Sequences encoding two trypsin inhibitors occur in strikingly similar genomic environments

I. Barry KINGSTON* and Stephen ANDERSON†

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, U.K.

The nucleotide sequences of two approx. 4 kilobase pair segments of the bovine genome are presented. One segment contains a coding region for bovine pancreatic trypsin inhibitor (BPTI) and the other segment contains a coding region for a BPTI homologue. The two 4 kilobase pair sequences are strikingly similar over approx. 3.4 kilobase pairs of their sequence, including putative intron sequences, suggesting that they have evolved from a gene duplication event.

INTRODUCTION

Trypsin inhibitors belonging to the pancreatic trypsin inhibitor (Kunitz) family occur in organisms as diverse as sea anemones, snails, snakes, cattle and man (Kunitz, 1947; Laskowski & Kato, 1980) suggesting that the gene coding for them is both very old and very widely distributed. The most well characterized member of the family is bovine pancreatic trypsin inhibitor, BPTI. This is a basic protein of 58 amino acids which has been found in many bovine organs (Fritz *et al.*, 1979) including lung, parotid gland, spleen, liver and pancreas. Its amino acid sequence is known (Kassell *et al.*, 1965) and the three-dimensional structure has been determined crystallographically for both the free inhibitor and the enzyme-inhibitor complex (Deisenhofer & Steigemann, 1975; Ruhlmann *et al.*, 1973; Huber *et al.*, 1974a, 1975). Although its exact physiological role is unclear, its interactions with trypsin and other serine proteinases such as kallikrein, chymotrypsin and plasmin have been well characterized. Because of its small size, high solubility and stability, BPTI has been an object of intense physical and theoretical study (Creighton, 1978; Karplus & McCammon, 1981; Wagner & Wuthrich, 1982). Previously we have described the synthesis of a unique sequence 86-bp DNA probe designed on the basis of the amino acid sequence of BPTI (Anderson & Kingston, 1983). Using this probe to screen a genomic library in a bacteriophage vector, we isolated a genomic clone for BPTI. The coding region comprised 174 nucleotides and was located in a contiguous sequence of 677 nucleotides from one of the positively hybridizing phage recombinants. Four positively hybridizing recombinants were purified using the synthetic probe and it was found that two of them represented partially overlapping segments of one region of the bovine genome and the other two represented partially overlapping segments of a second region. The two regions are probably separated in the genome by at least 11 kb (Anderson & Kingston, 1983). With a view to elucidating the genomic organization of these genes, more extensive DNA sequencing of two of the different phage recombinants was undertaken.

In the present paper we present the sequence of two approx. 4 kb segments of the bovine genome isolated from two of the phage recombinants. The two segments contain different trypsin inhibitor coding sequences. The sequences are compared and shown to be extensively similar, suggesting that they have evolved from a gene duplication event.

MATERIALS AND METHODS

Chemicals and reagents

Guanidinium thiocyanate was from Fluka AG and oligo(dT)-cellulose (type 3) was from Collaborative Research Inc. All other reagents were as described previously (Anderson & Kingston, 1983).

DNA preparation and library construction

Bovine DNA was prepared from bovine kidneys, partially digested with *Sau3A* I and a size fraction corresponding to 11–24 kb was used to construct a genomic library in the vector λ 1059 as described previously (Karn *et al.*, 1980; Anderson & Kingston, 1983).

Synthetic probe construction

The construction, and application in Southern blot hybridization, of a synthetic 86-bp DNA probe designed using the amino acid sequence of BPTI (Kassell *et al.*, 1965) has been described (Anderson & Kingston, 1983).

Subclones of λ recombinants

Restriction fragments from λ recombinants hybridizing with the synthetic probe were identified by Southern blot hybridization, isolated by preparative electrophoresis in 0.8% agarose gels (Girvitz *et al.*, 1980; Maniatis *et al.*, 1982) and subcloned into the vector pUC8 (Vieira & Messing, 1982). Recombinant plasmids were prepared from 10 ml of stationary phase cultures by the boiling method of Holmes & Quigley (1981) and the inserts purified by electrophoresis in 1% agarose gels.

Abbreviations used: BPTI, bovine pancreatic trypsin inhibitor; bp, base pairs; kb, kilobases (or base pairs); SCP, 0.1 M-NaCl/0.03 M-NaH₂PO₄/1 mM-EDTA, pH 6.2.

* To whom correspondence and reprint requests should be addressed at his present address: Department of Parasitology, Molteno Institute, University of Cambridge, Downing Street, Cambridge CB2 3EE, U.K.

† Present address: Genentech Inc., 460 Point San Bruno Boulevard, South San Francisco, CA 94080, U.S.A.

DNA sequencing

DNA sequences were determined by the dideoxynucleotide chain termination method (Sanger *et al.*, 1977; Bankier & Barrell, 1983) as described previously (Anderson & Kingston, 1983) except that the modifications of Biggin *et al.* (1983) were used and DNA fragmentation was carried out by sonication (Deininger, 1983; Bankier & Barrell, 1983). All data were compiled using the computer programs of Staden (1982a). The sequences were completely determined on both strands of the DNA and an average of seven independent assignments was made for each nucleotide.

