Biochem. J. (1989) 260, 249–258 (Printed in Great Britain)

DNA sequences of a bovine gene and of two related pseudogenes for the proteolipid subunit of mitochondrial ATP synthase

Mark R. DYER, Nicholas J. GAY* and John E. WALKER†
Medical Research Council Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, U.K.

The dicyclohexylcarbodi-imide-reactive proteolipid is a membrane subunit of mitochondrial ATP synthase. In cows it is encoded by two different nuclear genes known as P1 and P2. These genes are expressed in a tissue-specific fashion which reflects the embryonic origin of the tissues. The proteins that they encode are synthesized in the cytosol, and are precursors of the proteolipid that have different mitochondrial import sequences of 61 and 68 amino acids respectively. By use of gene-specific probes derived from the bovine P2 cDNA, regions containing corresponding parts of the bovine P2 gene have been isolated from a bovine genomic library, and their DNA sequences and those of flanking and intervening regions have been determined. The sequence contains four exons, which represent the cDNA sequence, spread over 3.8 kb of the bovine genome. Two of the introns are in the DNA sequence coding for the mitochondrial import sequence, and a third intron is in a sequence encoding an extramembranous structure between the two putative transmembrane α-helical domains of the mature proteolipid. An *Alu*-type repetitive element was detected at the extreme 5′ end of the sequence. The bovine P1 and P2 genes for the dicyclohexylcarbodi-imide-reactive proteolipid of ATP synthase are members of a multiple gene family that also contains many pseudogenes. The bovine P1 gene has not been isolated, but two distinct P1 pseudogenes have been cloned and their DNA sequences have been determined. Both of them contain 'in-phase' stop codons and frame-shift mutations, and one of them bears the hallmarks of retroposition; it has no introns, it contains a poly(A) tract at its 3′ end and it is flanked by direct DNA sequence repeats. The second P1 pseudogene is very unusual. It appears to be derived from a partially processed transcript and contains an intervening DNA sequence of 861 bp that corresponds in position with an intron in the human P1 gene. This pseudogene also could have been introduced by retroposition since its sequence is flanked by short direct repeats. However, it does not contain a poly(A) tract at its 3′ end. An alternative, but less likely, explanation is that rather than being a retroposon, this sequence arose by duplication of an expressed gene at a time when it had only one intron.

INTRODUCTION

The dicyclohexylcarbodi-imide (DCCD)-reactive proteolipid is an essential membrane protein component of the proton channel of bovine mitochondrial ATP synthase. It is 75 amino acids long (Sebald & Hoppe, 1981), and is a nuclear gene product, as are all but thirteen bovine mitochondrial proteins (Anderson *et al.*, 1982). Import into the organelle of these nuclear coded proteins is usually directed by an *N*-terminal extension known as the mitochondrial import sequence. This is removed during entry into the mitochondrion (Schatz & Butow, 1983). The DCCD-proteolipid is unique among mitochondrial proteins so far investigated in having two different, but weakly homologous, import presequences, which, when removed from the precursors, produce an identical mature protein. The precursors are the products of two different genes, P1 and P2, that are expressed in different ratios in various bovine tissues (Gay & Walker, 1985). In the present paper the characterization of the bovine P2 gene is presented. It is split into at least four exons, as are the human homologues (M. R. Dyer & J. E. Walker, unpublished work), and their sequence is spread over about 4 kb of bovine DNA. This gene is a member of a multigene family which includes at least two expressed genes and several pseudogenes. The sequences of two related pseudogenes are also presented. One of them has no intervening sequences and appears to be a retroposon; the other is a partly spliced pseudogene.

MATERIALS AND METHODS

Preparation of bovine DNA and genomic libraries

The preparation of bovine liver DNA has been described previously (Walker *et al.*, 1987). A phage library of partial *Sau*3AI fragments of bovine genomic DNA was made in λ2001 (Karn *et al.*, 1984).

DNA hybridization

Digests of DNA were fractionated by electrophoresis in 0.6% agarose gels, and fragments were transferred and fixed to nitrocellulose as described by Southern (1975). After transfer, the nitrocellulose filters were incubated at 65 °C for 1 h in a solution containing 6 × SSC (1 × SSC is 0.15 M-NaCl/0.015 M-trisodium

Abbreviation used: DCCD, dicyclohexylcarbodi-imide.

* Present address: Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1QW, U.K.
† To whom correspondence and reprint requests should be addressed.

These sequence data have been submitted to the EMBL/GenBank data libraries under the accession number X14203.

citrate), 0.2% bovine serum albumin (fraction V), 0.2% Ficoll, 0.2% polyvinylpyrrolidone, 0.5% *N*-laurylsarcosine and sonicated salmon testis DNA (100 μg/ml). Then filters were hybridized for 15–20 h at the same temperature in the presence of radioactive 'prime–cut' probes (Farrell *et al.*, 1983) dissolved in the same solution as used for pre-hybridization, except that it contained also 10% dextran sulphate. Subsequently, the filters were washed four times for 30 min at 65 °C in either 0.2 or 2 × SSC, each containing 0.5% laurylsarcosine. Autoradiographs of filters were exposed for 1–3 h at −70 °C in the presence of an intensifying screen.

Screening the genomic library

Plaques (about 10^6) were produced on *Escherichia coli* Q358 grown on 20 cm diameter agar plates, and were screened by the plaque hybridization method (Benton & Davis, 1977). Phage from each plate were transferred to two nitrocellulose filters placed sequentially on the agar. The preparation of 'prime–cut' probes and the hybridization conditions employed were the same as those described above. Recombinant phages were grown on *E. coli* Q358 in 500 ml cultures, and DNA was prepared from them according to Maniatis *et al.* (1982).

Identification and manipulation of genomic clones

Hybridization studies indicated that the recombinant λP4.21 contained the bovine P2 gene. From its DNA were excised a 5.3 kb *Nco*I fragment and an overlapping 2.8 kb *Xba*I fragment. They were purified by electrophoresis in low melting point agarose, and then broken up by sonication. The resultant fragments were fractionated by electrophoresis, and those that were greater than about 500 bp were cloned into the *Sma*I site of M13mp8 (Deininger, 1983). Two other recombinants, λP3.9 and λP3.17, contained sequences related to the bovine P1 cDNA. By sequence analysis it seemed that they both contained pseudogenes. From λP3.9 and λP3.17 respectively, a *Sac*I fragment of 5.2 kb and a 4.0 kb *Eco*RI fragment were purified, and libraries of random fragments produced by sonication were prepared from them as above.

DNA sequence analysis

DNA sequences were determined by the dideoxy chain termination method (Sanger *et al.*, 1977) as modified by Biggin *et al.* (1983) and a random strategy was employed. All sequences were determined minimally at least once in both senses of the DNA. This required that both clone turn-arounds and some long-runs be performed on selected clones. 'Compressed' sequences were resolved by substituting deoxyinosine triphosphate for deoxyGTP in the appropriate sequencing reactions (Mills & Kramer, 1979). DNA sequences were compiled with the help of the computer programs DBUTIL and DBAUTO (Staden, 1982), and were analysed with ANALYSEQ (Staden, 1985).

