

# Sequences of members of the human gene family for the c subunit of mitochondrial ATP synthase

Mark R. DYER\* and John E. WALKER†

Medical Research Council Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, U.K.

Subunit c is an intrinsic membrane component of ATP synthase, and in mammals it is encoded by two expressed nuclear genes, P1 and P2. Both genes encode the same mature c subunit, but the mitochondrial import pre-sequences in the precursors of subunit c are different. The DNA sequences of the human P1 and P2 genes are described. They occupy about 3.0 and 10.9 kb respectively of the human genome, and both genes are split into five exons. The human genome also contains about 14 related

spliced pseudogenes, and the sequence of one such pseudogene related to P2 is described. Sequences flanking the 5' ends of the human P1 and P2 coding sequences each contain a CpG-rich island. Potential promoter elements (TATA and CCAAT boxes) are present in the 5' sequences of the P1 gene, but not that of P2, although there is no direct experimental evidence to show the involvement of these sequences in transcription of the genes.

## INTRODUCTION

Bovine mitochondrial ATP synthase is a membrane-bound complex of 14 different polypeptides (Walker et al., 1991), and the c subunit (also known as the dicyclohexylcarbodi-imide-reactive proteolipid) is an essential part of the proton channel in the membrane sector (Sebald and Hoppe, 1981). It is a hydrophobic protein of 75 amino acids, probably folded into a hairpin of two transmembrane  $\alpha$ -helices linked by a  $\beta$ -turn near the membrane surface. A carboxyl group essential for functioning of the proton channel, and the site of reaction of dicyclohexylcarbodi-imide, is situated near to the middle of the C-terminal  $\alpha$ -helix. In mammals, *Neurospora crassa* (Jackl and Sebald, 1975) and *Aspergillus nidulans* (Turner et al., 1979), but not in *Saccharomyces cerevisiae* (Macino and Tzagoloff, 1979), subunit c is a nuclear gene product, synthesized on cytoplasmic ribosomes as a precursor with an N-terminal extension. The extension directs the protein into the mitochondrion and is cleaved during import. However, the proteolipid is highly unusual, if not unique, amongst nuclear-encoded mitochondrial proteins in having two different precursors derived from separate genes (Gay and Walker, 1985). Both cDNAs for the precursors contain a segment coding for the same mature proteolipid, but the N-terminal presequences, although related, differ extensively. The 3' non-coding regions of their cDNAs are only weakly related and so each can be employed as a specific hybridization probe (Gay and Walker, 1985). As described here, we have isolated and sequenced the human P1 and P2 genes. They are members of a complex gene family that includes numerous spliced pseudogenes for P2, and probably for P1 also. The expressed P1 gene is distributed over about 3.0 kb of DNA and the human P2 gene occupies about 10.9 kb of the genome. Both contain four introns at equivalent positions. Interest in this gene family has been increased by the recent finding that in the fatal human disease ceroid lipofuscinosis, or Batten's disease, subunit c accumulates in lysosomes (Palmer et al., 1992).

## MATERIALS AND METHODS

### DNA hybridization

Digests of human DNA prepared from a placenta (Walker et al., 1987) were fractionated by electrophoresis in 0.6% agarose gels, and fragments were transferred to nitrocellulose filters (Southern, 1975). The filters were incubated at 65 °C, first for 1 h in a solution containing 6 × SSC (1 × SSC is 0.15 M NaCl and 0.015 M trisodium citrate), 0.2% BSA (fraction V), 0.2% polyvinylpyrrolidone, 0.5% N-laurylsarcosine and sonicated salmon testis DNA (100 mg/ml), and then for 15–20 h in the presence of radioactive 'prime-cut' probes (Farrell et al., 1983) dissolved with 10% dextran sulphate in the same solution. The filters were washed four times for 30 min each at 65 °C in either 0.2 or 2 × SSC, each containing 0.5% N-laurylsarcosine. Autoradiographs of filters were exposed with an intensifying screen at –70 °C for either 1–7 days (genomic DNA) or 1–3 h (phage DNA).

### Screening of genomic libraries

The human genomic libraries SH, AT5 (LeFranc et al., 1986) and RPMI (Forster et al., 1987), consisting of partial *Sau3A* fragments cloned into the *Bam*HI site of  $\lambda$ 2001 (Karn et al., 1984), were gifts from Dr. T. H. Rabbitts. Plaques (approx.  $5 \times 10^6$ ) were produced on *Escherichia coli* Q358 grown on 20 cm diameter agar plates. Phage were transferred sequentially to two nitrocellulose filters per plate, and each library was screened (Benton and Davis, 1977) with the two prime-cut probes. DNA was prepared from recombinant phages (Maniatis et al., 1982) grown in 500 ml cultures of *E. coli* Q358.

### Sub-cloning and DNA sequencing

A 4.4 kb *Bam*HI fragment containing part of the human P1 gene was excised from  $\lambda$ HP1.9, sonicated and sub-cloned into the

\* Present address: Sandoz Ltd., CH-4002 Basle, Switzerland.

† To whom correspondence should be addressed.

The nucleotide sequence data reported will appear in the EMBL, GenBank and DDBJ Nucleotide Sequence Databases under accession nos. X69907, X69908 and X69909.

3' region of Alu repeat 1 (+)

GATCAAGACCATCTGGCCAACACGGTGAA	20	30	40	50	60	70	80	90	100	110	120	
Sau 3A												
CAGGAGAATGGCGTGA	130	140	150	160	170	180	190	200	210	220	230	240
CCCCCAGGGAGCTTGAGCTGAGCCGAGATCGC												
CACTGCACACTGAGCTGGGCAATGCTGAA	250	260	270	280	290	300	310	320	330	340	350	360
ACAGAGCTGGGAGGCTGAGGTAGGCCA	370	380	390	400	410	420	430	440	450	460	470	480
ACTGGGAGGCTGAGGTAGGCCA	490	500	510	520	530	540	550	560	570	580	590	600
AGAGAGGCCGGTGCAGTGGCTCA	610	620	630	640	650	660	670	680	690	700	710	720
Alu repeat 2 (+)												
CTATCTATCTATCTATCTATCTATCTA	730	740	750	760	770	780	790	800	810	820	830	840
TTAGCCGCAACTGCACTCCAGCTGGG	850	860	870	880	890	900	910	920	930	940	950	960
CCTAGCACAGTATTTGTCATAATGAGGT	970	980	990	1000	1010	1020	1030	1040	1050	1060	1070	1080
ATCGATAATTGATAGCTTAAACACTAAC	1090	1100	1110	1120	1130	1140	1150	1160	1170	1180	1190	1200
ACAGCTGGCTCCACTTCCCTGGG	1210	1220	1230	1240	1250	1260	1270	1280	1290	1300	1310	1320
GGCTATGTTGCCAGGGCTGGCTCC	1330	1340	1350	1360	1370	1380	1390	1400	1410	1420	1430	1440
AAATGTGCTGGCCAGCACAGTGGCT	1450	1460	1470	1480	1490	1500	1510	1520	1530	1540	1550	1560
Alu repeat 4 (+)												
TCTCTACTAAAAAATTTAAATTAGCTGGGGTGGTGGCCG	1570	1580	1590	1600	1610	1620	1630	1640	1650	1660	1670	1680
CTGAGCTGAGATCTGCCACTGCACTCCAG	1690	1700	1710	1720	1730	1740	1750	1760	1770	1780	1790	1800
CCAGGACACTTGGGAGGCCAACGGGGG	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920
Alu repeat 5 (+)												
TCCGCACCTGTAATCCCCAGCTACTGAGGAGCTGAGG	1930	1940	1950	1960	1970	1980	1990	2000	2010	2020	2030	2040
GACTCTATCTCAAAAAGAAAAAGAAAAAGAAAAAGAAA	2050	2060	2070	2080	2090	2100	2110	2120	2130	2140	2150	2160
CCAGAGGCTGAAGTGGGAGGATCTTGAGGCCAGGG	2170	2180	2190	2200	2210	2220	2230	2240	2250	2260	2270	2280
CCCAGGAGATAAGGTATAATAGTGAGGGATGATTG	2290	2300	2310	2320	2330	2340	2350	2360	2370	2380	2390	2400
AAGACTCCAGACTTGTAATTGCCAGATTAGTCA	2410	2420	2430	2440	2450	2460	2470	2480	2490	2500	2510	2520
ATTTCAACATTTCATCTTTCATGCATA	2530	2540	2550	2560	2570	2580	2590	2600	2610	2620	2630	2640
TGCCCTGCAATCCCAAGCACTTGAAGGCCAAGGG	2650	2660	2670	2680	2690	2700	2710	2720	2730	2740	2750	2760
Alu repeat 6 (+)												
GGGCATGGTGGTGCAGGCCCTGTAATCCCAAGCTACT	2770	2780	2790	2800	2810	2820	2830	2840	2850	2860	2870	2880
TTGGGAGGCTGAGGAGAGAAATTGCTGAACCGGGGAGG	2890	2900	2910	2920	2930	2940	2950	2960	2970	2980	2990	3000
Alu repeat 7 (+)												
CGGGATCACGAGGTCAAGGAGATCGAGGACCATCCG	3010	3020	3030	3040	3050	3060	3070	3080	3090	3100	3110	3120