RNA preparation

RNA was prepared from bovine lung, liver and parotid gland by the method of Chirgwin *et al.* (1979), using the modification described by them in which the RNA is separated from the guanidinium thiocyanate homogenate by centrifugation through a cushion of 5.7 M-CsCl/0.1 M-EDTA, pH 7.5. Poly(A)-containing RNA was selected by two cycles of binding to oligo(dT)-cellulose.

Hybridization to Northern blots

Samples of total RNA and lung poly(A)-containing RNA for Northern blot analysis were denatured in glyoxal, fractionated on a 1.1% agarose gel and blotted onto nitrocellulose (Thomas, 1983). Single strand DNA probes were prepared as described previously (Anderson & Kingston, 1983) from M13mp8 clones obtained during sequencing the bacteriophage recombinant containing the BPTI sequence. For hybridization, filters were first heated in 20 mM-Tris/HCl, pH 8.0 (Thomas, 1983) and then pretreated with 50% (v/v) deionized formamide/6.6 × SCP containing 100 µg of sonicated denatured *Escherichia coli* DNA and 20 µg of sonicated denatured M13mp8 DNA/ml, 5 × Denhardt's solution (1 × Denhardt's solution is 0.02% Ficoll, 0.02% polyvinylpyrrolidone and 0.02% bovine serum albumin) and 1% SDS for 4 h at 42 °C. Hybridizations were performed at 42 °C for 20 h in a solution identical with that used for the prehybridization except that dextran sulphate was also present at a concentration of 10% and a probe containing the complementary sequence of the RNA was present at a concentration of 10⁶ c.p.m./ml. After hybridization filters were washed first in 3 × SCP/0.1% SDS at room temperature and then in 0.15 × SCP/0.1% SDS at room temperature or at 65 °C.

Hybridization to Southern blots

Genomic Southern blots of bovine DNA and the preparation of single stranded ³²P-labelled probes were carried out as described previously (Anderson & Kingston, 1983). Briefly, hybridization was performed at 65 °C in 6.6 × SCP containing 100 µg of sonicated denatured *E. coli* DNA and 20 µg of sonicated denatured M13mp8 DNA/ml, 10 × Denhardt's solution, 1% Sarkosyl, 10% dextran sulphate and a single-stranded probe at a concentration of 10⁶ c.p.m./ml. After hybridization, filters were washed twice at room temperature and then four times at 65 °C in 6.6 × SCP containing 1% Sarkosyl.

RESULTS

DNA sequences of probe-hybridizing restriction fragments

When a genomic library containing about 2 × 10⁶ recombinant phage was screened with the synthetic 86 bp

probe, four positively hybridizing clones were isolated and plaque-purified (Anderson & Kingston, 1983). Restriction mapping and partial sequencing (results not shown) showed that two of the clones represented partially overlapping segments of one region of the genome and the other two represented partially overlapping segments of a second region (Anderson & Kingston, 1983). The restriction digests and Southern blots showed that the two different probe-hybridizing regions were located in two of the clones on 4 kb *Eco*RI fragments. These two *Eco*RI fragments were subcloned into pUC8 and after amplification each was excised and its DNA sequence determined. These sequences are presented in Fig. 1.

One of them contained an open reading frame, the translation of which perfectly matched the amino acid sequence of BPTI (Fig. 1). The other contained an open reading frame, the translation of which matched the amino acid sequence of an inhibitor isolated from bovine spleen, spleen inhibitor II (Fioretti *et al.*, 1985).

Identification of genes

Each of the two sequences was analysed for potential genes using the gene search programs in the general sequence program ANALYSEQ (Staden, 1984a,b). The regions of the sequences which correspond to the two inhibitor genes were picked out by this analysis as being coding regions. No other region was identified as having a higher probability of being a coding region but some regions were noted which showed a low probability. The most notable of these were between bases 800 and 900 and 3100 and 3300 approximately in the BPTI clone and between bases 450 and 550 and 3050 and 3250 approximately in the spleen inhibitor clone. In order to see if any of these regions correspond to a known protein, the sequences were translated using the translation option in ANALYSEQ (Staden, 1984a) and 20 amino acid stretches from each of them were screened against the protein databases of Doolittle (1981) and Dayhoff (National Biomedical Research Foundation, U.S.A.) using the rapid search techniques of Wilbur & Lipman (1983) and Lipman & Pearson (1985). None was shown to code for a protein in either of these databases.

A search of both sequences for potential splice sites revealed that both the BPTI gene and the spleen inhibitor II gene are flanked (Fig. 1) by sequences which resemble consensus intron/exon and exon/intron boundary sequences (Mount, 1982). In addition both contain a potential intron/exon splice signal (Keller & Noon, 1984; Ruskin *et al.*, 1984; Zeitlin & Efstratiadis, 1984), upstream from the potential intron/exon splice junction (Fig. 1).

Each clone contains two potential polyadenylation signals, AATAAA(N)₂₄₋₄₀GTGTY (Proudfoot & Brownlee, 1976; McLauchlan *et al.*, 1985), one located upstream of the protein coding region and one located downstream (Fig. 1).

