGTTGAGCTGACAGTACAGAAGGATCGGGAGTGTGAGTCCTACTTTAAAAATCGTTACAACAAAACCAATCAGATAT  1920
nGluValGluLeuThrValGlnLysAspArgGluCysGluSerTyrPheLysAsnArgTyrAsnLysThrAsnGlnIleC

GTGCGGGGGACCCAAAGACCAAACGTGCTTCCTTTCGGGTAAGTTGGGTTGCAGTCCCTCTGGGCTAAGTGGGAGGGGAA  2000
ysAlaGlyAspProLysThrLysArgAlaSerPheArg

AAGGAATCTGGGACCTAGAGACCCAAATATCAAGGGACTCCTTTACCCACTGGCTGTGATCTTTCTCCCTGGGAACAGCA  2080
GGTACTAGTAATGAACTGGGGCCCCCAGAGCTGACTAGGAGCCTCTGCTGAAGGTAGCTTGTACCAAAAGGAGGTGTTGG  2160
CAACACAGTACTGTCCACCCAGCTGTAGAAAGCTGGGCTCCCTGGTGTGTCATAACCACACCCATGTGTCTCCTCTGAGC  2240
AGACACACACTGTTGTCAGTGGGCTTCCCCTCCATGCTCACACCTGGCCCACTCACTGTCCCCATGTCCTTAAACAACAG  2320
CCTAGAGACGAGGGTCCCACACACCTTCCTAGAGCGTGGATCACAGACCTGGGGAAAGGCAGAGGCTGCCTCAGGAGGGA  2400
GGTGCAGCCCTAACCATGGTCCACAGTCAGAGCTGGGAAGCCGGGGGCCCCAGGACTGTCCTCTGACCTCCCATAAGCAT  2480
AAATCATGCTTCTCTGGGAGGAGCCTTGACATGAGGAGGTTGGGACCAGGGTGAACAATAGCTCAGTGCCTTGTATCCAC  2560

TCAATTCAACAGGGGGATTCTGGAGGCCCGCTTGTGTGTAAAAAAGTGGCTGCAGGCATAGTTTCCTATGGATATAAGGA  2640
            GlyAspSerGlyGlyProLeuValCysLysLysValAlaAlaGlyIleValSerTyrGlyTyrLysAs

TGGTTCACCTCCACGTGCTTTCACCAAAGTCTCGAGTTTCTTATCCTGGATAAAGAAAACAATGAAAAGCAGCTAACTAC  2720
pGlySerProProArgAlaPheThrLysValSerSerPheLeuSerTrpIleLysLysThrMetLysSerSerEnd