Figure 1 For legend see page 55.

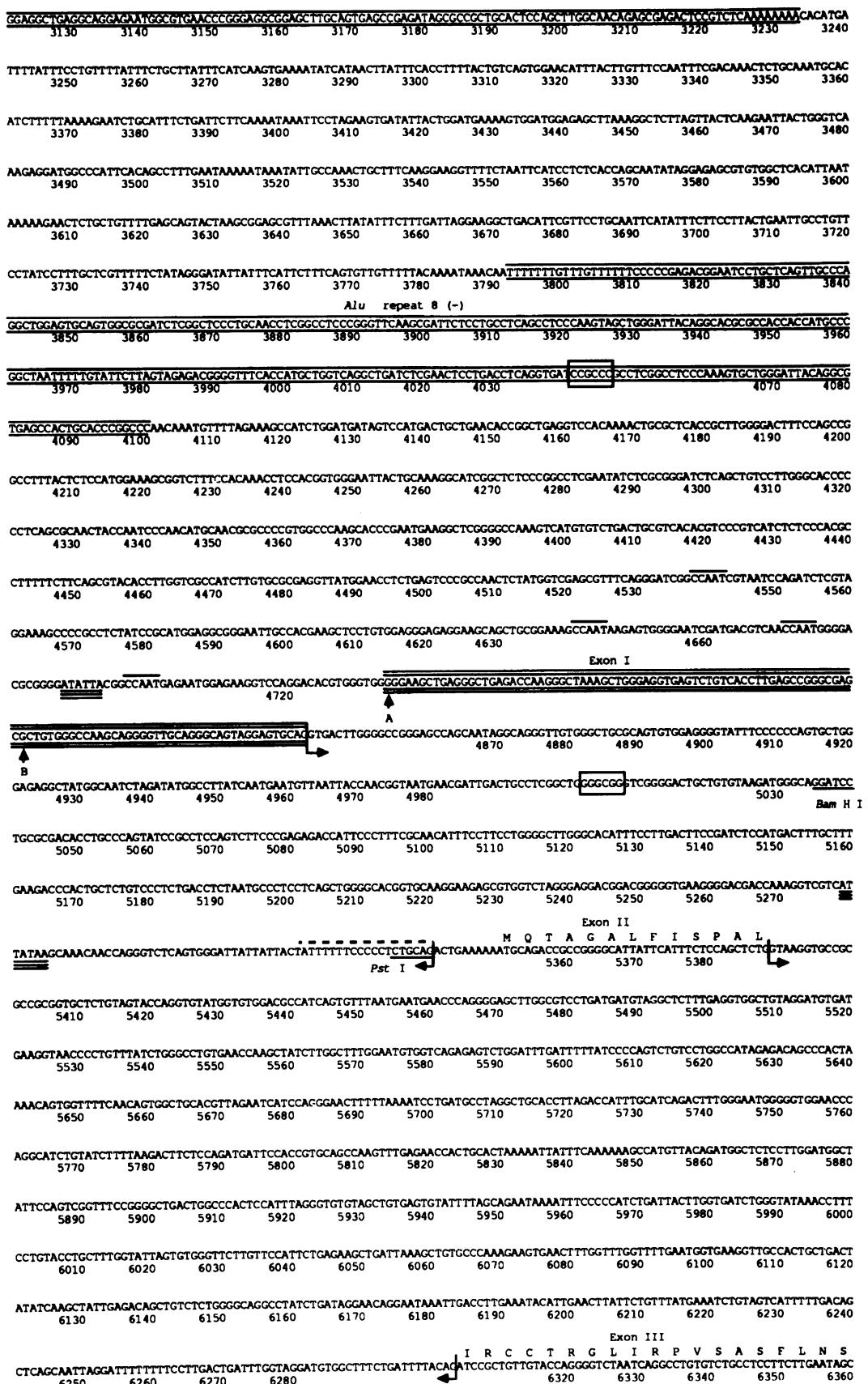


Figure 1 For legend see page 55.

P V N S S K O  
 CCAGTGAATTCACTAAACAGCTAAGGGAGGAATAGCCTCTCTAGGAATGTCGCCAAGGGCCCAGGATGGTGGCTCACACCTGTAAATCCAGCTTTAGAGGGCTAGGGAGAGGATCA  
 6370 6400 6410 6420 6430 6440 6450 6460 6470 6480

Alu repeat 9 (+)

CTTGAGGCCAGTTGTTCAAGACCAGCTGGGAAACACAGTGAACCCATCTCAACAAAAAAAAATTAGCCATGTGTTGGCATGTGGCTGTAGTCCAGTTACTCCAGGG  
 6490 6500 6510 6520 6530 6540 6550 6560 6570 6580 6590 6600

TGAGATGAGAGGGATTGGGTAATCCCAGAGGTGAGACTGCTGTGACTATGCCACTCCAGACTGGGTGACAGGCAAGACCCCTGACTCATAAACCCATTAAATTATTTTTA  
 6610 6620 6630 6640 6650 6660 6670 6680 6690 6700 6710 6720

AATCAAGGAATGTCAGGAGCTGGGCTTCAGCTTAAAGTGCCTTAAAGCCAAGTGGGAGGAGGTGGAGAACTGTATGCTTAAATAGTCTAAGTCCCCAAGGTAAACCTGT  
 6730 6740 6750 6760 6770 6780 6790 6800 6810 6820 6830 6840

AGCTAAAGAAACTATAAAAGTGAGGTGCTCTAACCCATGAATTTCCOCTAGTCCTGTCGCTGTCTGACAAACAGACTCTAAGACTAACTCCAGCATACTGCAATCAGAGTATTCTG  
 6850 6860 6870 6880 6890 6900 6910 6920 6930 6940 6950 6960

CATTACCCAGCGATTGCCAACAGTTCAAGGCTCAGGTGGGTGGGTGAAGGGAGTAGGTGACCCACCTGCTCTTATGCCACTATCTCTGCTATCTGCCCTGGCTTC  
 6970 6980 6990 7000 7010 7020 7030 7040 7050 7060 7070 7080

Exon IV

P S Y S N F P L Q V A R R E F Q T S T V V S R D I D T A A K F I G A G A A T V  
TTCTTACCTCTTACAGCAACTCCCACTCCAGGTGCCAGACGGGAGTCCAGACCCAGTGTGCTCCGGGACATTGACACAGCAGCAGTTATTGGTGTGGCAGCCACAGT  
 7100 7110 7120 7130 7140 7150 7160 7170 7180 7190 7200