RESULTS AND DISCUSSION

Gene cloning

About 10^6 recombinants in the bovine genomic library were screened with probes derived from the 3′ non-coding regions of the bovine P1 and P2 cDNAs. In contrast to the coding regions, these sequences are poorly conserved and would be expected to distinguish between the P1 and P2 genes. The probes employed were nucleotides 404–558 and 406–615 respectively of P1 and P2 cDNAs (Gay & Walker, 1985). After re-screening of initial isolates, 20 recombinants, λP3.1–λP3.20, were selected which hybridized with the P1 probe, and also 21 recombinants, λP4.1–λP4.21, which hybridized with the P2 probe. DNA was prepared from each isolate, and further restriction analysis and hybridization experiments showed that the P1 isolates appeared to fall into three related groups, and the P2 isolates could be put in four groups. Each group contained overlapping, but non-identical, DNA inserts. From the P1 isolates, λP3.9 and λP3.17 were subsequently found to contain different pseudogenes. Another isolate, λP3.3, which is not described in this paper, appears to contain part of exon IV of the P1 gene and a flanking sequence. However, part of the clone seems to result from a re-ligation event with unrelated DNA (N. J. Gay, unpublished work). No further attempts have been made to find other clones containing the bovine P1 gene. Restriction fragments in recombinant λP4.21 hybridized to the P2 probe derived from the non-coding region of the bovine cDNA, and other restriction fragments hybridized also to a second probe taken from the coding region. From these data it was deduced that this recombinant contained the regions of the expressed gene that correspond to the P2 cDNA, as was proved subsequently by sequence analysis.

DNA sequence of the bovine P2 gene

A digest with *Nco*I of DNA from λP4.21 was fractionated in a 0.6% agarose gel, and the fragments were hybridized to a 'prime–cut' probe containing nucleotides 120–409 of the cDNA sequence of bovine P2 (Gay & Walker, 1985). As shown in Fig. 1, this revealed a 5.3 kb *Nco*I restriction fragment, suggesting that most of the DNA sequence in the probe is within this fragment. This conclusion is based on the assumption that the *Nco*I site within the cDNA sequence was also present in genomic DNA, and that it had not been created by splicing of the primary transcript. The DNA sequences encoding the 15 *C*-terminal amino acids of the protein and 3′ untranslated region of the mRNA were not expected to be found in this restriction fragment. The DNA sequence of this fragment (see Fig. 2) contained the 5′ non-coding region present in the cDNA, and sequences representing the coding sequence of the cDNA up to and including the first base of the codon for amino acid 61 of the mature protein; as anticipated, the remainder of the 3′ end of the gene was absent. Therefore, in order to extend this sequence, an *Xba*I restriction fragment derived from λP4.21 DNA that overlaps the 3′ region of the *Nco*I fragment was identified by hybridization to a 'prime–cut' probe derived from the 3′ region of the P2 cDNA. This *Xba*I fragment was sequenced completely and corresponds to nucleotides 4831–7647 in Fig. 2. The extended sequence contained the missing 3′ region encoding 15 amino acids at the *C*-terminal end of the proteolipid, and the 3′ untranslated region present in the mRNA.

Each nucleotide in the sequence of the bovine P2 gene was determined six times on average, and at least once on each strand. The G + C content of this 7647 bp segment is 44%, in reasonable agreement with the estimated G + C content of 42% for the bovine genome (Chargaff & Lipshitz, 1953).

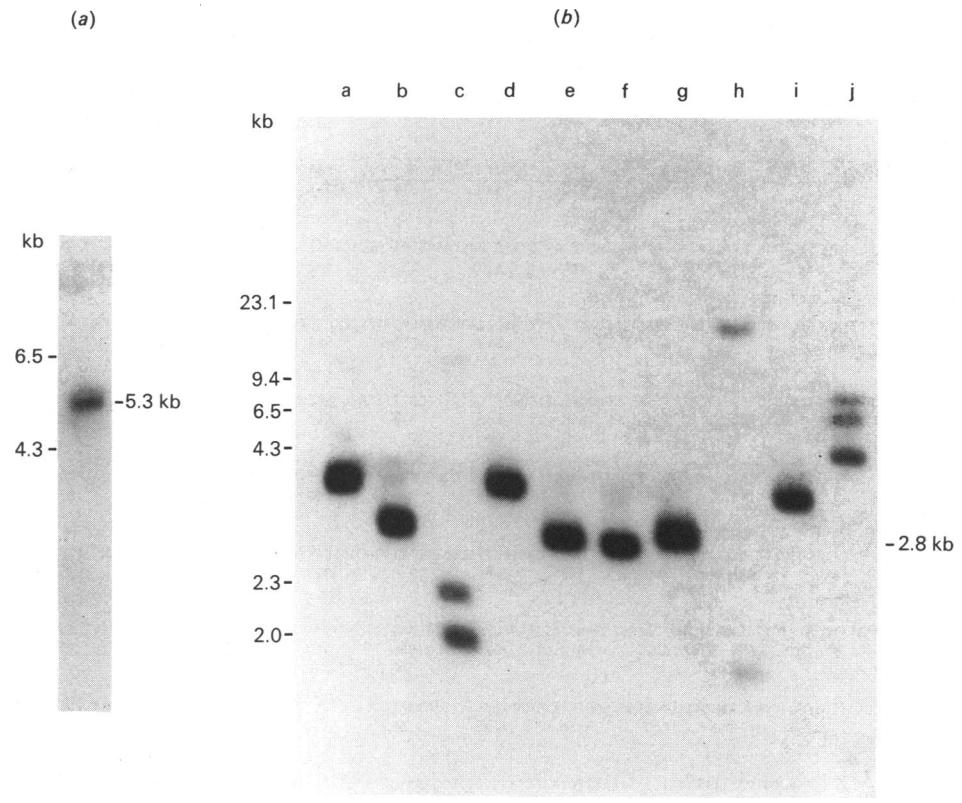


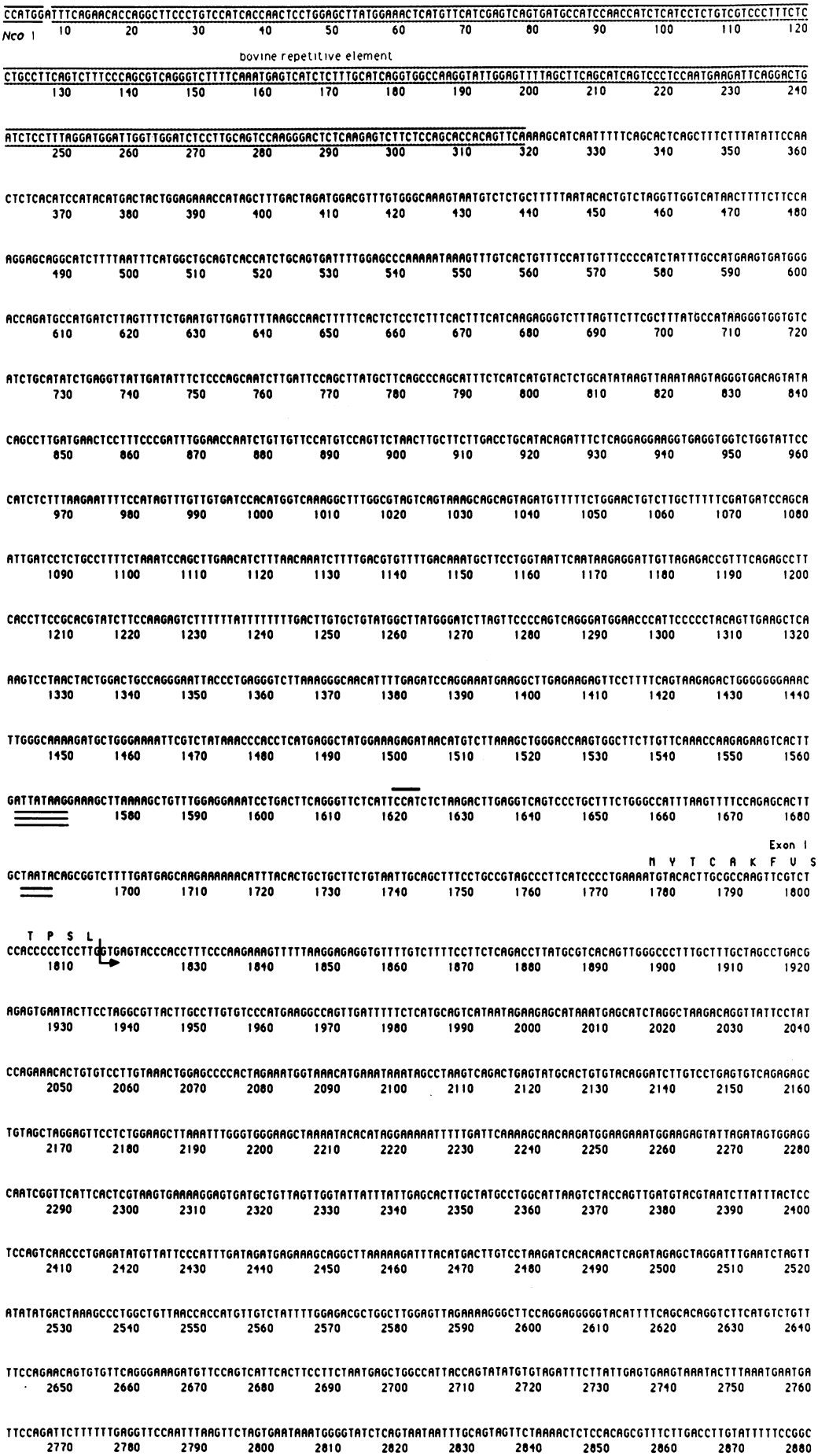
Fig. 1. Hybridizations of bovine DNA in a recombinant phage with probes for the bovine P2 gene