AGAAGCAACATGGATCCTGCTCTGATTACCCATCGTCCCTAGAGCTGAGTCCAGGATTGCTCTAGGACAGGTGGCAGGAT  2800
CTGAATAAAGGACTGCAAAGACTGGCTTCATGTCCATTCACAAGGACCAGCTCTGTCCTTGGCAGGCCAATGGAACACCT  2880
CTTCTGCCACCATGCTGTGACAACCCAACTGACATCTTCCTATGGAAGTTTGCCCTCTCCACAAAAGAAGTAGAATGTTT  2960
GCATTGGAGCTGGGCATGCTCTGCTTCCCCTCAGTGCCCCGAGAATGTTATCTAATGCTAGTCATCATTAATAGCTCCCT  3040
ACAGAACTTTCATACAGTTGCACCCAAGTTGCTGATGTGTTCTCTAGAATAGAGCAAGAAATAGTAAACAGAATTCCTTT  3120
TGCCTCTCTGTACTATTTTCCCCCAAATACCAAGATTTGTATGTTTTATAAAGCTAATTTCCTTATCAAATGACATCTTT  3200
TAATTTTTACATTAATGGCTTATTTTCAAGGTACAACCTGATTTTTTTATGGACAAAAATGATCGTAAAATCAAGTAAAA  3280
CTAATTAATATATCC                                                                     3290
```

FIGURE 2: Sequence of genomic C11. Subfragments of the genomic C11 DNA were recloned in M13 and sequenced by using the dideoxy method. The upper line shows the nucleotide residues, and the amino acids are given below. Nucleotide 1 corresponds to the first residue of the initiation codon. Intron sequences do not have amino acid residues below them. Splice sites that interrupt codons are indicated by separation of the letters of the three-letter code for amino acids. The final nucleotide residue corresponds to the site of polyadenylation.

1987), and rat mast cell protease II (Benfey et al., 1987). The activation sequence coding portion of serine proteases is often interrupted by an intron; however, the fact that this is even the case for a dipeptide indicates that its position may be important. Possibly it is advantageous to be able to generate variability in the activation sequence. The second exon encodes the remainder of the activation dipeptide plus the first 147 amino acids of the predicted mature protein, including the histidine residue that forms part of the catalytic site of the serine protease. Interestingly, the other catalytic triad residues are encoded by separate exons (3 and 5, respectively). This segregation of active-site domains between different exons is

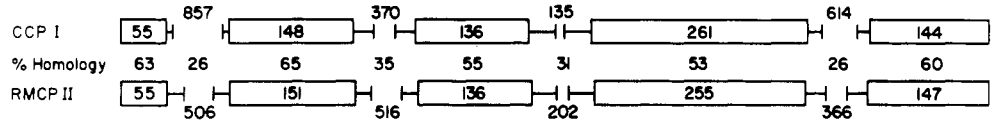

FIGURE 3: Comparison of the genomic organizations of CCPI and RMCPII. Boxes indicate exon sequences, and the numbers within correspond to the nucleotide residues. Introns are shown as broken lines with the number of nucleotides given above. Comparable regions were compared by using the Microgenie alignment program. The resulting percentage similarities are indicated between the two genes.

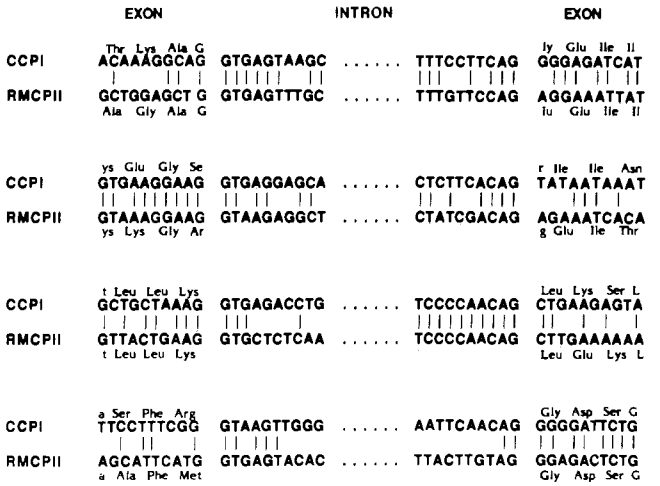

FIGURE 4: Sequence comparisons of exon/intron boundaries of CCPI and RMCPII. The nucleotides surrounding each of the exon–intron boundaries were aligned. Identical residues are indicated by vertical lines and the positions of the boundaries by a space in the nucleotide sequences.

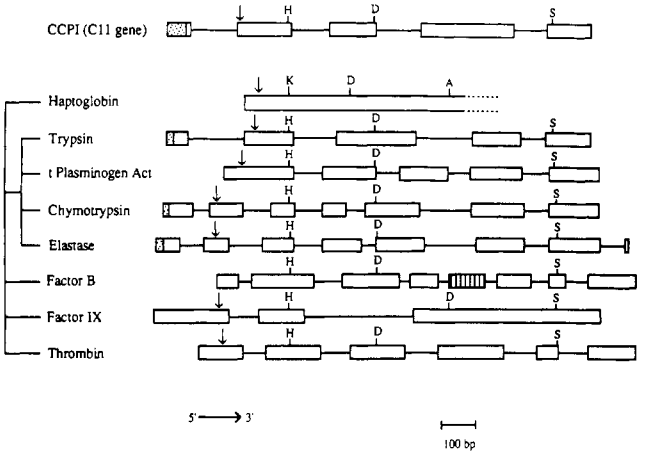

FIGURE 5: Comparison of the CCPI gene with other serine proteases. The CCPI-encoding C11 gene is shown at the top, and below it are serine protease genes representing each of the five groups, distinguished by gene organization. Exons are shown as boxes, while introns are depicted as lines. Locations of the codons for the active-site residues (His, H, Asp, D, and Ser, S) are shown. The vertical arrow designates the point of cleavage of the activation peptide. Complement factor B is not activated in this way. The signal peptide is denoted as a dotted box and the unique exon of complement B factor as a vertically stripped box. Exons are drawn to the scale shown, but introns are not drawn to scale.

a common feature of serine protease genes (Rogers, 1985).