G V A G S G A G I G T V F G S L I I I G Y A R  
TGGTGTGGCTGGTCAAGGGCTGGCATGGGAACCGTGTGGCAGCTTGTATGGCTATGCCAGTAAAGTTGGTGGCTACAGCATCTCCACTGTAAATTCCACCCGGTTGGG  
 7210 7220 7230 7240 7250 7260 7280 7290 7300 7310 7320

AAGGCTCAGCTGGAGGAGCTCCCTTCAGGAGCCTTCAGGTGATTCTACATCATAGTTCTCCAGAGAAACATGCACTCAGGCTGGCTACTAAACCTACAGGTTGGCCAACT  
 7330 7340 7350 7360 7370 7380 7390 7400 7410 7420 7430 7440

CTGCTTGTCATCCAAATCCCCAGGATCTGTGCGAGCTAGGCCCTCTCCAGGAGTAACAGTCCCACTTACCTCACCCTCTGTGCTCCCTCCCTCACCCCTCTCTCCCAA  
 7450 7460 7470 7480 7490 7500 7510 7520 7530 7540 7550 7560

Exon V

N P 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  
CTGGCAATGATCTGTCCCTCCAGAACCCGCTCTCAAGCAGCAGCTCTCTCCATGCGATCTTGGCTTGTGAGGCCATGGGGCTTTCTGTTGATGGTGGCTTC  
 7570 7580 7600 7610 7620 7630 7640 7650 7660 7670 7680

L I L F A M \*  
CTCATCCCTCTGCCATGTGAGGCTCCATGGGGCTACCCGCCCTGGCTACTGCAACTCACCCATCTTGGCTGTGGGTTGTGTTAACGTTTACATTAACACAACGTTCTCTA  
 7690 7700 7710 7720 7730 7740 7750 7760 7770

AACCCCTGTCTGTGCCCTGTGCTTGCACCTTCAGGGGCCCTGGTAGAGGGTGAGGAGGGAAATTGCTCAGGCATGGGTGATGGAACTCTCAGGGCTGAGAAGGAGCACTGCC  
 7810 7820 7830 7840 7850 7860 7870 7880 7890 7900 7910 7920

ATGACTGTAAGAACAGAGGGTTCTCATGTTCTCCAGGAGGAATATGGTTGTGAGTAATGAGATTGGCTTCTCCAGATTATACAGTTGGCCATCTGATCCATAGGT  
 7930 7940 7950 7960 7970 7980 7990 8000 8010 8020 8030 8040

TCCATATTTGTGATTCAACTAACAGTAGACAGAAAATATTAGGGAAATTCAGGGAAAAAGTGAATGGTTGTGCTGTACTGCACAAAGTACAGGTTTTTTCTGTGTCATTATTCCCTAAACATA  
 8050 8060 8070 8080 8090 8100 8110 8120 8130 8140 8150 8160

CAGTATGCCATTATATAGCATTTACCTAGTATTAGGTATTGTAAGTAATCTAGAGATGATTAAAGTATATGGAGGATGTGCATAGGTTATATGCAAATACTACACCAATT  
 8170 8180 8190 8200 8210 8220 8230 8240 8250 8260 8270 8280

TATAAGGGACTTGAGCATCTGTGTTATCCACTGGGGCCTGGAGGCCAATCTCCACAGATACCAAGGATACTGAAACACTGCATATTAGGTTACATTGGGCTTAAATT  
 8290 8300 8310 8320 8330 8340 8350 8360 8370 8380 8390 8400

ACACCAAGCATTTCACATTGTTATCTCAACACCCCTAAGAGTTAAAGCACAACCTTAAAGCTTGTGCAATGCAAGACCCAGGGTTGGGAACAAAATACCTTACAGTATCA  
 8410 8420 8430 8440 8450 8460 8470 8480 8490 8500 8510 8520

CAGACTACTAAGTGGCAGATTGGGATTCTCCCCAGGCCAGCTTAAACCTTAACTGCCACTTGTGCTATCTATTCTTTTTTTGTCATCTATTCTGCT  
 8530 8540 8550 8560 8570 8580 8590 8600 8610 8620 8630 8640

CACACAACATTAAATGTGGCACAGAAGGGAAACTTAGTCCAAAGAAGTTGTATAGCAGGTGAAAGGTAGAGGTAGGCTCTATCTGCCGTGTTAGCCCCATTGTAGGTTCTT  
 8650 8660 8670 8680 8690 8700 8710 8720 8730 8740 8750 8760

TCATCATCAATTAGCCCTTCTCAGATCTAAATGTTCTTAAACGAGATAAAGTACACATCTGGTCTTGGAGAATTACCTTACGGTGTGAGGATAAAAGACTATCTGGAC  
 8770 8780 8790 8800 8810 8820 8830 8840 8850 8860 8870 8880

Alu repeat 10 (+)

CAGGCATGGTGGCTCACGCCATGTAACTCCAGCACTTGGGAGGCTGAGGCCGCCGATCACCTGAGGTGAGGACTTCGAGACCACCCATGAAACATGGAAACCCAGTCTACTAA  
 8890 8900 8910 8920 8930 8940 8950 8960 8970 8980 8990 9000

AATACAAAGTAAACCAGCCATGGTGGCCATGCTGTAACTCCAGCTACTTGGGAGGCTGAGGCCGCCGATCACCTGAGGTGAGGACTTCGAGACCACCCATGAAACATGGAAACCCAGTCTACTAA  
 9010 9020 9030 9040 9050 9060 9070 9080 9090 9100 9110 9120

TGCACTCCAGCCCTGGGAAACAGAGGAAACACTCCGCTCTAAAGGAGGAAATGATGATTGGGGCTGTGCAAAATAATGAAATTCCATAAAGAGCTGTCAGCATCTTCTG  
 9130 9140 9150 9160 9170 9180 9190 9200 9210 9220 9230 9240

GGTTCTCAAGTCTATGCCACCAGCTTGGCCCTCTGGCTAAAGTGGGACTGATAACCCATGTAAAATAGCTAGTTGACATATCACAGTTGCTGCTCTCTAAGTCCCTGAGGA  
 9250 9260 9270 9280 9290 9300 9310 9320 9330 9340 9350 9360

GCAACCATGGCAGAGAACTCATGCAAATGCTGGGACCTGGTCTGCTCAAATTGGATGGCTCAAAAGGAGTAGTAAAGCTAAAGTGGATCC  
 9370 9380 9390 9400 9410 9420 9430 9440 9450 Bam H I

Figure 1 For legend see opposite.

*Sma*I site of M13mp8 (Deininger, 1983). A 5.3 kb *Xho*I–*Pst*I fragment extended the sequence in a 5' direction. Similarly, a 12.2 kb *Sac*I fragment in  $\lambda$ AT5P2.1 appeared to contain at least a substantial part of the human P2 gene. It was amplified in pUC12 (Messing, 1983), sonicated and sub-cloned into the *Sma*I site of M13mp8 (Deininger, 1983). Subsequently the sequence was extended beyond the 5' end of this fragment by sequencing an overlapping 3.8 kb *Pst*I fragment.

DNA sequences were determined at least once in both senses of the DNA, and on average five and six times in P1 and P2 respectively, by the modified dideoxy chain termination method (Sanger et al., 1977; Biggin et al., 1983). Problematic sequences were resolved by substituting either deoxyinosine triphosphate (Mills and Kramer, 1979) or deoxy-7-deazaguanosine triphosphate (Mizusawa et al., 1986) for dGTP in the sequencing reactions. Data were compiled with programs DBAUTO and DBUTIL (Staden, 1982) and analysed with ANALYSEQ (Staden, 1985). Sequences were aligned with programs NUCALN and PRTALN (Wilbur and Lipman, 1983).