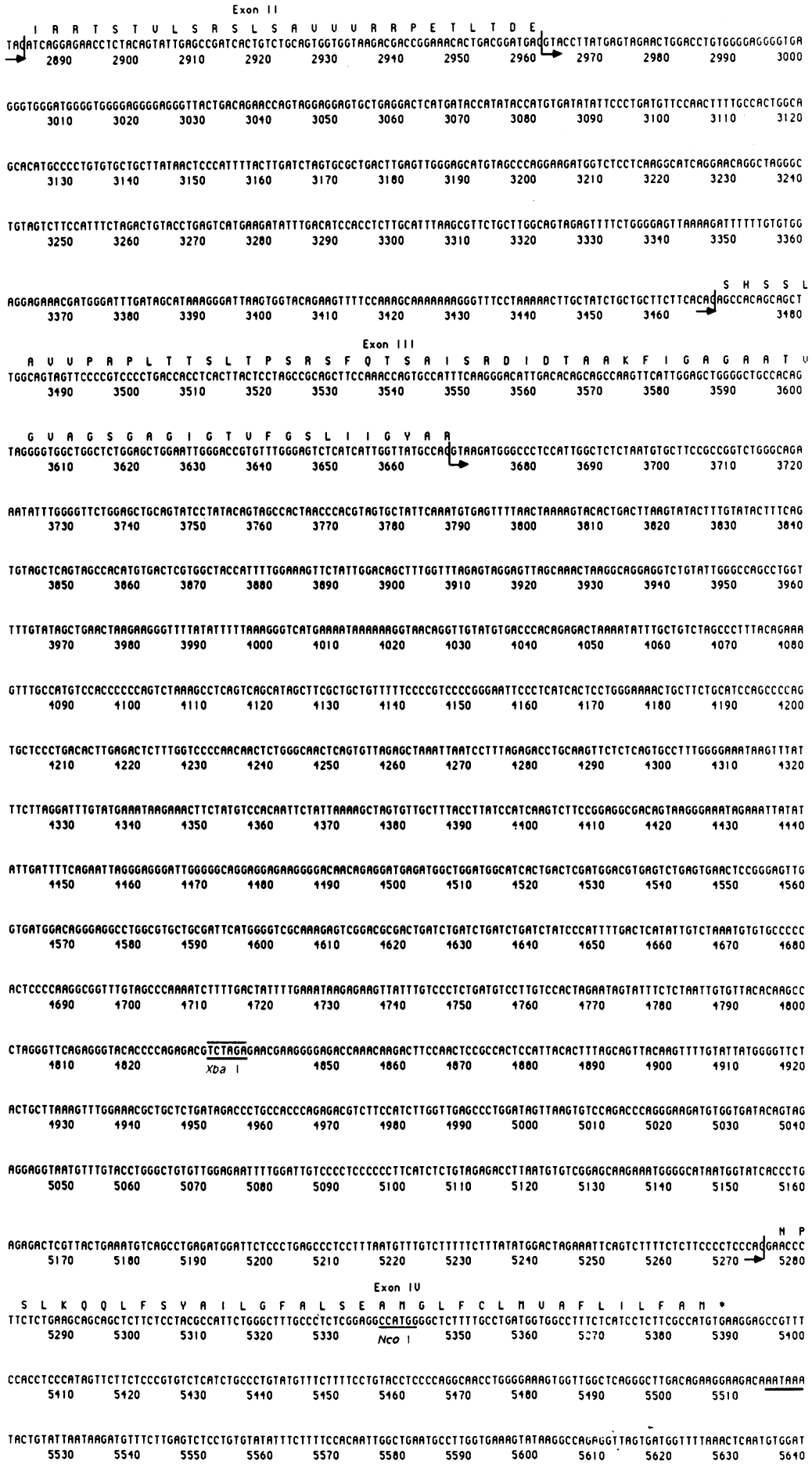
In panel (*a*), DNA from recombinant λP4.21 (0.5 μg of DNA/digest) was restricted with *Nco*I, and in panel (*b*) DNA (0.5 μg of DNA/digest) from the same phage was digested with *Sac*I and *Eco*RI (lane a), *Eco*RI and *Xba*I (lane b), *Stu*I and *Eco*RI (lane c), *Nci*I and *Sac*I (lane d), *Aha*I and *Sac*I (lane e), *Xba*I (lane f), *Taq*I (lane g), *Stu*I (lane h), *Nci*I (lane i), and *Aha*II (lane j). The DNA digests were fractionated on a 0.6% agarose gel and the restriction fragments were transferred to nitrocellulose. In panels (*a*) and (*b*), the DNA fragments were hybridized with 'prime–cut' probes containing nucleotides 120–409 and 406–615 respectively from the bovine P2 cDNA (Gay & Walker, 1985). Both of these blots were washed at 65 °C in 0.2 × SSC, and then autoradiographed for 1 h at −70 °C using Fuji X-ray film. In panels (*a*) and (*b*), bands that correspond to a 5.3 kb *Nco*I and 2.8 kb *Xba*I restriction fragments, respectively, are labelled.