Genomic Organizations of CCPI and RMCPII Are Similar. In our original discovery that CCPI was a serine protease, the most homologous member of the family was RMCPII (Lobe et al., 1986b). The genomic sequence and organization of RMCPII have recently been reported (Benfey et al., 1987), and in Figure 3 the two are compared. The size of each of the exons is very similar, and their high degree of sequence similarity is a reflection of what we originally observed at the protein level. Presumably, the murine equivalent of RMCPII would be even more similar in the coding portions of the gene. The sizes and sequences of the introns differ markedly; however, their positions are very similar between the two genes. Indeed, when the sequences surrounding each of the boundaries were aligned, the positions of the splice sites were identical in all cases (Figure 4). Clearly these two genes share an evolutionary ancestor. It will be interesting to see how the organization of the Hanukah factor (Gershenfeld & Weissman, 1986), other granzymes (Masson & Tschopp, 1987), and human cathepsin G genes (Salveson et al., 1987) compares with those of CCPI and CCPII.

CCPI Intron 3 Is in an Unusual Position. The organization of many of the serine protease genes is known, ranging from the ancestral, intron-less bacterial serine protease genes to the more complex vertebrate genes. The vertebrate serine protease genes have been grouped according to their intron organization, since the genes seem to have evolved by exon shuffling and intron insertion (Rogers, 1985; see Figure 5). The first group consists of the haptoglobin gene, which has no introns interrupting the catalytic triad residues. The second group is the trypsin family. In these genes, an intron occurs just downstream of the His codon, another occurs just downstream of the Asp codon, and another occurs just upstream of the Ser codon. Therefore, there is one exon to each of the catalytic triad residues and a fourth between the Asp and Ser exons. Three variations on this basic pattern exist, which probably arose by intron insertion. One is exemplified by the tissue-type plasminogen activator gene, which has an extra intron just downstream from the Asp codon. Another variation is seen in the chymotrypsin gene, which has two extra introns, one upstream of the Asp codon and one between the His and activation peptide encoding sequences. Finally, the elastase gene has the two additional introns of the chymotrypsin gene plus another intron splitting the exon downstream of the Ser codon. The third family of serine protease genes consists of the complement factor B gene, which contains seven introns throughout the catalytic region. The fourth group is exemplified by the factor IX gene, which has two introns, one separating the sequence encoding the activation peptide and the His codon and one between the His and Asp codons, leaving the Asp and Ser codons on the same exon. Last, the thrombin gene comprises the fifth group, in which the catalytic region is split by five exons. It would appear that the CCPI gene fits into the trypsin subfamily of organization except for one major difference, that is, the site of intron 3.

Previous analyses of serine protease genes (Craik et al., 1983) have concluded that introns map to areas of variability, as defined by differences (insertions or deletions) between eukaryotic and prokaryotic proteases, and to surface regions of the proteins. The position of intron 3 of CCPI does not correspond with either. Although the positions of introns 1, 2, and 4 are similar to those in chymotrypsin, the third intron site of CCPI does not correspond to a region of variability as defined above. The second unusual feature of intron 3 is its position with respect to the folding pattern of a typical serine protease (Read & James, 1988). The molecule consists of two domains, each of which has a central core region and a number of external loops. The residues at which the CCPI gene is