## RESULTS AND DISCUSSION

### Characterization of the genes

Attempts to clone the human P1 and P2 genes were hampered by the presence in the genome of numerous related spliced pseudogenes. In consequence, almost all recombinants identified by screening of genomic libraries contained spliced pseudogenes. A similar obstacle had been encountered previously in cloning the bovine genes, and no recombinant containing the expressed P1 gene was identified (Dyer et al., 1989). In the case of the human P1 gene, but not the P2 gene, this problem was surmounted by rescreening restriction digests of positive clones with a second probe derived from the 5' region of the P1 bovine cDNAs, and searching for recombinants in which the 5' and 3' probes hybridized either with the same large fragments in various digests, or with more than one fragment in the same digest. Thus, isolate  $\lambda$ HPI.9 from the SH library was found to contain only large restriction fragments (> 3 kb) that hybridized with both the 5' and 3' probes, indicating that the hybridizing sequences were distributed in several kilobases of DNA. Amongst hybridizing fragments in  $\lambda$ HPI.9 was a *Bam*H1 fragment of 4.4 kb, and a fragment of this size also was detected in digests of human DNA (results not shown). It was sequenced and proved to contain the expressed human P1 gene.

A clone containing the human P2 expressed gene was isolated by probing with a sequence at the 5' end of sequence determined in the bovine P2 gene (Dyer et al., 1989) which is not present in the bovine P2 cDNA clone. This sequence is now known to be part of intron A of the bovine and human P2 genes. Large restriction fragments (> 12 kb) in recombinant  $\lambda$ AT5P2.1 hybridized both with this probe and with the P2 probe derived from the 3' end of the bovine cDNA. A 12.2 kb *Sac*I fragment

from  $\lambda$ AT5P2.1 was sequenced and contained the expressed gene. A fragment of similar size hybridized with the P2 probes in a digest of human DNA (see Figure 6).

### The human P1 and P2 genes

The human genomic sequences containing the expressed P1 and P2 genes are 9457 and 15 016 bases in length respectively (Figures 1 and 2). Their G+C contents are 47.8% (9.4 kb sequence) and 46.9% (15 kb sequence). There is one ambiguity in the P1 gene at nucleotide 9444, where an A residue was found in one clone, and a G residue in two others. It is assumed that the correct assignment is G. Nucleotide sequences of partial cDNAs for human P1 and P2 have been described (Farrell and Nagley, 1987), but both differ from the corresponding genomic sequences at several positions (see legend to Figures 1 and 2). The P1 cDNA clone confirms that poly(A) is added after nucleotide 7799.

The protein sequences of the human and bovine P1 and P2 precursors contain an identical mature c subunit. Assuming that the sites of cleavage of the pre-sequences are the same as in the bovine proteins, the human pre-sequences of P1 and P2 are 61 and 66 amino acids long respectively. In contrast to the mature proteins, the pre-sequences are not conserved. Those of the P1 proteins are the same length, but the sequences differ in 11 amino acids. The bovine P2 pre-sequence is two amino acids longer than the human homologue, and the sequences differ in 17 amino acids.

The human P1 and P2 genes are both divided into five exons (Figure 3). In common with the rather narrow range of exon lengths observed in other eukaryotic genes (Naora and Deacon, 1982), their sizes range between 29 and 259 bp (Table 1). Introns B–D in both genes are found at almost identical positions to those in the bovine P2 gene (Dyer et al., 1989). In the human P1 and P2 genes (and also in bovine P2), exons I are in the 5' non-coding region, and exons II correspond to the rest of the 5' non-coding regions present in the mRNAs and to a region encoding part of the import pre-sequence. The rest of the pre-sequences are encoded in exons III and part of exons IV, which code for the N-terminal 38 amino acids of the mature protein. In none of the three genes is an intron found at the boundary between the processed import sequence and the mature protein. In contrast, the import sequence and the 5' non-coding region of the mRNA of another mitochondrial protein, subunit IV of cytochrome c oxidase, are encoded in separate exons (Bachman et al., 1987), and an intron separates almost all of the DNA coding for the import sequence of the human  $\beta$ -subunit of ATP synthase from the region coding for the N-terminal end of the mature protein (Ohta et al., 1988).

Intron D in the human pre-proteolipid genes almost certainly does correspond to a boundary between structural domains in subunit c. It interrupts the sequence coding for Arg-Asn-Pro, which is believed to form a  $\beta$ -turn outside the lipid bilayer and

**Figure 1** DNA sequence of a fragment of human DNA containing the P1 gene

Exon I (marked with double lines above and beneath) is homologous to sequences in the 5' regions of a bovine processed pseudogene (Dyer et al., 1989) and of an ovine P1 cDNA (Medd et al., 1993) from sites A and B respectively. Protein sequences are shown over exons II–V, and the small arrows denote exon–intron boundaries. The part of intron A (marked with a broken line; nucleotides 5322–5340) is very similar to nucleotides 1–31 of bovine P1 cDNA (Gay and Walker, 1985). The sequence differences in the 5' regions of a bovine pseudogene and of bovine and ovine cDNA can be explained by two alternate transcription initiation sites in the human sequence, corresponding to two TATA boxes (triple underlines; nucleotides 4688–4693 and 5279–5285). CCAAT boxes have a single line above them, and the sequence GGGCGG and its complement are boxed. The doubly underlined sequence (nucleotides 7780–7785) is a polyadenylation signal, and poly(A) is added between nucleotides 7800 and 7802 (Farrell and Nagley, 1987). The *Alu* repeats on the displayed and complementary DNA strands are denoted by (+) and (–) respectively. The *Sau*3A site at the 5' end of the insert in  $\lambda$ HPI.9 is shown, as are the *Pst*I and *Bam*H1 sites used in cloning this sequence. A partial human P1 cDNA sequence (Farrell and Nagley, 1987) covers the coding nucleotides from 6372 to 7785. At the positions corresponding to nucleotides 7126, 7128, 7129, 7702 and 7785, Farrell and Nagley (1987) report C, G, C, A and T in the cDNA. In addition, the cDNA sequence lacks 33 nucleotides that are present in the gene sequence, from bases 7711 to 7743.



TGCTTATCCCACCTCGGTAGTTGAAACCTGACCAGGGCTGGCAACCCCTGGACTGACCTCGAACCTTACACGGGGTCCACGGTCCCCTTATTCTAGACTGTCGCCAGGCCAGC  
 3250 3260 3270 3280 3290 3300 3310 3320 3330 3340 3350 3360  
  
 AAAGGGGATTCCGATCATATTAACCTATCTCTGGCTCGTGTGCTGCTCTCATGGTTTATCACCGGAATGGAAAGGAGGAAACGGGAGTGGGGAAAGGGCAAGCCACCA  
 3370 3380 3390 3400 3410 3420 3430 3440 3450 3460 3470 3480  
  
 GAAGCAGCTGGGCATTTTGCTGACTATTCGCCCATGGAGCGTTTGCCATCCAGTGTGCTGCCAGAGTCCCGGGGACAGAAAGCAGGACTCCAAAGATCTCCCTCGTCATTTA  
 3490 3500 3510 3520 3530 3540 3550 3560 3570 3580 3590 3600  
  
 GGCACGTAAACCCATCCCGACCGTATGTAACACTGAACTCTGAGTAGGAGTTTGGCAGTGTGACACTGCCATGCTCGGAAAGGGACCCAGTCCGAGATCCCAACCGG  
 3610 3620 3630 3640 3650 3660 3670 3680 3690 3700 3710 3720  
  
 GTACAGACCAAACCCAGTCCACGGTACGGCTTACTCCCGGAGTGGCCTCATTCCTGAGTCAGTGGCTCCCTGTAGTTCTCTCTGCAACGCCAGGGAGCAACGGCG  
 3730 3740 3750 3760 3770 *Pst I* 3800 3810 3820 3830 3840  
  