Gene structure

Comparison of the sequence with that of the bovine P2 cDNA showed that the gene is split into at least four exons (see Fig. 3), and the sequences of the exons together agree exactly with that of the bovine P2 cDNA (Gay & Walker, 1985). The exons are in precisely the same positions as those in the human P1 and P2 genes (M. R. Dyer & J. E. Walker, unpublished work). The exact location of the 5′ end of exon I is not known, as the transcriptional start site(s) of bovine P2 has (have) not been determined experimentally. However, eukaryotic promoters contain DNA sequences which regulate the rate of transcription by RNA polymerase II, and these elements are usually located within a 100 bp to the 5′ side of the transcriptional initiation site (Cochran & Weissmann, 1984; McKnight & Kingsbury, 1982). Many eukaryotic promoters contain a TATA box, a conserved AT-rich sequence which is centred about 25 bp to the 5′ side of the cap site. The TATA box has been shown to control the precise position of transcriptional initiation in some eukaryotic genes (Grosschedl & Birnstiel, 1980; Benoist & Chambon, 1981; McKnight & Kingsbury, 1982). A second conserved sequence, the CCAAT box, is located 70–90 bp to the 5′ side of the cap site in many genes. Its precise function is unknown but in certain genes it has been shown to bind different cellular factors (McKnight & Tijian, 1986; Dorn *et al.*, 1987). Both of these sequences are present in the region to the 5′ side of the furthest established extent of exon I (see Fig. 2). Nonetheless, the possibility remains that the 5′ non-coding region present in the mRNA is more extensive than that characterized in the bovine P2 cDNA, and that additional intervening sequences are present towards or beyond the 5′ end of the present sequence.

Two of the introns that have been detected are found in sequences encoding the mitochondrial import sequence, and none is found at the boundary between the import sequence and the mature proteolipid. A third intron, however, is at a position which is within the protein sequence ARNP (amino acids 37–40 of the mature protein). This is thought to form a β-turn at the membrane periphery, and to link the two transmembrane α-helices into which the proteolipid is probably folded. The location of introns in genes encoding membrane proteins in sequences that are believed to be extramembranous links between transmembrane segments has been noted also in rhodopsins (Nathans & Hogness, 1984), in the band III protein from mouse red cell membranes (Kopito *et al.*, 1987) and in the mitochondrial


Nco I
bovine repetitive element
Exon I
M Y T C A K F V S
T P S L


Exon II
I A R T S T V L S R S L S A V V V R R P E T L T D E
TAGATCAGGAGAACCTCTACAGTATTGAGCCGATCACTGTCTGCAGTGGTGGTAAGACGACCGGAAAACACTGACGGATGAGGTACCTTATGAGTAGAACTGGACCTGTGGGGAGGGGTGA
2890 2900 2910 2920 2930 2940 2950 2960 2970 2980 2990 3000
GGGTGGGATGGGGTGGGGAGGGGAGGGTTACTGACAGAACCAGTAGGAGGAGTGCTGAGGACTCATGATACCATATACCATGTGATATATTCCCTGATGTTCCAACTTTTGCCACTGGCA
3010 3020 3030 3040 3050 3060 3070 3080 3090 3100 3110 3120
GCACATGCCCCTGTGTGCTGCTTATAACTCCCATTTTACTTGATCTAGTGCGCTGACTTGAGTTGGGAGCATGTAGCCCAGGAAGATGGTCTCCTCAAGGCATCAGGAACAGGCTAGGGC
3130 3140 3150 3160 3170 3180 3190 3200 3210 3220 3230 3240
TGTAGTCTTCCATTTCTAGACTGTACCTGAGTCATGAAGATATTTGACATCCACCTCTTGCATTTAAGCGTTCTGCTTGGCAGTAGAGTTTTCTGGGGAGTTAAAAGATTTTTTGTGTGG
3250 3260 3270 3280 3290 3300 3310 3320 3330 3340 3350 3360
S H S S L
AGGAGAAACGATGGGATTTGATAGCATAAAGGGATTAAGTGGTACAGAAGTTTTCCAAAGCAAAAAAAGGGTTTCCTAAAAACTTGCTATCTGCTGCTTCTTCACAGAGCCACAGCAGCT
3370 3380 3390 3400 3410 3420 3430 3440 3450 3460 3480
Exon III
A V V P R P L T T S L T P S R S F Q T S A I S R D I D T A A K F I G A G A A T V
TGGCAGTAGTTCCCCGTCCCCTGACCACCTCACTTACTCCTAGCCGCAGCTTCCAAACCAGTGCCATTTCAAGGGACATTGACACAGCAGCCAAGTTCATTGGAGCTGGGGCTGCCACAG
3490 3500 3510 3520 3530 3540 3550 3560 3570 3580 3590 3600
G V A G S G A G I G T V F G S L I I G Y A A
TAGGGGTGGCTGGCTCTGGAGCTGGAATTGGGACCGTGTTTGGGAGTCTCATCATTGGTTATGCCAGGTAAGATGGGCCCTCCATTGGCTCTCTAATGTGCTTCCGCCGGTCTGGGCAGA
3610 3620 3630 3640 3650 3660 3680 3690 3700 3710 3720
AATATTTGGGGTTCTGGAGCTGCAGTATCCTATACAGTAGCCACTAACCCACGTAGTGCTATTCAAATGTGAGTTTTAACTAAAAGTACACTGACTTAAGTATACTTTGTATACTTTCAG
3730 3740 3750 3760 3770 3780 3790 3800 3810 3820 3830 3840
TGTAGCTCAGTAGCCACATGTGACTCGTGGCTACCATTTTGGAAAGTTCTATTGGACAGCTTTGGTTTAGAGTAGGAGTTAGCAAACTAAGGCAGGAGGTCTGTATTGGGCCAGCCTGGT
3850 3860 3870 3880 3890 3900 3910 3920 3930 3940 3950 3960
TTTGTATAGCTGAACTAAGAAGGGTTTTATATTTTAAAGGGTCATGAAAATAAAAAAGGTAACAGGTTGTATGTGACCCACAGAGACTAAAATATTTGCTGTCTAGCCCTTTACAGAAA
3970 3980 3990 4000 4010 4020 4030 4040 4050 4060 4070 4080
GTTTGCCATGTCCACCCCCCAGTCTAAAGCCTCAGTCAGCATAGCTTCGCTGCTGTTTTTCCCCGTCCCCGGGAATTCCCTCATCACTCCTGGGAAAACTGCTTCTGCATCCAGCCCCAG
4090 4100 4110 4120 4130 4140 4150 4160 4170 4180 4190 4200
TGCTCCCTGACACTTGAGACTCTTTGGTCCCCAACAACTCTGGGCAACTCAGTGTTAGAGCTAAATTAATCCTTTAGAGACCTGCAAGTTCTCTCAGTGCCTTTGGGGAAATAAGTTTAT
4210 4220 4230 4240 4250 4260 4270 4280 4290 4300 4310 4320
TTCTTAGGATTTGTATGAAATAAGAAACTTCTATGTCCACAATTCTATTAAAAGCTAGTGTTGCTTTACCTTATCCATCAAGTCTTCCGGAGGCGACAGTAAGGGAAATAGAAATTATAT
4330 4340 4350 4360 4370 4380 4390 4400 4410 4420 4430 4440
ATTGATTTTCAGAATTAGGGAGGGATTGGGGGCAGGAGGAGAAGGGGACAACAGAGGATGAGATGGCTGGATGGCATCACTGACTCGATGGACGTGAGTCTGAGTGAACTCCGGGAGTTG
4450 4460 4470 4480 4490 4500 4510 4520 4530 4540 4550 4560
GTGATGGACAGGGAGGCCTGGCGTGCTGCGATTCATGGGGTCGCAAAGAGTCGGACGCGACTGATCTGATCTGATCTGATCTATCCCATTTTGACTCATATTGTCTAAATGTGTGCCCCC
4570 4580 4590 4600 4610 4620 4630 4640 4650 4660 4670 4680
ACTCCCCAAGGCGGTTTGTAGCCCAAAATCTTTTGACTATTTTGAAATAAGAGAAGTTATTTGTCCCTCTGATGTCCTTGTCCACTAGAATAGTATTTCTCTAATTGTGTTACACAAGCC
4690 4700 4710 4720 4730 4740 4750 4760 4770 4780 4790 4800
CTAGGGTTCAGAGGGTACACCCCAGAGACGTCTAGAGAACGAAGGGGAGACCAAACAAGACTTCCAACTCCGCCACTCCATTACACTTTAGCAGTTACAAGTTTTGTATTATGGGGTTCT
4810 4820 Xba I 4850 4860 4870 4880 4890 4900 4910 4920
ACTGCTTAAAGTTTGGAAACGCTGCTCTGATAGACCCTGCCACCCAGAGACGTCTTCCATCTTGGTTGAGCCCTGGATAGTTAAGTGTCCAGACCCAGGGAAGATGTGGTGATACAGTAG
4930 4940 4950 4960 4970 4980 4990 5000 5010 5020 5030 5040
AGGAGGTAATGTTTGTACCTGGGCTGTGTTGGAGAATTTTGGATTGTCCCCTCCCCCCTTCATCTCTGTAGAGACCTTAATGTGTCGGAGCAAGAAATGGGGCATAATGGTATCACCCTG
5050 5060 5070 5080 5090 5100 5110 5120 5130 5140 5150 5160
M P
AGAGACTCGTTACTGAAATGTCAGCCTGAGATGGATTCTCCCTGAGCCCTCCTTTAATGTTTGTCTTTTTCTTTATATGGACTAGAAATTCAGTCTTTTCTCTTCCCCTCCCAGGAACCC
5170 5180 5190 5200 5210 5220 5230 5240 5250 5260 5270 5280
Exon IV
S L K Q Q L F S Y A I L G F A L S E A M G L F C L M V A F L I L F A M *
TTCTCTGAAGCAGCAGCTCTTCTCCTACGCCATTCTGGGCTTTGCCCTCTCGGAGGCCATGGGGCTCTTTTGCCTGATGGTGGCCTTTCTCATCCTCTTCGCCATGTGAAGGAGCCGTTT
5290 5300 5310 5320 5330 Nco I 5350 5360 5370 5380 5390 5400
CCACCTCCCATAGTTCTTCTCCCGTGTCTCATCTGCCCTGTATGTTTCTTTTCCTGTACCTCCCCAGGCAACCTGGGGAAAGTGGTTGGCTCAGGGCTTGACAGAAGGAAGACAAATAAA
5410 5420 5430 5440 5450 5460 5470 5480 5490 5500 5510
TACTGTATTAATAAGATGTTTCTTGAGTCTCCTGTGTATATTTCTTTTCCACAATTGGCTGAATGCCTTGGTGAAAGTATAAGGCCAGAGGTTAGTGATGGTTTTAAACTCAATGTGGAT
5530 5540 5550 5560 5570 5580 5590 5600 5610 5620 5630 5640