Northern blots

The results of Northern blot analyses of RNA isolated from bovine lung, parotid gland and liver are shown in Fig. 2. Two probes were used (see the Materials and methods section for their preparation): one was the complement of the sequence from nucleotides 2344 to 2652 in the sequence containing BPTI (Fig. 1) and the other was the complement of the sequence from 2558 to

AATTCTGATAATGCAGAGAACTGGTAAGGAGTTCTGATTGTTCTGCTTGATTAAATGGGTTGTAACAGGAT	71
AGTGTCTTGTCTGATCCTAGCATTTCATATGGTGTGTGTTCTGGGGCAAGTCATCTGCAGTTTCTTCACCTGAACAGGGGGACCAGGTTA	161
AATTCAGTGTATTTTGTAAAGTCATGAGATGAGTGAAGGATGGGACCTTAGATAGTGTCTGGACGGGACTGGATTCTAGAACATGCTAG	89
CATGAGTTTCTTAAAGATTACCAAGTCATGAGTATGAAGAGTTTACACTTTCCTGATCAATGACGTCCATTTCCCATCAAAATATTTTAG	251
AACGTTCTAGAACGTCAGGTGATTAAAGGGTTAGAAGTATCAACCTGTCCCTCTGGGGAGAGGAGAACTACTGTGTTCAAGTCAACAATG	179
TCCAAAAGACTCATCTATCTAATGTAGATCATTTTCTCACCACCCCTCTAAAAAATTTATCTTTCAGATATGATCATTTCTCTATTATGA	341
GCCAATGATTCTACCAAAATTACCTACCTCCAATCCTTAATCAAACTCATACACACTGTGTTTCAGAGAGCTTCAGGGGAGGTGCAACAT	269
AATTAATCAGAGAGTTGAGTGACAGCTGAGTGTCTTCCCTCCAAAGGCAACTGCAGGAAGAGCAAGAAATGCAATACTTTTCTATGAGTT	431
CAAGGGATTGTGAAGGGGAAAGTCTTAAGAGGGTCTGGAGGATCAGCACCCCAATCCCATTTGACCGATATTTCTTCTCCATTTGT	359
TGCTCGTGGGGCCAAGACTGCTTTTCCAGGCTGGTACAATAGTAATCAAACTCAAGATATTCTTCTTCTCCTGGCCAGACTATTA	521
CTGTCCCTGGAGTTTATGCTTTATAAAGTGAAGGATGAGGACAGTAAGTAAAGGGTTTTTATAGAGTCCATGAATGACTCTAATAAGTTATTGA	449
TTTTATTTTCTATCAAGATATAGAAAGTTAGAAGTAGACTCATAATTATATAGGCAGGCCTCATCATCAATAGACTAACAAGAATTTT	611
ACGTGGGAGGTGGGTCGTGGAACCTCTGAATTTAGAGCTAGTTGATCACACGTGTAAGTGACCACAGAAATGTTATCAACACTTGAAGT	539
ATTTTATCTGCCTTTTCAATGACTGTGCCTTGGCATGAGGATGAAATGGGAGATTTATCCCTTGATAAATATTCATGAAATACTTATG	701
GTGGGTGAAACCTTTTCAATGACTGTACACTTGGCATGAGGATGAAATGGGAGATTTATCCCTTGATAAATATTCATCAAAATACTTATG	629
CTTTTGTCCCTAAAAAGCATATTTCTTGATATAGGAAAACAGCTGTAACAAAAGGATAGTAAATAATATGCCTTCTAAGAGGGATACA	791
CTTTTGTCCCTAAAAAGCATATTTCTTGATATAGGAAAACAGCTGTAACAAAAGGATAGTAAATAATATGCCTTCTAAGAGGGATACA	719
GACAATAAAGACGGGGGGCGATTCTATACCAGGTGATGATGATGATGATGAGGAAGGTGAGTTATGGGG-TTCAGGATGCTGTAGAGGA	880
GACAATAAAGACGGGGGCAAA-TCCTGTGCCAGGTGATGATGATGATGATGAGGAAGGTGAGTTATGGGGATTCAGGAGGCTGTAGAGGA	808