 ATACGCCACAGCCCTGGCAGGGGGCTGTGTGCTGAGCTGATCTGATCTGAACTCACTCTATCCGTGGCTGAGCAGCTGGCCGGTACGGCATGCCCTGAGCCCTCT  
 3850 3860 3870 3880 3890 3900 3910 3920 3930 3940 3950 3960  
 Exon I  
 TTGCTCTTCCGCTGCTCGCCCGGCTGTGTGCTGAGCTGGGGAGCTGAAAGCTCTGAGCTGGCTGGCTGGCTGGCTGGGGGACACCTCACCTTGCGGG  
 4080 4090 4100 4110 4120 4130 4140 4150 4160 4170 4180 4190 4200  
 A  
 TGCTTGGAGGCTTACCTGGCTGCTCCACCCCTGGCTCTCGCCCTTCTGCTCCGCTATCCCTAGGCTTAGGTTGGCTTCCCTGACCTGGCTGGGGGACCCCTGACCT  
 4210 4220 4230 4240 4250 4260 4270 4280 4290 4300 4310 4320  
  
 CTGGCCCATCTGGATGCTCGCAGGGGGAGGATACCTGGAGCTGGCAAAGTGTAGCTCTGGAGTGTGCTGGCTGGCTGGCTGGGGGACACCTTCTGGCTGGGG  
 4330 4340 4350 4360 4370 4380 4390 4400 4410 4420 4430 4440  
  
 GGTGCTCCGAGCAATGGAGATAGTGGATCGCTAAAGTCACCAATCCCTAGTCAGCACCCCTCAATTAAAAGTAACCTGAACTGGTAGAAGTGGGAAAAAGCCAAACATC  
 4450 4460 4470 4480 4490 4500 4510 4520 4530 4540 4550 4560  
  
 GCCTCCGGTTAGGTTGGCTGGGATGCTAAGGAGAAGAGAGAGGCTGGTGGCTGGGATCACTAGTTTACCACTGGCTTCTCGCATCCGGCACGGCTGGTAGGTAA  
 4570 4580 4590 4600 4610 4620 4630 4640 4650 4660 4670 4680  
  
 AGTCACAAATGGTGGCTAGAGATGAGGGTTCTGTGAGGCCAGGGCCAGAAGGTGGGTAAGGCAAGCTAGTGGGCCAGCATGAGACCCCTGGCTGCCAGCTTCTGTGAA  
 4690 4700 4710 4720 4730 4740 4750 4760 4770 4780 4790 4800  
  
 ATTAATTAACCTCCCCAGATTCTGCTGTACAAGGAAAAGAAGGTTAAGGAAAGGAATAGTTAGTTACTACCCCTGTGACTCATCTACACTGAGGAAAAGCCGTGATTTGGA  
 4810 4820 4830 4840 4850 4860 4870 4880 4890 4900 4910 4920  
  
 GTCTGATTCTGGATAGTGGCTTGGCCAGTTGAGCTAACATTGCAAGTGGCTGGGATAGCCATGGTAGATAACCAATTGGGGGACTCAATTGCCACCCCTGTTGGGTTTATAT  
 4930 4940 4950 4960 4970 4980 4990 5000 5010 5020 5030 5040  
  
 GCCAGCACCACTAAACTCTTATGTTGTTAAATGAAACATACTTAATCCCTAAGAGCTGCAATTACAAAAGATAATTAGGATAATTCCCTAGAAAAGAACCC  
 5050 5060 5070 5080 5090 5100 5110 5120 5130 5140 5150 5160  
  
 CCTCTTTCTAAAGCCTTAAAGAAGTATTGGATTGGCCAGAAAACCTGAATAGTACGGCAGAAAATTGGGTTAGAGTACATAGGCAAAATCAGGTTCTAGTAGGC  
 5170 5180 5190 5200 5210 5220 5230 5240 5250 5260 5270 5280  
 Alu repeat 2 (+)  
 AAGAGTTGAAGACATCCATTAACTTAAAGTAATTAAAGCTGGGTGGTGGCTAACACCTGTAATCCCAACACTGGGGGGCAAGGTGGGAAAGTCCTTGTGATGCCAGGAGTCCACAA  
 5290 5300 5310 5320 5330 5340 5350 5360 5370 5380 5390 5400  
  
 TAGCTGGGCAAAATAAAATTAAAGACCAATTAAAGATGTTGGATTCAGAAGTGTGTTAACCTGGCAACTCTGAGTTCACTGTTTAAATGAGGGTATTAAAGCCCTAACATTCA  
 5410 5420 5430 5440 5450 5460 5470 5480 5490 5500 5510 5520  
  
 CTGAGTATTTCAGCTGTGACATATTAGAAGCATGTAACCTAAAGATCCAAGTTCAGTCTATCCCTACCACTACCAACAGTTCAAGACTGGTTAGGGTACCGCTTGGGTTAA  
 5530 5540 5550 5560 5570 5580 5590 5600 5610 5620 5630 5640  
  
 GTATACTTAATACTCATCTGGAAACTTGTGAAATACAGGGCACTTGAGCTTATCCCAAGAGATCTTGTGTTATAGACAGGGTTCACTGGCCATACTGGACAC  
 5650 5660 5670 5680 5690 5700 5710 5720 5730 5740 5750 5760  
 Alu repeat 3 (-)  
 TGGCTGATCATAGCTCACTGAAGCCCTCAGCTCTGGCTCAAGCACTCTCTGGCTCAGCTTCTGGAGTAGCTGGGACTACAGGTGGCACTTCCACCTGGCTAAATTCTGTT  
 5770 5780 5790 5800 5810 5820 5830 5840 5850 5860 5870 5880  
  
 TTGAGAGATGAGGTCTTGTGTTGGCCAGGCTGGCTCAAACTCCCTGGCTCAAGTTCTTCCGGCCAGCTCTCTGGGTTACAGGAATGAGGCACTGAACTGGCTCC  
 5890 5900 5910 5920 5930 5940 5950 5960 5970 5980 5990 6000  
  
 CCCAGAGATCTGACCAAGGAGGTCAAGTGGCCCAAGAATATGCAATTCTTATACCTACACCCAGATAATTCTAGTTAGTGGGTTCTCACTTCACTTGGGAAAACCTG  
 6010 6020 6030 6040 6050 6060 6070 6080 6090 6100 6110 6120  
  
 CTTAGGTTTAACTCACCTCTGTTGGAGGTAAGTCACTTCACTAACCTAGATACCTATAGGGTGTGAGTCTCTTATAGCTTGGGAGACACAATTAAACTTCTTCTT  
 6130 6140 6150 6160 6170 6180 6190 6200 6210 6220 6230 6240  
 Alu repeat 4 (-)  
 TTTTGAGACAGTTCCCTCTGTTGGCCAGGCTGGAGTCAATGGGCAATCTGGCTCACCGCACCCCTGGCTCCTGAGTCAGCTGAGTCAGCTGAGTCAGCTG  
 6250 6260 6270 6280 6290 6300 6310 6320 6330 6340 6350 6360  
  
 GGATTACAGGCATGCCACCACTGGCCAGCTAAATTGTTAGAGAGCAGGGTTCTCCATGGTGGTCAGGCCCTGGCTCGAACCTCTGACCTCGTGTACCCGCTCAGGCT  
 6370 6380 6390 6400 6410 6420 6430 6440 6450 6460 6470 6480

Figure 2 For legend see page 60.