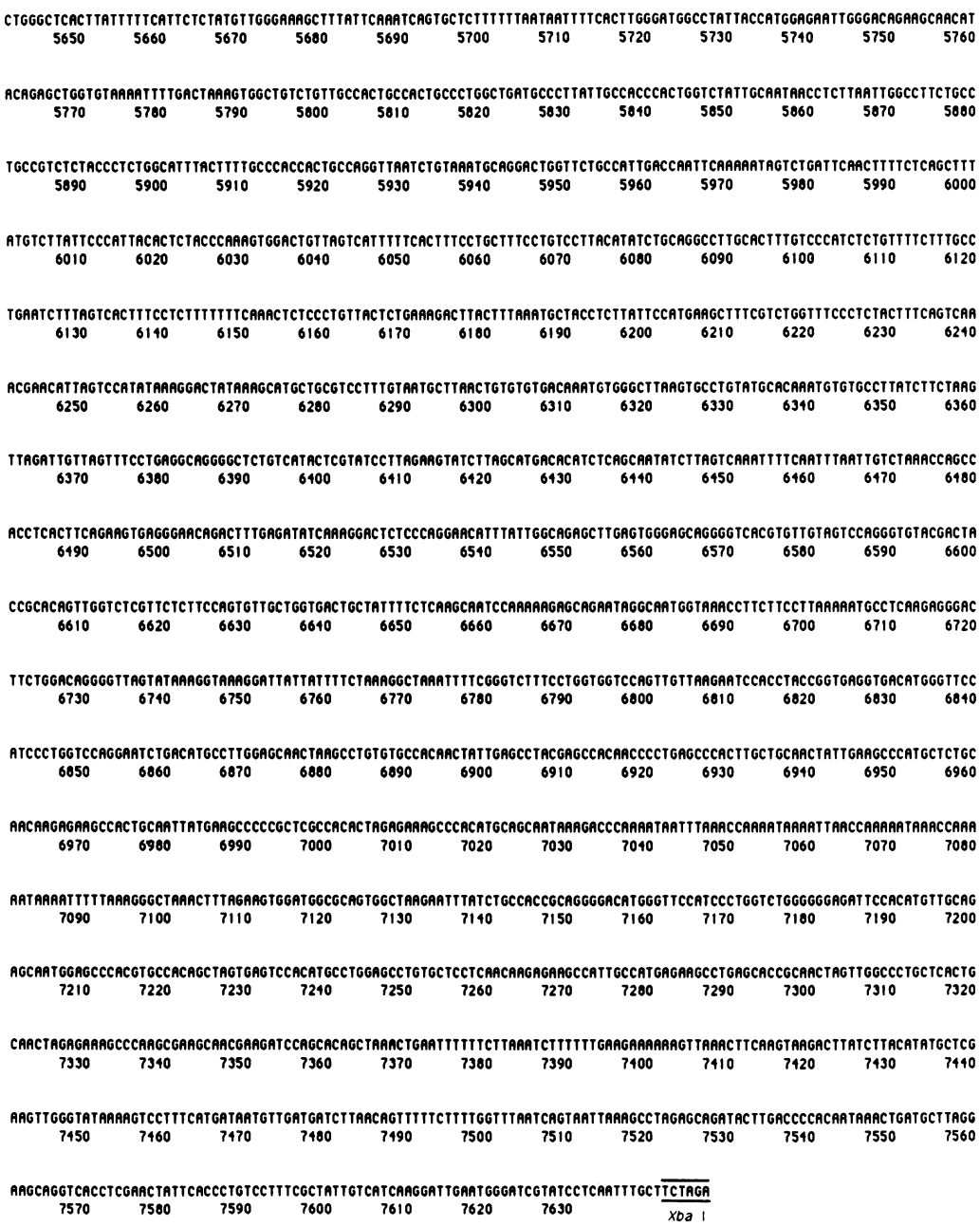


Fig. 2. DNA sequence of a segment of bovine DNA containing the bovine P2 gene for the proteolipid subunit of ATP synthase

The numbers refer to the nucleotide sequence. The *Nco*I and *Xba*I restriction sites that were used to clone the two overlapping DNA fragments are shown. *Cis*-acting DNA sequences that are of importance in control of gene expression have not as yet been identified. However, the double- and triple-underlined sequences are conserved at corresponding positions in the human P2 gene (M. R. Dyer & J. E. Walker, unpublished work) and might contain TATA boxes (Breathnach & Chambon, 1981). The overlined sequence is found 63 bp upstream from one of these conserved AT-rich elements and might be a functional CCAAT box (Efstratiadis *et al.*, 1980). The single underlined sequence denotes a potential signal for polyadenylation (Proudfoot & Brownlee, 1977). Intron–exon boundaries are denoted by arrows. A bovine repetitive element is shown also. It is related to the repetitive sequence in the bovine foetal β-globin gene (Duncan, 1987; see Fig. 4).