```
gb10fromATG

ATGCCACCAGTCCTGATTCTCCTGACCCTACTTCTGCCTCTCAGAGCTGGAGCAGGTGAGTGAACACCCCTGTGGCAGTG    80
MetProProValLeuIleLeuLeuThrLeuLeuLeuProLeuArgAlaGlyAlaG

ATGCCGTCCCATGGCACCTCTGCTGACTTTCTGCCCAGCGGTAGTCAGGAATTGTTCTGGCACTGGACCCAGGATGTTTC   160
TGAATGTCAGGTCCCATAACAAACCCAGACTTAGAAATCTGACACTGGCCTAATTTTCAACAAGGTCTGAAGAAAGGAGA   240
TGCTTTCAGAAAACTTAGTAGGCACTAGCTTCTCCCAACATTTTTTCATCACCTGGAGACTGGATCATCCTTACAAAGAG   320
ATGTGACAGAGAATGTAAAGATACAAAAATGTGCACCATTTACTTTATTTCAGCCTAGGGAATGAGCAGCTTTAATGAGG   400
CTTAAGATTCAGCCAGGATTATTCTATGTGCAAGGAATCCAGTCTTCAAAAGCCTTGAGGGACTTCAGATTCCACTCTTA   480

ATTATTCCCTCCCTAGATTCCTCTTGGCCACAAATATTGCTCCCAGACACCAAATATTCTAGGCTTTTCTTTCAGAGGAG   560
                                                                          luGlu

ATAATCGGAGGCAATGAGATCAGTCCACATTCCCGTCCCTACATGGCATATTATGAGTTTCTGAAAGTTGGTGGGAAGAA   640
IleIleGlyGlyAsnGluIleSerProHisSerArgProTyrMetAlaTyrTyrGluPheLeuLysValGlyGlyLysLy

GATGTTCTGCGGAGGCTTCCTGGTTCGAGACAAATTCGTGCTAACAGCTGCTCACTGCAAAGGAAGGTGAGGAGAAGCAG   720
sMetPheCysGlyGlyPheLeuValArgAspLysPheValLeuThrAlaAlaHisCysLysGlySe

ATAGCTCACATCTTCTGAAACATCCACAGGGGCTTCTGTCCTCTATTGTGGAAGCTCCCCTGAGGGCTCACTAGAAGTCA   800
GGGTTGCAAAAATTAATGTCTGACAAGAGGTATATTTGAGGTAAAAGGGTGAGAGGTTAAATCCAGCATCATACTTGAGT   880
TGATCAACTGCACTGGCCAAACAAAAGTAGTGGTTGAGATTAAAAGTTCATGTAGCTGTCTCTTGTGAGACTATGCTGGG   960
GCCTAGCAAACACAGAAGTGGATGCTCACAGTCAGCTATTGGATGGATCACAGGGCTCCCAATGGAGGAGCTAGAGAAAG  1040
TACCCAAGGAGCTAAAGGGATCTGCAACCCTATAGGTGGAACAACATCATGAACTAACCAGTACCCCGGAGCTCTTGACT  1120
CTAGCTGCATATGTATCAAAAGATTGCCTAGTCGGCCATCACTGGAAAGAGAGGCCCATTGGACACGCAAACTGTATATG  1200
CCCCAGTACAGGGGAACGCCAGGGCCATAAAAATGGGAATGGGTGGGTAGGGAAGTGGGGGGGAGAGTATGGGGGACTTT  1280
TGGGATATCATTGGAAATGTAATTGAGGAAAATATGTAATAAAAAATATTTTTTAAAAAAGTTCACATTAAGCCGGGCTG  1360
TGGTGGCGCACACCTTTAATCCCAGCACTTAGGAGGCAGGGACAGGCAGATTTCTGAGTCCAAGGAGCAGGGACAGGCAG  1440
ATTTCTGAGTCCGAGGCCAGCCTGGTCTACAAAGTGAGTTCCAGGACAGCCAGGGCTATACAGAGAAACCCTGTCTCGAA  1520
AAACCAAAAAAAAAAAAACAAAAAAAAAAAAACACATTAAAAGTTTATGTAAACTTCAGAATAATGCCTAGATGCAAGAGTG  1600
AACCTCCTGGAATTTCATGTCTTTGAGACCCTCCTGTGTCCTCAAGTCCCCTCAGCTACAGTTCCATCCTGCCTACCTTG  1680