TCAGGGAAACCTCTGTGATGAGGAGACATTAAGCAGAAAGCTGCCAAAAGGAGGCTGGTGTGTTTGAGCAGCAGCCAGGACCAGGGTGG	970
TCAGGGAAACCCA-TGTGATGAGGAGACATTAAGCAGAAAGCTGCCAAAAGGAGGCTGG--GTGTTTGAGCAGCAGCCAGGACCAGGGTGT	895
CTGGAGCTGAGTGGGTGAGGGGAGGGGAGTGGAAAGGGGATGAAGCAGAGAGGCCATGGGGGCAGGTGATGAGGAACCTTCTAGGACTTTA	1060
GTGAAGCTGAGTGGGTGAGGGGAGGGGAGTGGAAAGGGGATGAAGCAGAGAGGCCATGGGGGCAGGTGATGAGGAACCTTCTAGGACTTTA	985
TAAGGATAAAATTTGACTCTGAGAGAGCTGGGAAACCACTGAGGGACTGGTCTGTTGGAACAACGAGATAGGACTGGAGTTTAAACAGGG	1150
TAAGGATAAAGATTTGACTCTGAGAGAGCTGGGAAACCACTGAGGGACTGGTCTGTTGGAACAACGAGATAGGACTGGAGTTTAAACAGGG	1075
TCCTTGCGACTGCAGTGTGGAGCGTGGCTCTAGGGGGCGAAAAGCAGGGACAGGGGGCCCTGGGCAGGTGGCTGCAGGGGTCCAGTGA	1240
TCCTTGCGACTGCAGTGTGGAGCGTGGCTCTAGGGGGCGAAAAGCAGGGACAGGGGGCCCTGGGCAGGCAGCTGCAGGGGTCCAGTGA	1165
GCTATGATGGAGAATATACCTGTGTTGTTCCCTGGGTTGATTCCAGTTCTCTTGAATAACCTGAATGACTTGTCTATATTAAGATATCT	1330
GCCATGATGGAGAATATA--CCTGTGTTGTTCCCTGGGTTGATTCCAGTTCTCTTGAATAACCTGAATGACTTGTCTATATTAAGATATCT	1253
GGGGAGGCTTCATCACAATCATGATTCAATTAATCTTTAGTCATTTGTGATTGATTCAACCTCCAATCCCTCTCCCTTCTCTGAGATGT	1420
GGGGAGGCTTCATCACAATCATGATTCAATTAATCTTTAGTCATTTGTGATTGATTCAACCTCCAATCCCTCTCCCTTCTCTGAGATGT	1343
GGATACAATGAAAG-----AAGTAGGAATGAAATTTCCACACCCCACTCAGGCAGTTGTTTCCCTGACAACCTATCCCCATTCTTG	1504
GGATACAATGAAAGAGAGGAAAGTAGGAATGAAATTTCCACACCCAGCTCCGGCAGTTGTTTCACTGACAACCTATCCCCATTCTTG	1433
CTTTGCTTGAGGCTTTACAAAACCTCATCTCCCTCACATGATAAAGGACTCCCTTTGCTCTCATCTCTTAGGAAATTTCAATGTTTTAGG	1594
CTTTGCTTGAGGCTTTACAAAACCTCATCTCCCTCACATGATAAAGGACTCCCTTTGCTCTCATCTCTTAGGAAATTTCAATGTTTTAGG	1523
AGCTCTGTGGCAGGAATGGGATGCAGACCAAGTTAATATTTCTTTTATAAGTCACAGTATCAATATTTCTCAATATTCTATTATTCCAG	1684
AGCTCTGTGGCAGGAATGGGATGCAGACCAAGTTAATATTTCTTTTATAAGTCACAGTATCAATATTTCTCAATATTCTATTATTCCAG	1613
TCTCCATGAGGTAACCAAGTAACACCGGTGTGTTTTCTACCATGCCCTTCTCCATTTATGGCATGATTTCTTACACTTTTGTAAATAG	1774
TCTCCATGAGGTAACCAAGTAACACCGGTGTGTTTTCTACCATGCCCTTCTCCATTTATGGCATGATTTCTTACACTTTTGTAAATAG	1703
TGACGGGTACGCAGGCCTATCAACCATTGGCTGGCATCCAGGTGGGCACCTTCATCAAGGGATAACTGTAAATGAGCAACCCCTTGGTGG	1864
TGACGGGTACGCAGGCCTATCAACCATTGGCTGGCATCCAGGTGGGCACCTTCATCAAGGGATAACTGTAAATGAGCAACCCCTTGGTGG	1793