CCTAAAGTGCTAGGATTACAGGGTGA  
6490 6500 6510 6520 6530 6540 6550 6560 6570 6580 6590 6600

ATTTCAAGGAATCAGATACTTCCAGGGATCAGCAAAAGAGATTGTTGGAGACCTTCCAGTGTCTGGCCTCAACAAAATCTCCAAA  
6610 6620 6630 6640 6650 6660 6670 6680 6690 6700 6710 6720

TGTTTGTTGTTGTTGTTGTTGAGATGGGGTCTCGCTCTGTCACCCAGGGTGGAGCTGGCACTGGCTTGGCTCGCCTCGCCT  
6730 6740 6750 6760 6770 6780 6790 6800 6810 6820 6830 6840

Alu repeat 5 (-)

GCTCAGCCNCTGAGTAGCTGGGATTACAGGGGGTGTCTACCCACCTGGCTAATTTTGTAATTTTGTAAGAGTGGGGTT  
6850 6860 6870 6880 6890 6900 6910 6920 6930 6940 6950 6960

TCTGATCCGCCCTGGCTCCCAAAGTGCTGGATTACAGGTGTGAACCACGTGCCAGCTAACGGCAACTTTAAAGGTCAGATGATGA  
6970 6980 6990 7000 7010 7020 7030 7040 7050 7060 7070 7080

AATAAAAACAAGGAAACTTGGCAAAAGACACTCTGGGAAATTCTGTGGCAACTGCAAATCATGAGGCTATGAAATGGTAACAGG  
7090 7100 7110 7120 7130 7140 7150 7160 7170 7180 7190 7200

CAGACCAGGGAAAGTCACTTCATTAAAGGAAATCTTGGGAGCTGTGGGGTTCTGGAAAGAAGTCATGACTTCAGGGTTAATTCT  
7210 7220 7230 7240 7250 7260 7270 7280 7290 7300 7310 7320

TTTCCTGGCCAGTTAACCTTCTAGGGCCCCCTTCTGCTAATAAGGATACTTTAGTGAATAAGGAGCGTTAACATACTGCTCTGTA  
7330 7340 7350 7360 7370 7380 7390 7400 ← 7440

Exon II

M F A C S K F V S T P S L  
CCCCCTGAAAATGTTGCCTGCTCCAAGTTGCTCCACCTCCCTCTGAGTACCTGCTTTCTGGAGAGTTTAAAGGAGAGACATCTTGCTCTCTCTCCATCAGACCTTGT  
7450 7460 7470 7480 → 7510 7520 7530 7540 7550 7560

TCTTCACAGCAAGGCTGGGGAGAGGGTCCAGGGGGAGATCACITGGAGGCAAGGAGCTGGCAACATGGCGAACCCATCTCTACTA  
7570 7580 7590 7600 7610 7620 7630 7640 7650 7660 7670 7680

Alu repeat 6 (+)

TTGGCAATGGTGGCGGGCCCTGTAACTCCAGCTACTCAGGAGGTGAGGAACAGATACTGAGGCTTGAGGTGAGGCTGAGGCA  
7690 7700 7710 7720 7730 7740 7750 7760 7770 7780 7790 7800

GGCGTCAGAGCAAGACTCTGTCAAAAAAATAAAATAAGTAATAAGTAAGAAAGATTCTGGGGAGGGAGGCCCTGCTATGATGGCTAACAGGATAGAGTGA  
7810 7820 7830 7840 7850 7860 7870 7880 7890 7900 7910 7920

AGATACTCTGCCCATATCCCTGGGGCCAGTTAATTTTTTTTTTGAGACAGAGTTGCGCTCTGTCAGGCTGCAATGGCGGATCCACCAACCTCCG  
7930 7940 7950 7960 7970 7980 7990 8000 8010 8020 8030 8040

Alu repeat 7 (-)

CCTCCGGGTTCAACCACTTCACCTGGCTCAGACTCCAAAGTAGCTGGGGTTACAGGCCACCTGCCACCATACCCAGCTAATT  
8050 8060 8070 8080 8090 8100 8110 8120 8130 8140 8150 8160

CGGGCTGGCTCAACCTCTGACCTCGTGTAGAGACAGGGTTCTCATGTTGGCTAGGCTGGCTCTGCAACTCTGACCTCAGGTGATCCGGGCTCTGGCCCTCTAAATT  
8170 8180 8190 8200 8210 8220 8230 8240 8250 8260 8270 8280

GCTGATATTAGGCGATGAGGCACCGCTGGCTGTGCCAGTTAATTCTACCCAGCTTCTAGGAGAAGAATAGAGATGAGAATCTAGACCAAGATCAAGTTACCTCTG  
8290 8300 8310 8320 8330 8340 8350 8360 8370 8380 8390 8400

AACTATATGTCGTGAACTGGCCATCCAGAGTGGTAATAAGTAGATTACATACCTGGACTGAGTAGCTGACTGTGAGGATCTGTCCTGAGTTAGCC  
8410 8420 8430 8440 8450 8460 8470 8480 8490 8500 8510 8520

TTCTGGGACTTAAATTGGATAGGAAGCTAAATTACATAAATACACATTGGAAAAGTATTGACTCAAGAAACACAGTATGGAAGAAAATT  
8530 8540 8550 8560 8570 8580 8590 8600 8610 8620 8630 8640

GGTCATTCACAAATAAAAAGGAGTACCATTTATTGAGCATTTACCATGTCCTTAGCACTACTGAGTTACTAATTGGTATACATACTCTCA  
8650 8660 8670 8680 8690 8700 8710 8720 8730 8740 8750 8760

ATGATATTGATTTGATTAATGAGAAATCACGGCTTAGAAAGGTTAAGTGAATTACCTCTGAGAATACACTGCTGATAATAGAATTAAAGATT  
8770 8780 8790 8800 8810 8820 8830 8840 8850 8860 8870 8880

CCCTGGCTATGAATCACCATGTCCTCTCTGAGGCTTACGGCTGGGGTAGTCAGAAAAGGCTACCAAGAAGGGTGA  
8890 8900 8910 8920 8930 8940 8950 8960 8970 8980 8990 9000

GTGTCAGGGAAAGTTACTGGCTCTCTGAGGCTGACCATACCCATACAGAAGGGCCCCACTTAAAGTAAAGTTTTTTAACTTATGATGATG  
9010 9020 9030 9040 9050 9060 9070 9080 9090 9100 9110 9120

ATTCATAGAACCATCTGGCTGGGGCTCATGCCCTGTAATCCACCCAGCTGGAAAGCCCAGGGTGGCTGGATCACCTGAGGTAGGAA  
9130 9140 9150 9160 9170 9180 9190 9200 9210 9220 9230 9240

Alu repeat 8 (+)

ATGGTGAACCTGTCCTACTAAATAACAAAATTAGCTGGGGCTGCTTGGCCGGCCCTGTAGTCCACGCTACTTGGGAGGCTGAGGCA  
9250 9260 9270 9280 9290 9300 9310 9320 9330 9340 9350 9360

GTTGACTGAGCCGAGCCAGATCACGACACTGCATTCCAGCTGGCGATACAGCAAGACTCCGCTCAAA  
9370 9380 9390 9400 9410 9420 9430 9440 9450 9460 9470 9480

ATCTTTCACCAAGCTAGCAATGTTGGTACATACTCTGTCATACACTCAGCAGGAAACCCACGCTCCAGTCACCTGCCAATCACAGGGTAA  
9490 9500 9510 9520 9530 9540 9550 9560 9570 9580 9590 9600

Figure 2 For legend see page 60.

GTGTTGCTAGGTGATTTCACCTAACTGTAAAGCTAATGTAAGCTAGTGAAGCTTTGAGCACATTGAAGGCCAGACTATGGTGTTGGTAGGTTAGCAGTATTAATGGTTTTCAAG  
 9610 9620 9630 9640 9650 9660 9670 9680 9690 9700 9710 9720  
  
 TATGATGGATTATTGGCACAATAACCTATGTAAGTTGAGGAGCATCTGTATATATTCCTTGTGAGTGAAGCCAAATGCTCTGAGGCTTAGATTCTCTCTCAAGACCTAT  
 9730 9740 9750 9760 9770 9780 9790 9800 9810 9820 9830 9840  
  
 AAAAGGATATGCAAGGTTGGATGCTCCCCTATGCCCAAAGGAAGTAGACTTCTTCCCTAAACCAAGCTGAGAACAGAACAGAAAATGGTGTGGTCTTTAAATTAGACTTGCCTAA  
 9850 9860 9870 9880 9890 9900 9910 9920 9930 9940 9950 9960  
  