CTTTACACAGGTCCTGCACTACATCCCCTCTGACTCCACATCTCTGCTCTCCACTCTGCTCTCTCCACAGCTCAATGACA  1760
                                                                    rSerMetThr

GTCACACTGGGGGCTCACAACATCAAGGCTAAGGAGGAGACACAGCAGATCATCCCTGTGGCAAAAGCCATTCCCCATCC  1840
ValThrLeuGlyAlaHisAsnIleLysAlaLysGluGluThrGlnGlnIleIleProValAlaLysAlaIleProHisPr

AGACTATAATCCTGATGACCGTTCTAATGACATCATGCTATTAAAGGTGAGACCTGCCATCCTCCCAGTCACATACCCCC  1920
oAspTyrAsnProAspAspArgSerAsnAspIleMetLeuLeuLys

ACCCCTCATCCCACCTCTGGTGCCTGTGCTCTGCTTGGTACAGGGCTCCTCCCTGCCCTCCTTCTCTGATCTTGCTTCTT  2000
CCCTTCTTCCATTACAAGTTTCAGTGTGACCCAGAGCATCACCCTCTTGGCTGAGCTTCTTTCTCCTGTCTCTTCCCTAT  2080

CAGCTGGTGAGAAATGCCAAGAGGACTAGAGCTGTGAGGCCCCTCAACCTGCCCAGGCGCAATGCTCATGTGAAGCCAGG  2160
   LeuValArgAsnAlaLysArgThrArgAlaValArgProLeuAsnLeuProArgArgAsnAlaHisValLysProGl

GGATGAGTGCTATGTGGCTGGTTGGGGAAAGGTAACCCCGGACGGGGAATTCCCAAAAACACTGCACGAAGTTAAGCTGA  2240
yAspGluCysTyrValAlaGlyTrpGlyLysValThrProAspGlyGluPheProLysThrLeuHisGluValLysLeuT

CAGTACAGAAGGATCAGGTGTGTGAGTCCCAGTTCCAAAGTTCTTACAACAGAGCTAATGAGATATGTGTGGGAGACTCA  2320
hrValGlnLysAspGlnValCysGluSerGlnPheGlnSerSerTyrAsnArgAlaAsnGluIleCysValGlyAspSer

AAGATCAAGGGAGCTTCCTTTGAGGTAAGTTGGATTGCCTTCAACACTGGGCTCAGTTGGAGGGAAAAGGAACCTGGGAC  2400
LysIleLysGlyAlaSerPheGlu

CTAGAGACCTCAAGGAACTCTTTTGTCCACTGGCTGTGATCTTTCTCCCTGGGAACAGCAGGAATCAGAAACTAAGCAGG  2480
GCCCCCAGAGCTGACTAAGGAGTCTCTGCTGAAGGTAGCTTGTACAAAAGGAGGTGTTGGCAACACAATACCTGTCTCCA  2560
GGCCCAGGCTGTAGAAAGCTGGGCTCCCTGGGTGTGTCATAACACACCATGTGTCTCCTCTGAGCAGACACACACTTGTC  2640
AGTGGCTTCCCCTCCATGCTCACACCTGGCCCACTCACTGTCCCCATGTCCTTAAACAATAGCCTAGAGACGAGGGTCCC  2720
ACACACCTTCCTAGAGAGTGGATCACAGACCTGGGGAAAGGCAGAGGCTGCCTCAGGAGGGAGGTGCAGCCCTAACCATG  2800
TCCACAGTCAGAGCTGACTGACCTCCCATAAGCATAAGTCATGCTTCTCTGAGAGGAGCCTTGACATGAGAAGGAGGTTG  2880