1986

CAGGTCACCTTACTTTAGCCCTATCTCATCCAGTTTGTCTTAAGCACCATGAAAGCAAATCTCCCTTTGTCCTTCACACTTCCCACAA	3390
CAAGTCACCTTACTTTAGCCCTATCTCATCCAGTTTGTCTTAAGCACCATGAAAGCAAATCTCCCTTTGTCCTTCACACTTCCCACAA	3322
TTCTGGCACAAGGAGAGGTCAGAAATATTGGAGGAAGGAAGGAATGAAGTTCCCATGACTGGAGCATCTGTAGAGTCTGAGATT	3480
TTCTGGCACAAGGAGAGGTCAGAAATATTGGAGGAAGGAAGGAATGAAGTTCCCATGACTGGAGCATCTGTAGAGTCTGAGATT	3412
AATCTGGATTCTTGTCTAATCTTCTCCTCACGGCATCCTTACCTTCATCTCCACCCACCATCACTGCTCTCCCTCTACTGGCGAAA	3570
AATCTGGATTCTTGTCTAATCTTCTCCTCACGGCATCCTTACCTTCATCTCCACCCACCATCACTGCTCTCCCTCTACTGGCGAAA	3502
GTAGAATTTCCATCATCGAGTTTTCAGCTCAGTGGTGGGAGAGGTCCTTTTCATGAACGAAACCTCCTCTCACATTGATTTGAAGGCTG	3660
GTAGAATTTCCATCATCGAGTTTTCAGCTCAGTGGTGGGAGAGGTCCTTTTCATG-ACGAAGCCTCCTCTCACATTGATTTGAAGGCTG	3591
TGGCTTCAAAGAGTCTGGCCTTATCTTTAATAAATTCATATTTTAATTAACTAACTGGAGTGGATTGTGTGTTTGCACCTAAGAACC	3750
TGGCTTCAAAGAGTCTGGCCTTATCTTTAATAAATTCATATTTTAATTAACTAACTGGAGTGGATTGTGTGTTTGCAGCTAAGAAC-	3680
TTAACCCATAGGTTCCAT-GGAAACGGTGGTCTTTCTCATTTTATGCAGATGGGTGGGCAGCTCTCCATCACCTCTCCTCAGACTCAGCC	3839
-TAAC-----CCATAGGAAACGGTGGTCTTTCTCATTTTATGCAGATGGGTGGGCAGCTCCCCATCACCTCTCCTCAGACTCAGCC	3760
CTACCAAGTAGAAGGAGCCAAACCCCTTACACTGACATCTACCTCTTATGGCCGTGCCAGTGTACATGAAAACTGGATGAGAGACACCTC	3929
CTACCAAGTAGAAGGAGCCAAACCCCTTACACTGACATCTACCTCTTATGGCCGTGCCAGTGTACATGAAAACTGGATGAGAGACACCTC	3850
AACAAGAAAACCTTTTGTCTTCACTTCTTGGGCCAGGTCAAACCTTTGGGGTGTGTTATTTCCCTGAATT	3998
AACAAGAAAACCTTTTGTCTTCACTTCTTGGGCCAGGTCAAACCTTTGGGGTGTGTTATTTCCCTGAATT	3912

Fig. 1. Nucleotide sequences of two 4 kb segments of the bovine genome containing genes for BPTI and spleen inhibitor II

The upper sequence contains the BPTI coding region and the lower one contains the spleen inhibitor II coding region. The sequences are aligned to maximize homology and dashes (–) indicate where compensatory gaps have been introduced. A colon (:) indicates where the sequences are identical. The derived amino acid sequences spanning the inhibitor genes are shown and the portion of those sequences which match the amino acid sequences of BPTI and spleen inhibitor II (SI) are indicated. Amino acids which differ between the two sequences are boxed. Asterisks indicate in-frame termination codons. Arrows indicate potential intron/exon (Acceptor) and exon/intron (Donor) splice junctions and a potential splice signal is overlined (Keller & Noon, 1984; Ruskin *et al.*, 1984; Zeitlin & Efstratiadis, 1984). The boxed nucleotide sequences are potential polyadenylation signals (Proudfoot & Brownlee, 1976; McLauchlan *et al.*, 1985).

Fig. 2. Northern blots of bovine RNA

Samples of RNA were denatured using glyoxal and fractionated by electrophoresis on a 1.1% agarose gel. The probe corresponded to the complement of the sequence from nucleotides 2344 to 2652 in the sequence containing BPTI (Fig. 1). The positions of the 28S and 18S rRNA

2691. Both probes gave the same result and a major band can be seen to be present in RNA isolated from bovine lung and also, more weakly, in RNA isolated from bovine parotid gland. The size of this mRNA, estimated from the positions of the 18S and 28S rRNA bands, was approx. 800 nucleotides. Southern blot analysis of oligo(dT)-primed first-strand cDNA from lung mRNA after electrophoresis in the presence of labelled restriction fragment markers supports this estimation (I. B. Kingston, unpublished work). There appeared to be only a faint trace of this mRNA component present in liver RNA.

Southern blots

When Southern blots of bovine genomic DNA were hybridized under conditions of low stringency to a BPTI-specific probe, 16 or more restriction fragments were found to hybridize (Fig. 3). This suggests that the bovine genome may contain a substantial family of genes that are related to the BPTI gene. This is in agreement with the protein data of Fioretti *et al.* (1983), who have isolated four different Kunitz-type inhibitors from bovine

bands are indicated. (a) Approx. 10 μ g of total RNA from bovine lung (track 1), liver (track 2) and parotid gland (track 3); the blot was washed at room temperature. (b) Poly(A)-containing RNA (2 μ g) from bovine lung (track 1) and 10 μ g of total RNA from bovine lung (track 2); the blot was washed at 65 °C.

Fig. 3. Genomic Southern blot of bovine DNA

(a) *Eco*RI, (b) *Hind*III and (c) *Bam*HI restriction digests of bovine DNA were fractionated by electrophoresis on a 0.5% agarose gel and blotted onto nitrocellulose. The 32 P-labelled probe was prepared from an M13mp8 clone which contained the sequence from nucleotides 2344 to 2652 in the BPTI-containing fragment (Fig. 1). The numbers on the left are the lengths (kb) of restriction fragments which are present in this blot and were previously detected by hybridization to a synthetic BPTI probe (Anderson & Kingston, 1983).

spleen and with the occurrence of Kunitz-type inhibitors in bovine serum (Wachter *et al.*, 1980; Hochstrasser *et al.*, 1978) and bovine colostrum (Laskowski *et al.*, 1952; Cechova *et al.*, 1971; Jonakova & Cechova, 1977).