ADP/ATP translocase (Cozens *et al.*, 1989). These observations support the view that some exons encode structural domains (Gilbert, 1978; Blake, 1979).

The nucleotide sequences adjacent to the 5′ and 3′ boundaries of the introns in the bovine P2 gene are conserved (see Table 1). They all begin with the dinucleotide GT and end with the dinucleotide AG, and so they conform exactly with the consensus sequences adjacent to splice junctions (Breathnach & Chambon, 1981). Furthermore, the conservation extends for 8–10 bp beyond the splice junctions in the sequences of the introns, and these extended sequences agree rather well with the consensus for sequences around splice sites (Mount, 1982).

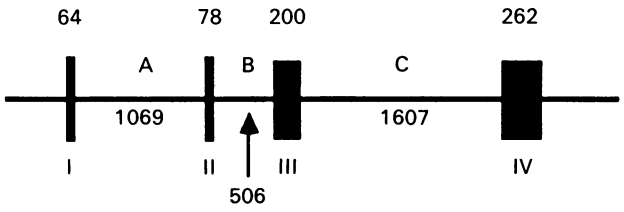


Fig. 3. Structure of the bovine P2 gene encoding the mitochondrial import precursor of the proteolipid subunit of ATP synthase

Exons I–IV and introns A–C are denoted by black boxes and solid lines, respectively. The sizes of exons and introns are given in base pairs. The site of initiation of transcription, and therefore the structure of the gene in the 5′ non-coding region, are not known at present.

An *Alu*-type repeated sequence in the bovine P2 gene

In mammalian DNA the largest class of intermediate repeated sequences is known as the *Alu* family. In humans and other primates the *Alu* sequences are well conserved (Deininger *et al.*, 1981; Daniels *et al.*, 1983), although in the prosimian species *Galago* a second distinct but related repeated sequence is also present (Daniels & Deininger, 1983). The bovine genome, in common with those of other ruminants, contains *Alu*-type elements which can be composed of monomers, dimers and trimers of a 120 bp DNA sequence, and the bovine *Alu*-type repeat shares similarity with the human *Alu* sequence over a 40 bp segment of DNA (Watanabe *et al.*, 1982). Recently, DNA sequencing studies of the bovine β-globin locus have demonstrated a second and distinct bovine *Alu*-type family (Duncan, 1987). Members of this family of repeated DNA sequences are about 500 bp long and there are about 10^5 members in the bovine genome. They are related to the original bovine *Alu*-type repeated sequence described by Watanabe *et al.* (1982) over a 75 bp segment of DNA (Duncan, 1987), but the remainder of the sequences from these two families are quite different. This second family of bovine *Alu* repeats is not similar in its DNA sequence or structure to human *Alu* repeats, nor is it composed of two related monomers, and it lacks a poly(A) tract near its 3′ end. The sequence that encompasses the bovine P2 gene contains most of the nucleotide sequence of an *Alu*-type repeat belonging to the family described by Duncan (1987; see Fig. 4). It lacks 45 bp found in the full-length repeat, presumably because it lies at the extremity of the sequence that has been determined.

DNA sequences of two bovine pseudogenes

As mentioned above, a 4.0 kb *Eco*RI fragment in λP3.17 and a 5.2 kb *Sac*I fragment in isolate λP3.9 hybridized with the bovine P1 specific gene probe. Sequencing experiments showed that each fragment appears to contain a different pseudogene, and their DNA sequences are presented in Figs. 5 and 6. These sequences have been determined in both senses of the DNA. The first fragment appears to contain the sequence of a spliced pseudogene related to the bovine P1 cDNA sequence. However, it differs in 56 nucleotide positions from the cDNA, and some of these changes result in 29 amino acid differences. The changes also give rise to two 'in-phase' stop codons, and the sequence contains a frame-shift and a 6 bp deletion. These features form part of the basis of the assignment of this sequence as a pseudogene. In addition, this pseudogene has a poly(A) tract near to its 3′ end which is preceded by a potential polyadenylation signal (Proudfoot & Brownlee, 1977), and the sequence is flanked by a direct repeat of 10 base pairs. These latter characteristics indicate that this sequence was introduced into the bovine genome by retroposition (Rogers, 1985; Weiner *et al.*, 1986).

The second sequence (Fig. 6) also is related to that of the bovine P1 cDNA, and in all probability it is a pseudogene, although other explanations of its origin are possible. In common with the spliced pseudogene described above, it contains an 'in-phase' stop-codon, and it has four small deletions relative to the bovine P1 cDNA sequence, from which it differs by 50 nucleotides; some of these differences give rise to 31 changes in amino acid sequence. However, it differs in a number of ways from the other pseudogenes for P1 and P2 that have been characterized from the human (M. R. Dyer & J. E. Walker, unpublished work) and bovine genomes. Firstly, no poly(A) tract is found near to the 3′ end of the sequence, although a potential polyadenylation signal is present; secondly, it is flanked by a shorter than usual direct repeat. This is six bases long, with the sequence CTGGGA, and the position of the direct repeat at the 3′

Table 1. Introns in the bovine P2 pre-proteolipid gene

Gene	Intron	Size	Class	Sequence 5′ boundary	Sequence 3′ boundary
Bovine P2	A	1069	0	tcc.ttg.GTGAGTACCC...	...TTCCGGCTAG.atc.agg
				S L	I R
Bovine P2	B	506	0	gat.gag.GTACCTTACA...	...TTCTTCACAG.agc.cac
				D E	S H
Bovine P2	C	1607	2	gcc.ag.GTAAGATGGG...	...CCCCTCCCAG.g.aac
				A R	N
Consensus sequence				cag.GTAAGT...	...YYYYYYYYNCAG.g

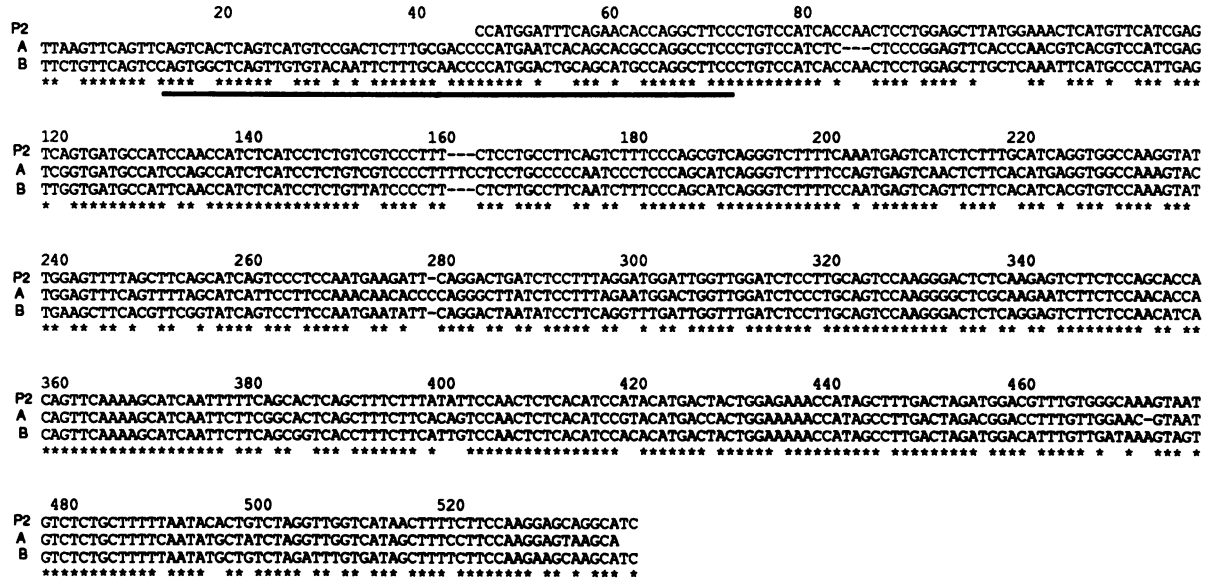


Fig. 4. An *Alu*-type repeat in the bovine P2 gene

The 5′ region of the fragment of bovine genome containing the P2 gene contains most of the DNA sequence of an *Alu*-type repeat first described by Duncan (1987). This sequence is aligned with two other members of this family of repeats from the cow β-globin locus. Sequences A and B are *Alu*-type elements from the 5′ and 3′ flanking regions of the cow foetal globin gene respectively (Duncan, 1987). The nucleotide sequences are numbered according to the *Alu*-type element A. –, Insertions needed to improve these alignments; *, conserved nucleotides. The underlined sequence is conserved in the original cow *Alu*-type family described by Watanabe *et al.* (1982).