GGAACAACAGCTCAGTGTCCCGTACCCACTCAATTCACAGGAGGATTCTGGAGGCCCGCTTGTGTGTAAAAGAGCAGCTG  2960
                                        GluAspSerGlyGlyProLeuValCysLysArgAlaAlaA

CAGGCATCGTCTCCTACGGGCAAACTGATGGATCAGCTCCGCAAGTCTTCACAAGAGTTTTGAGTTTTGTATCGTGGATA  3040
laGlyIleValSerTyrGlyGlnThrAspGlySerAlaProGlnValPheThrArgValLeuSerPheValSerTrpIle

AAGAAAACGATGAAACACAGCTAACTACAAGAAGCAACTAGATCCTGACTGACAGCCATCTTCCCATAGCTGAGTCCAGG  3120
LysLysThrMetLysHisSerEnd

ATTGCTCTAGGACAGATGGCAGGCAACTGAATAAAGAACTTTCTCTGACTGC                              3170
```

FIGURE 6: Sequence of genomic B10. Subfragments of the genomic B10 DNA were recloned in M13 and sequenced by using the dideoxy method. The upper line shows the nucleotide residues, while the amino acids are given below. The positions of the introns were predicted by comparisons with the organizations of the CCPI and RMCPII genes. Intron sequences do not have amino acid residues underneath. Splice sites that interrupt codons are indicated by separation of the letters of the three-letter code for amino acids. The final nucleotide residue corresponds to the site of polyadenylation.

interrupted by introns 2 and 4 correspond very closely with those found in chymotrypsin, that is, in loop structures which are found on the surface of the majority of serine proteases. However, intron 3 of CCPI corresponds to an internal β-sheet region running from Ala104 to Leu108 in chymotrypsin. The implications of an interruption in the gene coding sequence in this region are unclear; however, the position of this intron sets CCPI apart from other serine protease genes. It should also be noted that RMCPII (Benfey et al., 1987), adipsin (Min & Spiegelman, 1986), and CCPII (see below) have their third

intron in positions identical with those of CCPI.

Isolation and Sequencing of the B10 Gene. A partial *Eco*RI mouse genomic library was screened with the B10 cDNA probe (Lobe et al., 1986b). Positive plaques were isolated, and DNA was isolated corresponding to the B10 genomic pattern. Fragments were subcloned into pUC13, and a series of deletions were generated by *Exo*III digestions and sequenced in M13 by the dideoxy method (Sanger et al., 1977). The sequence of the B10 gene is presented in Figure 6. By comparison with the organization of the C11 gene and with the B10 cDNA sequence already determined, it was possible to predict the protein coding portions corresponding to CCPII. The genomic organization of B10 very closely resembles that of C11 and RMCPII. Indeed, the sizes and positions of the exons predicted for B10 are almost identical with those of C11. The similarities between the two genes are 56, 66, 79, 82, and 80% for exons 1–5 and 30% or lower for the introns with the exception of 85% for intron 4. It would appear that the 3′ halves of the genes, including even the final intron, are more conserved between B10 and C11 than the 5′ regions. The positioning of intron 1 in B10 predicts that the activation peptide of CCPII will be GluGlu, as was found for RMCPII (Benfey et al., 1987), and in contrast to GlyGlu for CCPI. However, it should be noted that if the boundaries of intron 1 for B10 were shifted over by two residues, an activation peptide of GlyGlu would be generated for CCPII also. The isolation and characterization of a full-length B10 cDNA should resolve this issue.