 AAGGAAGAATGATCCAAAACCTAACGTCATGTGAAGATGGAAGGCCATGAGAGGAAGAAAATCTTGTCTGATGATGTAACAACGGCATAGTCTGAAACTCTCCAGTCCTC  
 9970 9980 9990 10000 10010 10020 10030 10040 10050 10060 10070 10080

Exon III

V K S T S Q L L S R P L S A V V L K R P E I L T D E  
 CTTGACCTCCCTGCCCTGTATGTCATTCTGCTAGTCAGAGGCCCTCACAGCTGCTGAGCGCTCGCGCTATCTGAGTGTGCTGAAACGACCGGAGATCTGACAGATGAGGTACCT  
 10090 10100 10110 ← 10130 10140 10150 10160 10170 10180 → 10190

TACACTGGAGTTGGGACTCTGGTTTGCGGGTGTAGCGGTCGGGAGGGTTACCTGCCAGAACCCAGGGTAGAATGCTGAGGCACTCATCACAGGCTAAATTCAATGGCAAGGCC  
 10210 10220 10230 10240 10250 10260 10270 10280 10290 10300 10310 10320

ATGGCTTAAAGTGGAGATGGAGGACACTTAAACCGCTGTATGCTATAGTGTCCAGCTTGGCACTGGCAGCACATGCCCTGTGACTGTGCTAACTGGTATTACTGGCTAGTC  
 10330 10340 10350 10360 10370 10380 10390 10400 10410 10420 10430 10440

TACTGACTTCAGTGGAAAGCAATTAGCTCTAGGAAGATGATCTCCCTGGGAGCAAGAATAGACTAGGGCCCTAGCTTCAACTCTTGTGACCACACCTGAATCTAAAGGCAATTGAA  
 10450 10460 10470 10480 10490 10500 10510 10520 10530 10540 10550 10560

TCTACATCCCTACGTTAAAGGTTCTACTTGCAGCATTAGGGCTTAACAAATTCTGTGTTGGGGAGAACAAATGGGTTGGCAGCTAAAGGTTTAAATGGTACAAAAGTTA  
 10570 10580 10590 10600 10610 10620 10630 10640 10650 10660 10670 10680

S L S S L A V V S C P L T S L V S S R S F Q T S A I S  
 CCAAAGCAAAAAAACTTACCTCTACCTGCCCTTTTCAACAGGCCCTCAGCAGCTGCTGAGCTTCATGCTCCCTAACCTCACTGTCTAGGCCGACCTCCAAACAGGCCATTTC  
 10690 10700 10710 ← 10740 10750 10760 10770 10780 10790 10800

Exon IV

R D I D T A A K F I G A G A A T V G V A G S G G A G I G T V T F G S L I I G Y A R  
 AAGGACATCGCACACAGCAAGCTTCATTGGAGCTGGGCTGCCACAGTTGGGTGCGTGGGACTGTTGGGATGGAACTGTGTTGGAGGCTCATCTGGTTATGCCAGTA  
 10810 10820 10830 10840 10850 10860 10870 10880 10890 10900 10910 →

AGATAATGGACCCCTCACTGGTATCTGATATGCTTCAAGGTCAGAAAATTTGGGGCTAGAACTATACTATCCCACACTGTAGCCACTAGCTACATGTTGGTTGTTGTT  
 10930 10940 10950 10960 10970 10980 10990 11000 11010 11020 11030 11040

GTTGTGAGACAGTAACGCTCTGTTCCCAGGTGGAGTGCAGTGGCCAATCTCAGCTCACTGTACCTCTACCTCTGGGTCAGGGATTCTGTGCTCAGGCTCCAAAGTAGC  
 11050 11060 11070 11080 11090 11100 11110 11120 11130 11140 11150 11160

Alu repeat 9 (-)

TGGGATTACAGGACACACAAATGAGCCCACTTAAATTGTGTTTAGAGATGGGTTTCGCCATGTGAGGCTGGGCTCTAACCTCTGACCTCAAGTGTACCCACCTGGCCT  
 11170 11180 11190 11200 11210 11220 11230 11240 11250 11260 11270 11280

GGCCCTCTAACTGCCAGGATTACAGGCATGAGCCCCGGCTTGCCTGAATTAGTTAAITAATAAAATACATGTTGAGTTCTCAGTCAGCTAACCATATT  
 11290 11300 11310 11320 11330 11340 11350 11360 11370 11380 11390 11400

CGAGTGCCTAGTACCCAGATGGCTAGAGGCTACCATATTCAACATGCAAATATAAGATGTTGGCTGGCACAGTGGCTACTAACTCCACCACTTGGGAGGCCAGGGCAGGTGA  
 11410 11420 11430 11440 11450 11460 11470 11480 11490 11500 11510 11520

TCCCTTGGGTCAAGGAGTCAGGAGCTGGACAGCAGCTGGCCACATGGCAAACCCCATCTCTACTAAAAAATACAAAAAATTACCTGGTAAGGGGGCTGCACCTATAATCCAGCT  
 11530 11540 11550 11560 11570 11580 11590 11600 11610 11620 11630 11640

Alu repeat 10 (+)

GGCTGAGGCAAGAAATCCTGAAACCGGGAGGTGGAGGTGCGAGTGTGAGGCTGAGCTGACAGCTGACAGTGTGACACTTGGGAAAAA  
 11650 11660 11670 11680 11690 11700 11710 11720 11730 11740 11750 11760

AGAAAAAAACAAATATAAGCTGGGGAGGTGGTCAAGGAGCTGGTCAAGGAGGCTGGTCAAGGAGGCTGGTCAAGGAGGCTGGTCAAGGAGGCTGGTCA  
 11770 11780 11790 11800 11810 11820 11830 11840 11850 11860 11870 11880

Alu repeat 11 (+)

AAACCCCTGTCCTGCTAAACAAATACAAAAATTAGCCAGGGCTGTGGTGGGGCCCTGTAGTCCCAGCTACTCGGGAGGATGAGACAGGAGAATTGCTGAAACCCGGAGACAGGGTACA  
 11890 11900 11910 11920 11930 11940 11950 11960 11970 11980 11990 12000

GTGACCCAGATGGCTCCACTCATTCCAGCTGGCCACAGAGTGTAGACCATCAAAAAAAAAAAATTAT  
 12010 12020 12030 12040 12050 12060 12070 12080 12090 12100 12110 12120

ATGTAATGATGATTAGAATGTTATATCATGTGTCAATTGAAGAAAGTTCATGAACAGCTTGGCTAGACCATGTGTTGCAAACTAAGGCCAAATCTGGCCATAGCC  
 12130 12140 12150 12160 12170 12180 12190 12200 12210 12220 12230 12240

AGGTTTGTATGATTAGCTAAATTAATTATGGCCGGGTGGTGCCTACGCTATAATCCCAGACTTGGGAGGCCAGGGCTAACAGGAGTCAAGGAGTTGGAGACTAG  
 12250 12260 12270 12280 12290 12300 12310 12320 12330 12340 12350 12360

Alu repeat 12 (+)

CCTGGGCAACATGGTGAACCCCTCTACTAAAAATACAGAAATTAGCCAGGCATGGTGGAGATGCCCTGATATCCCACTACTCAGGGAGCTGAGGCAGGAAAATCACTTGAACCC  
 12370 12380 12390 12400 12410 12420 12430 12440 12450 12460 12470 12480

GAAGGGAGGTGCAAGTGTAGAGATCAGGCCACTGCACTCAGCTGGCAAAAGAGTGAACCTCAGTCAAAAAAAAAAAATAGTTGCAATTAAAAGGCTTCACTTTGTT  
 12490 12500 12510 12520 12530 12540 12550 12560 12570 12580 12590 12600

TAAAGCAAAAGAGACTATATGCGGCCACAAAACCTAAATTATACATGTGGCCCTTACAGAAAAGTTACCTCCCTGTCTAGAGACTCAGTGAACATAGTGGCTTACTGCT  
 12610 12620 12630 12640 12650 12660 12670 12680 12690 12700 12710 12720

Figure 2 For legend see page 60.