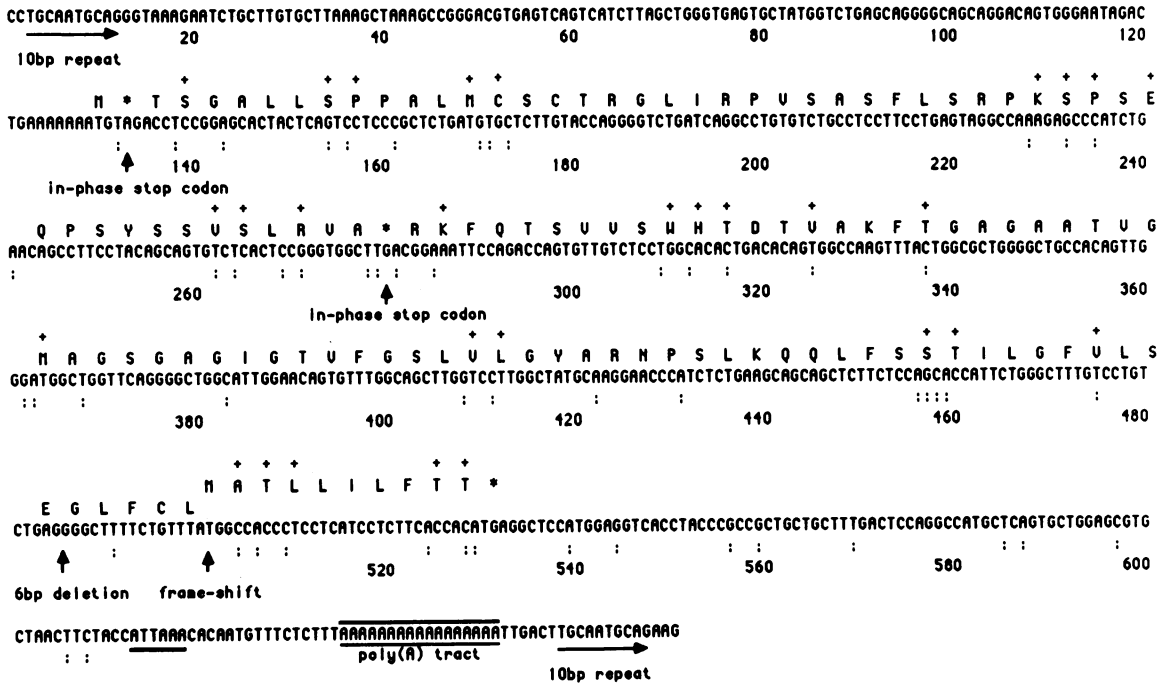


Fig. 5. DNA sequence and translation of a processed P1 pseudogene

The numbers refer to the nucleotide sequence. :, Nucleotide differences with the bovine P1 cDNA sequence (Gay & Walker, 1985); +, amino acid changes with the P1 pre-proteolipid. The potential polyadenylation signal is underlined. This gene is unlikely to encode a functional polypeptide since it has two in-phase stops, a frame-shift and a 6 bp deletion within the potential protein coding region. This pseudogene has two features which are diagnostic of retroposition (Rogers, 1985; Weiner *et al.*, 1986); it is flanked by 10 bp direct repeats and has a poly(A) tract at a position corresponding with the poly(A) tail in the mRNA.