DISCUSSION

In this paper we present the sequence of two *Eco*RI fragments from the bovine genome which contain coding regions that correspond to the amino acid sequence of BPTI and bovine spleen inhibitor II. Each coding region is present in an almost identical genetic environment. The fragment containing the BPTI coding region is 3998 nucleotides long and that containing spleen inhibitor II is 3919 nucleotides long. Neither gene is preceded by an initiation codon or followed by a termination codon but rather each is embedded in a longer open reading frame. Each gene is however flanked by sequences which resemble intron/exon and exon/intron boundary sequences and in addition both have a potential splice signal

(Keller & Noon, 1984; Ruskin *et al.*, 1984; Zeitlin & Efstratiadis, 1984) present 28 and 46 nucleotides upstream from the two potential intron/exon splice junctions. It is therefore likely, as was proposed earlier for the BPTI gene (Anderson & Kingston, 1983), that both trypsin inhibitor coding regions exist as part of an exon in the gene coding for a longer polypeptide. The nature of this precursor is as yet unknown, although cDNA analysis is currently in progress. Each gene is present only once in the haploid bovine genome (Anderson & Kingston, 1983) and neither of the sequences have any of the characteristics of pseudogenes such as being flanked by short direct repeats or possessing poly(A) tracts at the 3' end. Searching each of the 4 kb sequences for other potential coding sequences failed to reveal any obvious ones, although clearly this does not rule out the possibility that they are present since there are numerous open reading frames. In addition, very short exons would probably not be picked up by the gene search programs.

The probability that BPTI and spleen inhibitor II are released from larger precursors is not inconsistent with the Northern blot results (Fig. 3) in which BPTI probes hybridized to a mRNA isolated from bovine lung of approx. 800 nucleotides. Such a precursor for members of the BPTI family of inhibitors has a precedent in inter- α -trypsin inhibitor. This protein (M_r 160000) occurs in both man and cattle and has at its C-terminus two BPTI homologues which are released *in vivo* and may also be released *in vitro* by limited proteolysis (Hochstrasser *et al.*, 1976, 1978; Wachter & Hochstrasser, 1979). The Northern blot results suggest that the lung is a major site of synthesis of the BPTI precursor (Fig. 2).

A striking feature of the two sequences presented in Fig. 1 is their dramatic similarity. Over approx. 3.4 kb they differ by only about 100 nucleotides (2.9%). The distribution of substitutions between the two sequences appears to be random throughout the sequence, but it is notable that there are no substitutions in the spleen inhibitor coding sequence which are likely to disrupt the BPTI-like conformation (Creighton, 1975). An interesting feature of the variant bases in the protein coding regions is that some of them occur in codons for contact site residues in the inhibitor-enzyme complex. Thus in the BPTI-trypsin complex, BPTI residues Arg-17, Ile-18 and Arg-39 are in close contact with trypsin (Huber *et al.*, 1974b). In the spleen inhibitor, these residues are replaced by Lys, Met and Lys respectively. Although these are conservative changes, they may affect the interaction of the inhibitor with various proteinases (Wachter *et al.*, 1980), and Fioretti *et al.* (1985) have observed that the affinity constant of spleen inhibitor II for porcine kallikrein is about one-third that of BPTI. Similar effects have been observed with variants of avian ovomucoid third domain and serine proteinases, where amino acid substitutions at contact sites are responsible for the variants having optimal inhibitory effects upon different serine proteinases (Empie & Laskowski, 1982). As this Kunitz family of inhibitors to which BPTI and the spleen inhibitor belong inhibit kallikreins, one may speculate for example, that each may inhibit optimally a different kallikrein, as the members of this family of serine proteinases appear to differ mainly at residues which are implicated in substrate specificity (Mason *et al.*, 1983).

As has been stated above, the distribution of substitutions between the two sequences appears to be

random and therefore the putative exons and introns of this pair of sequences appear to have had very similar rates of accumulation of base changes. It is also apparent that of the nine differences which exist between the two coding regions, seven are 'replacement' changes and only two are 'silent' changes (with no corresponding amino acid change). Although we are analysing only a short sequence, these data indicate a situation different to that found in some 50 other homologous gene pairs, where the rates of intron and silent position divergence are several times greater than the rate of replacement changes (Miyata *et al.*, 1980; Perler *et al.*, 1980). A high rate of replacement compared with neutral changes has been reported in four other gene pairs (Efstratiadis *et al.*, 1977; Bernard *et al.*, 1978; Hardison *et al.*, 1979; van Ooyen *et al.*, 1979; Bregegere *et al.*, 1981; Talmadge *et al.*, 1984) and it has been suggested that this situation may reflect the effects of evolutionary selection (Perler *et al.*, 1980; Talmadge *et al.*, 1984).