CCPI and CCPII Genes, Together with RMCPII, Represent a New Subfamily of Serine Protease Genes. The analysis of the exon/intron organizations of CCPI and CCPII presented here reveals that they are very similar to each other and also to another recently characterized protease gene, RMCPII. Notably, the first intron of all three interrupts the coding portion of the genes within the postulated activation dipeptide. A comparison with the organizations of other protease genes suggested that these three belong in the trypsin subfamily. However, the position of one of the introns (closest to the active-site Asp) sets them apart from this group. This intron does not correspond to a region of sequence variability between bacterial and eukaryotic proteases and moreover is located in the core region of the protein. In addition, the primary protein sequences of CCPI, CCPII, and RMCPII suggest that they are structurally unusual compared with the overall family of serine proteases. This even extends to the lack of a disulfide bond, which functions in other proteases to stabilize the substrate binding pocket. We therefore conclude that the three genes are closely related to each other and represent archetypal members of a new subclass of serine protease genes. It will be interesting to see which other cellular proteases also fall into this same group.

ACKNOWLEDGMENTS

We thank Dr. Mark Davis for the mouse genomic DNA library and Beverly Bellamy and Dawn Oare for their preparation of the manuscript.

REFERENCES

Amara, S. G., Jonas, V., Rosenfeld, M., Org, E., & Evans, R. (1982) *Nature (London) 298*, 240–244.

Benfey, P. N., Yin, F. H., & Leder, P. (1987) *J. Biol. Chem. 262*, 5377–5384.

Cool, D. E., & MacGillivray, R. T. A. (1987) *J. Biol. Chem. 262*, 13662–13673.

Craik, C. S., Rutter, W. J., & Fletterick, R. (1983) *Science (Washington, D.C.) 220*, 1125–1129.

Deininger, P. L. (1983) *Anal. Biochem. 129*, 216–223.

Feinberg, A. P., & Volgelstein, B. (1983) *Anal. Biochem. 132*, 6–13.

Gershenfeld, H. K., & Weissman, I. L. (1986) *Science (Washington, D.C.) 232*, 854–858.

Henikoff, S. (1984) *Gene 28*, 351–359.

Lobe, C. G., Havele, C., & Bleackley, R. C. (1986a) *Proc. Natl. Acad. Sci. U.S.A. 83*, 1448–1452.

Lobe, C. G., Finlay, B. B., Paranchych, W., Paetkau, V. H., & Bleackley, R. C. (1986b) *Science (Washington, D.C.) 232*, 858–861.

Maniatis, T., Fritsch, E. F., & Sambrook, J. (1982) in *Molecular Cloning, a Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Masson, D., & Tschopp, J. (1987) *Cell (Cambridge, Mass.) 49*, 679–685.

Min, H. Y., & Spiegelman, B. M. (1986) *Nucleic Acids Res. 14*, 8879–8892.

Neurath, H., & Walsh, K. A. (1976) *Proc. Natl. Acad. Sci. U.S.A. 73*, 3825–3832.

Read, R. J., & James, M. N. G. (1988) *J. Mol. Biol.* (in press).

Redelman, D., & Hudig, D. (1980) *J. Immunol. 124*, 870–878.

Redmond, M. J., Letellier, M., Parker, J. M. R., Lobe, C. G., Havele, C., Paetkau, V., & Bleackley, R. C. (1987) *J. Immunol. 139*, 3184–3188.

Reid, K. B. M. (1986) *Nature (London) 322*, 684.

Rogers, J. (1985) *Nature (London) 315*, 458–459.

Salveson, G., Farley, D., Shuman, J., Przybyla, A., Reilly, C., & Travis, J. (1987) *Biochemistry 26*, 2289–2293.

Sanger, F., Hicklen, S., & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 5463–5467.

Schleif, R. F., & Wensink, P. C. (1981) in *Practical Methods in Molecular Biology*, Springer-Verlag, New York.

Simon, M. M., Fruth, U., Simon, H. G., & Kramer, M. D. (1987) *Ann. Inst. Pasteur 138*, 309–314.

Southern, E. M. (1975) *J. Mol. Biol. 98*, 503–517.

Woodbury, R. G., & Neurath, H. (1980) *FEBS Lett. 114*, 189–196.

Woodbury, R. G., Katunuma, N., Kobayashi, K., Titani, K., & Neurath, H. (1978) *Biochemistry 17*, 811–819.