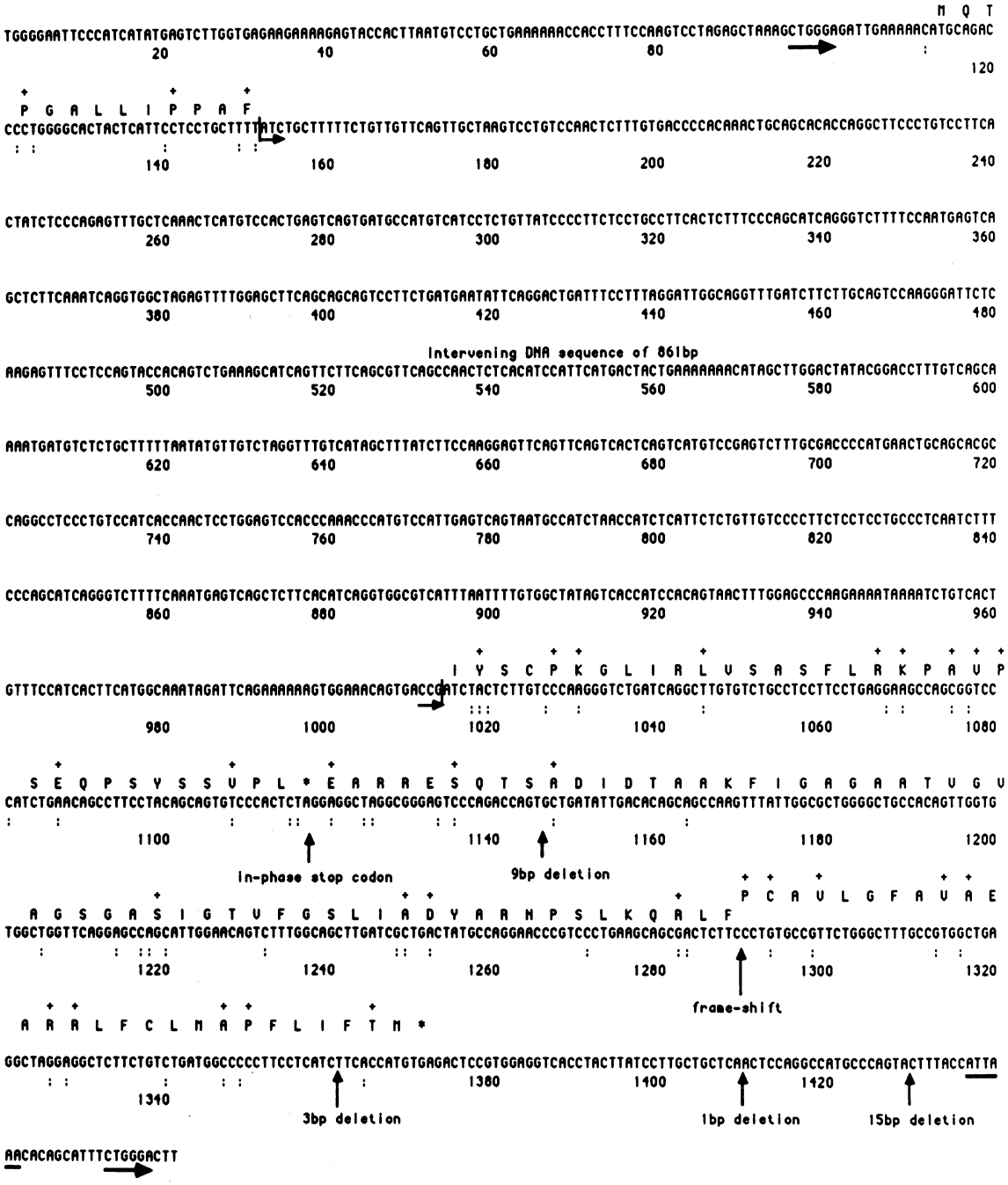


Fig. 6. DNA sequence and translation of a partially processed P1 pseudogene

The numbers refer to the nucleotide sequence. :, Nucleotide differences with the sequence of the bovine P1 cDNA (Gay & Walker, 1985); +, amino acid changes in comparison with the P1 pre-proteolipid. A potential signal for polyadenylation is underlined. This pseudogene is unusual since it contains an 861 bp intervening DNA sequence corresponding in position to an intron in the human P1 gene (M. R. Dyer & J. E. Walker, unpublished work). Boundaries between this intervening sequence and the P1 coding sequence are denoted by arrows. This pseudogene is unlikely to encode a functional polypeptide as it contains an in-phase stop codon, frame-shifts and 9 bp and 3 bp deletions within its potential protein coding region. Two other deletions are found in the DNA sequence corresponding to the 3′ untranslated region of the message. This pseudogene is flanked by short direct repeats of 6 bp.

end of the pseudogene suggests that the poly(A) sequence might have been removed during retroposition (Rogers, 1985; Weiner *et al.*, 1986). A third difference is that it contains an intervening sequence of 861 bp. This sequence is found in a position that corresponds to that of an intron in the human P1 gene (M. R. Dyer & J. E. Walker, unpublished work), but the sequence is not related to that of the equivalent intron in human P1. This bovine intervening sequence is flanked at its 5′ and 3′ ends respectively by the dinucleotides, AT and CG; these could be mutated canonical splice junction sequences, GT and AG (Breathnach & Chambon, 1981).

Another mammalian gene which could have been formed from a partially processed mRNA encodes insulin in mice and rats (Soares *et al.*, 1985). Both species have two expressed non-allelic insulin genes. The gene for preproinsulin II contains two introns and therefore is similar in its structure to the single copy genes for this protein in other mammalian species. However, the gene for preproinsulin I contains only a single intron and has distinctive features that are diagnostic of a retrogene; its sequence is flanked by 41 bp direct repeats and there is the remnant of a poly(A) tract downstream from the polyadenylation signal. There is no evidence that suggests expression of a protein from the partially spliced bovine P1 gene.

A second explanation for the origin of the partially spliced P1 sequence can be advanced. It is possible that this sequence is not a retroposon, but rather that it is a relic of a duplicated version of an early expressed P1 gene, which at the time of duplication had only one intron. However, since its sequence is closer to the bovine P1 cDNA than it is to the human P1 cDNA sequence (M. R. Dyer & J. E. Walker, unpublished work) this explanation is less plausible.

M.R.D. was supported by an M.R.C. research studentship and N.J.G. by an M.R.C. research training fellowship.

REFERENCES

Anderson, S., de Bruijn, M. H. L., Coulson, A. R., Eperon, I. C., Sanger, F. & Young, I. G. (1982) J. Mol. Biol. **156**, 683–717

Benoist, C. & Chambon, P. (1981) Nature (London) **290**, 304–310

Benton, W. D. & Davis, R. W. (1977) Science **196**, 180–182

Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) Proc. Natl. Acad. Sci. U.S.A. **80**, 3963–3965

Blake, C. C. F. (1979) Nature (London) **309**, 179–182

Breathnach, R. & Chambon, P. (1981) Annu. Rev. Biochem. **50**, 349–383

Chargaff, E. & Lipshitz, R. (1953) J. Am. Chem. Soc. **75**, 3658–3661

Cochran, M. D. & Weissmann, C. (1984) EMBO J. **3**, 2453–2459

Cozens, A., Runswick, M. J. & Walker, J. E. (1989) J. Mol. Biol. **206**, 261–280

Daniels, G. R. & Deininger, P. L. (1983) Nucleic Acids Res. **11**, 7595–7610

Daniels, G. R., Fox, G. M., Loewensteiner, D., Schmid, C. W. & Deininger, P. L. (1983) Nucleic Acids Res. **11**, 7579–7593

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

Deininger, P. L., Jolly, D. J., Rubin, C. M., Freidmann, T. & Schmid, C. W. (1981) J. Mol. Biol. **151**, 17–33

Dorn, A., Bollekens, J., Staub, A., Benoist, C. & Mathis, D. (1987) Cell **50**, 863–872

Duncan, C. H. (1987) Nucleic Acids Res. **15**, 1340

Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Conell, C., Spritz, R. A., DeReil, J. K., Forget, B. G., Weissman, S. M., Slightom, J. L., Blechl, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C. & Proudfoot, N. J. (1980) Cell **21**, 653–668

Farrell, P. J., Deininger, P. L., Bankier, A. & Barrell, B. G. (1983) Proc. Natl. Acad. Sci. U.S.A. **80**, 1565–1569

Gay, N. J. & Walker, J. E. (1985) EMBO J. **4**, 3519–3524

Gilbert, W. (1978) Nature (London) **271**, 50

Grosschedl, R. & Birnstiel, M. L. (1980) Proc. Natl. Acad. Sci. U.S.A. **77**, 7102–7106

Karn, J., Matthes, H. W. D., Gait, M. J. & Brenner, S. (1984) Gene **32**, 217–224

Kopito, R. R., Andersson, M. & Lodish, H. F. (1987) J. Biol. Chem. **262**, 8035–8040

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

McKnight, S. L. & Kingsbury, R. (1982) Science **217**, 316

McKnight, S. L. & Tijian, R. (1986) Cell **46**, 795–805

Mills, D. R. & Kramer, F. R. (1979) Proc. Natl. Acad. Sci. U.S.A. **76**, 2232–2235

Mount, S. M. (1982) Nucleic Acids Res. **10**, 459–472

Nathans, J. & Hogness, D. S. (1984) Proc. Natl. Acad. Sci. U.S.A. **81**, 4852–4855

Proudfoot, N. J. & Brownlee, G. G. (1977) Nature (London) **263**, 211–214

Rogers, J. H. (1985) Int. Rev. Cytol. **93**, 187–279

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

Sebald, W. & Hoppe, J. (1981) Curr. Top. Bioenerg. **12**, 2–64

Schatz, G. & Butow, R. A. (1983) Cell **32**, 316–318

Soares, M. B., Schon, E., Henderson, A., Karathanasis, S. K., Cate, R., Zeitlin, S., Chirgwin, J. & Efstratiadis, A. (1985) Mol. Cell Biol. **5**, 2090–2130

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

Staden, R. (1982) Nucleic Acids Res. **10**, 4731–4751

Staden, R. (1985) in Genetic Engineering: Principles and Methods (Setlow, J. K. & Hollaender, A., eds.), vol. 7, pp. 67–114, Plenum, New York and London

Walker, J. E., Gay, N. J., Powell, S. J., Kostina, M. & Dyer, M. R. (1987) Biochemistry **26**, 8613–8619

Watanabe, Y., Tsukada, T., Notake, M., Nakanishi, S. & Numa, S. (1982) Nucleic Acids Res. **10**, 1459–1469

Weiner, A. M., Deininger, P. L. & Efstratiadis, A. (1986) Annu. Rev. Biochem. **55**, 631–661

Received 11 November 1988; accepted 15 December 1988