The similarity between the two sequences suggests that this segment of the bovine genome has been duplicated. The extent of the similarity between these two sequences however suggests that either this duplication was a comparatively recent event or that gene conversion has rendered the two sequences more alike than would be expected from the time elapsed since the original duplication event. The latter appears to have happened in the human foetal globin gene where initially a 5 kb region was duplicated and then a subsequent gene conversion replaced part of the diverging γ -gene with the corresponding part of the δ -gene (Slightom *et al.*, 1980; Shen *et al.*, 1981). In the foetal globin gene a sequence of repeated dinucleotides, (TG)_n, occurs at the boundary of the conversion, but in the sequences presented here, the 3' end of the potentially converted region was not obtained. For the foetal globin gene there were no sequence peculiarities close to the putative 5' boundary of the conversion as notable as the repeated TG dinucleotides found at the 3' boundary, but the sequence TCAAAAAT occurred at the 5' boundary and was repeated within the converted region at a distance of 51 nucleotides (Shen *et al.*, 1981). No such sequence was found in the present work, though two sequences close to the boundary between the conserved and non-conserved regions (nucleotide 621 in the BPTI-containing sequence) are also present a short distance upstream. In the sequence containing BPTI, ATTTTATTTT occurs almost at the boundary and is also present 86 nucleotides upstream; in the corresponding position, the sequence containing spleen inhibitor II has TTGAAGT at the boundary and a similar sequence 86 nucleotides upstream. The two segments contain other direct repeats, the longest in both being of 11 nucleotides. In the sequence containing BPTI, there is only one such repeat and this occurs at positions 837 and 1035. In the sequence containing spleen inhibitor II, there are five and they occur at positions 342 and 3178, 204 and 1313, 267 and 1757, 160 and 234 and 366 and 1059. Both sequences contain one region which has some similarity with part of the 120 bp bovine *Alu*-like consensus sequence (Watanabe *et al.*, 1982). In the case of the BPTI-containing sequence, this is a stretch of 16 bases, AGGAGCCTGG-TCTGTT, which starts at position 930 and matches a 16 base stretch in the *Alu*-like consensus sequence at 14 out of the 16 positions. In the sequence containing spleen inhibitor II, a similar sequence occurs at position 857.

We are grateful to Drs. A. Bankier, T. Gibson, M. Biggin, and R. Staden for discussions and advice and Dr. F. Sanger for encouragement and support. S.A. was financially supported by a National Research Service Award from the National Institute of General Medical Sciences.

REFERENCES

- Anderson, S. & Kingston, I. B. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 6839–6842
- Bankier, A. T. & Barrell, B. G. (1983) in *Techniques in Nucleic Acid Biochemistry* (Flavell, R. A., ed.), pp. B508/1–B508/31, Elsevier Scientific Publishers, Ireland
- Bernard, O., Hozumi, N. & Tonegawa, S. (1978) *Cell* **15**, 1133–1144
- Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 3963–3965
- Bregegere, F., Abastado, J. P., Kvist, S., Rask, L., Lalanne, J. L., Garoff, H., Cami, B., Wilman, K., Larhammar, D., Peterson, P. A., Gachelin, G., Kourlisky, P. & Dobberstein, B. (1981) *Nature (London)* **292**, 78–81
- Cechova, D., Jonakova, V. & Sorm, F. (1971) *Collect. Czechoslov. Chem. Commun.* **36**, 3342–3357
- Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. & Rutter, W. J. (1979) *Biochemistry* **18**, 5294–5299
- Creighton, T. E. (1975) *Nature (London)* **255**, 743–745
- Creighton, T. E. (1978) *Prog. Biophys. Mol. Biol.* **33**, 231–297
- Deininger, P. L. (1983) *Anal. Biochem.* **129**, 216–223
- Deisenhofer, J. & Steigemann, W. (1975) *Acta. Crystallogr. Sect. B* **31**, 238–250
- Doolittle, R. F. (1981) *Science* **214**, 149–150
- Efstratiadis, A., Kafatos, F. C. & Maniatis, T. (1977) *Cell* **10**, 571–585
- Empie, M. W. & Laskowski, M. (1982) *Biochemistry* **21**, 2274–2284
- Fioretti, E., Binotti, I., Barra, D., Citro, G., Ascoli, F. & Antonini, E. (1983) *Eur. J. Biochem.* **130**, 13–18
- Fioretti, E., Iacopino, G., Angletti, M., Barra, D., Bossa, F. & Ascoli, F. (1985) *J. Biol. Chem.*, in the press
- Fritz, H., Fink, E. & Truscheit, E. (1979) *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **38**, 2752–2759
- Girvitz, S. C., Bacchetti, S., Rainbow, A. J. & Graham, F. L. (1980) *Anal. Biochem.* **106**, 492–496
- Hardison, R. C., Butler, E. T., Lacy, E., Maniatis, T., Rosenthal, N. & Efstratiadis, A. (1979) *Cell* **18**, 1285–1297
- Hochstrasser, K., Wachter, E. & Bretzel, G. (1976) *Hoppe-Seyler's Z. Physiol. Chem.* **357**, 1659–1661
- Hochstrasser, K., Wachter, E. & Bretzel, G. (1978) in *Proc. 11th FEBS Meet. Regulatory Proteolytic Enzymes and their Inhibitors* (Magnusson, S., Ottesen, M., Foltmann, B., Dano, K. & Neurath, H., eds.), pp. 225–234, Pergamon Press, Oxford
- Holmes, D. S. & Quigley, M. (1981) *Anal. Biochem.* **114**, 193–197
- Huber, R., Kukla, D., Bode, W., Schwager, P., Bartels, K., Deisenhofer, J. & Steigemann, W. (1974a) *J. Mol. Biol.* **89**, 73–101
- Huber, R., Kukla, D., Steigemann, W., Deisenhofer, J. & Jones, A. (1974b) in *Protease Inhibitors: Bayer Symp. 5* (Fritz, H., Tschesche, H., Green, L. J. & Truscheit, E., eds.), pp. 497–512, Springer-Verlag, New York
- Huber, R., Bode, W., Kukla, D., Kohl, U. & Ryan, C. A. (1975) *Biophys. Struct. Mechanism* **1**, 189–201
- Jonakova, V. & Cechova, D. (1977) *Collect. Czechoslov. Chem. Commun.* **42**, 759–769
- Karn, J., Brenner, S., Barnett, L. & Cesareni, G. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 5172–5176
- Karplus, M. & McCammon, J. A. (1981) *Crit. Rev. Biochem.* **9**, 293–349
- Kassell, B., Radicevic, M., Ansfield, M. J. & Laskowski, M., Sr. (1965) *Biochem. Biophys. Res. Commun.* **18**, 255–258

- Keller, E. B. & Noon, W. A. (1984) *Proc. Natl. Acad. Sci. U.S.A.* **81**, 7417–7420
- Kunitz, M. (1947) *J. Gen. Physiol.* **30**, 311–320
- Laskowski, M., Jr., Mars, P. H. & Laskowski, M., Sr. (1952) *J. Biol. Chem.* **198**, 745–752
- Laskowski, M., Jr. & Kato, I. (1980) *Annu. Rev. Biochem.* **49**, 593–626
- Lipman, D. J. & Pearson, W. R. (1985) *Science* **277**, 1435–1441
- McLauchlan, J., Gaffney, D., Whitton, J. L. & Clements, J. B. (1985) *Nucleic Acids Res.* **13**, 1347–1368
- Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning*, pp. 168–169, Cold Spring Harbor Laboratory, New York
- Mason, A. J., Evans, B. A., Cox, D. R., Shine, J. & Richards, R. I. (1983) *Nature (London)* **303**, 300–307
- Miyata, T., Yasunaga, T. & Nishida, T. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 7328–7332
- Mount, S. M. (1982) *Nucleic Acids Res.* **10**, 459–472
- Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. & Dodgson, J. (1980) *Cell* **20**, 555–566
- Proudfoot, N. J. & Brownlee, G. G. (1976) *Nature (London)* **263**, 211–214
- Ruhlmann, A., Kukla, D., Schwager, P., Bartels, K. & Huber, R. (1973) *J. Mol. Biol.* **77**, 417–436
- Ruskin, B., Krainer, A. R., Maniatis, T. & Green, M. R. (1984) *Cell* **38**, 317–331
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463–5467
- Shen, S., Slightom, J. L. & Smithies, O. (1981) *Cell* **26**, 191–203
- Slightom, J. L., Blechl, A. E. & Smithies, O. (1980) *Cell* **21**, 627–638
- Staden, R. (1982a) *Nucleic Acids Res.* **10**, 4731–4751
- Staden, R. (1982b) *Nucleic Acids Res.* **10**, 2951–2961
- Staden, R. (1984a) *Nucleic Acids Res.* **12**, 521–538
- Staden, R. (1984b) *Nucleic Acids Res.* **12**, 551–567
- Talmadge, K., Vamvakopoulos, N. C. & Fiddes, J. C. (1984) *Nature (London)* **307**, 37–40
- Thomas, P. S. (1983) *Methods Enzymol.* **100**, 255–266
- Van Ooyen, A., Van den Berg, J., Mantei, N. & Weissmann, C. (1979) *Science* **206**, 337–344
- Vieira, J. & Messing, J. (1982) *Gene* **19**, 259–268
- Wachter, E. & Hochstrasser, K. (1979) *Hoppe-Seyler's Z. Physiol. Chem.* **306**, 1305–1311
- Wachter, E., Deppner, K., Hochstrasser, K., Lempart, K. & Geiger, R. (1980) *FEBS Lett.* **119**, 58–62
- Wagner, G. & Wuthrich, K. (1982) *J. Mol. Biol.* **155**, 347–366
- Watanabe, Y., Tsukada, T., Notake, M., Nakanishi, S. & Numa, S. (1982) *Nucleic Acids Res.* **10**, 1459–1469
- Wilbur, W. J. & Lipman, D. J. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 726–730
- Zeitlin, S. & Efstratiadis, A. (1984) *Cell* **39**, 589–602

Received 29 July 1985/2 September 1985; accepted 13 September 1985