ATTTTCCCACATCTGGAAATTCCCTGTCACTCTGGAAAATGGACTGCAGCCAGCCCCGTTCTGACACCTGGACTGTTAGTACTCCACCTGGATAACTCAGTTA  
 12730 12740 12750 12760 12770 12780 12790 12800 12810 12820 12830 12840

AAACCAAAATTAATCCTAGAGACCAGGAAGTCTCTTAATGCTTTGAGAAATAGAGTCTTTAAGAATTGTATTAAACAGAGTCTGACTGCTGCTTATTCACTATTCTGTT  
 12850 12860 12870 12880 12890 12900 12910 12920 12930 12940 12950 12960

AAATGTTGGTCGATTTACCTAACCATCAAGACTCTGGAGGTATCAGAGTAAGGAAATACAGATTATATATGGCCCTAACACTGGAGTCCTTATCATACTACTCACTCA  
 12970 12980 12990 13000 13010 13020 13030 13040 13050 13060 13070 13080

TAAACCCCATAGACCATTGAACTCTTTTTTTTTGAGACTGAGTCAGCTCGCTGCTCCACAGCTGAAGTGAGTCAGGGACAATCTCAGCCCCACTGCAATCTGCCCTCCGG  
 13090 13100 13110 13120 13130 13140 13150 13160 13170 13180 13190 13200

GTTCAAATGATTCTCCGTGCTCAGCTCCCAAGTAGCTGGATTACAGGTGCCACCCACGCCGGCTAATTTTATTTTATTTTATTTTATTTTATTTGAGACAGAGTC  
 13210 13220 13230 13240 13250 13260 13270 13280 13290 13300 13310 13320

Alu repeat 13 (-)

TGACTCTGTCACCCAGGCTGGAGTCAGTGTAACTCTAGCTACTGCAACCTCTGCCCTCGGGCTCAAGCAATTCTTGCCTCGCCCTCCAGTGGCTGGGATTACAGGTACGC  
 13330 13340 13350 13360 13370 13380 13390 13400 13410 13420 13430 13440

ACCGCTACCCCCAGTAATTTGATTTTTCTTTGAGACAGAGTCCTGCTCTGCCAGGGCTGGAGTACAGTGGCTGCCCTGGATCACTGCATCCCT  
 13450 13460 13470 13480 13490 13500 13510 13520 13530 13540 13550 13560

GCCTCCGGGTTCACGCCCTCTCGCTCAGCTCTGAGTAGCTGGACTACAAGGCCCTGCCAACAGGCCCTGGCTAATTTTGTATTAGTAGAGACAGGGTGTACCGTATT  
 13570 13580 13590 13600 13610 13620 13630 13640 13650 13660 13670 13680

Alu repeat 14 (-)

AGCTAGGATGGCTCGATTCTGACCTCGTACTGCCCTCCAAAGTGTGGGATTACAGACATGAGCCACTGCCAGCAATTGTTGATTTGGTAGAGACACGG  
 13690 13700 13710 13720 13730 13740 13750 13760 13770 13780 13790 13800

TTTACCCAGGTTGGCCAGGCTCTCGAACCTCGACCTCAAGAACATCTACTCATCGCCCTCCAAAGTGTGGGATTACAGGCGTGAAGCCACCGGCCCTGCCCTAATTGTTGATT  
 13810 13820 13830 13840 13850 13860 13870 13880 13890 13900 13910 13920

TTTAGTAGAGACGGGTTTACCATGTTGCCAGGCTGGCTCCAACTCTGGCTCAAGTGATCCGCCTGCTTGGCTCCAAAGTGTGGGATTACAGGTGTAGGCCACCGAACCA  
 13930 13940 13950 13960 13970 13980 13990 14000 14010 14020 14030 14040

GCCCCATTGTTAGTTCTAAAGCCCCAGATCTCTGACTATTGAAATGAGAGAACATACTGTCCCTCTACTCTGTCTCTAGAAGAGCGGTGTTCCATAATCCCTAGGATTCTG  
 14050 14060 14070 14080 14090 14100 14110 14120 14130 14140 14150 14160

AGGTTATGCCAGAGACTGTTAGAGATAAAGGGAGACCAAGCCGTTAAATTCCCCACTACTTTGTACCATCCAGTTGGCTTTAGATGTTACTATATTGGAGTTCTGCT  
 14170 14180 14190 14200 14210 14220 14230 14240 14250 14260 14270 14280

TAAAGTTGAAAAACACTGCTCTAGATAGACCCCTCCATCTATTGGCCCTGGATATTAAAGTGTCTGGCCAGAGGTCTTAATTGTTGTAATGAGATGGGTGAACCAATTAGGAAG  
 14290 14300 14310 14320 14330 14340 14350 14360 14370 14380 14390 14400

TCATGATTACCTGGCCATGTTACAGGATTTAGATTGCCGCTCCCCCTCATTCAGTTCTGTAGAGCCTTGGGAATCAGGCAAGAAATTGGCATGTTGTTACCCCTAA  
 14410 14420 14430 14440 14450 14460 14470 14480 14490 14500 14510 14520

AGCTTCTTATTATGTGAGATAATCTGAGAGGGGATTCTCCCTGAGCCCATCTAGATATTCTCTTCTTGTGTAAGTAAAGTCTCTTCTCTCTCTCTCTACCCAG  
 14450 14450 14450 14450 14450 14450 14450 14450 14450 14450 14450 14450

Exon V

N P S L K Q Q L F S Y A I L G C F A L S E A M G L F C L M V A F L I L F A M \*

GAACCCCTCTGAAACCAACAGCTCTCTCTACGCCATTCTGGCTTGGCCCTCGGAGCCATGGGCTTGTGCTGATGGTAGCCCTCTCATCTCTTGGCATGTAAGGAG  
 14460 14460 14460 14460 14460 14460 14460 14460 14460 14460 14460 14460

CCGCTCCACCTCCCAGTTCTCCCGCTCTGGTGGCCCGTGTGTTCTTCTATACCTCCCCAGGCAGCCTGGGAACGTGGTGGCTCAGGGTTGACAGAGAAAAGACAAAT  
 14470 14470 14470 14470 14470 14470 14470 14470 14470 14470 14470 14470

AAATACTGTATTAATAAGATGTTCTGAGTCCTCTGTATATTCTTCCACAGTTGGCTGAGTGTGCTCGTGAAGAGTACAAGGCCAGGGTACTGATGGCTAAACTCAACAT  
 14480 14490 14490 14490 14490 14490 14490 14490 14490 14490 14490 14490

GGATTTGGCTGAGCTC

*Sac I*

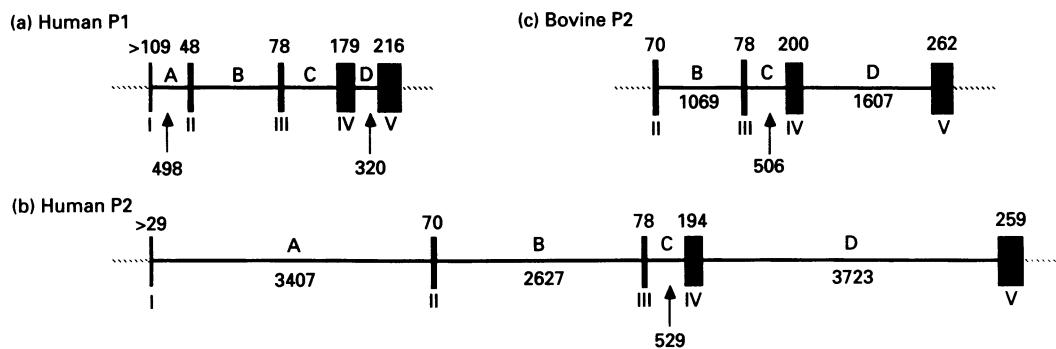
Figure 2 DNA sequence of a fragment of human DNA containing the P2 gene

Exon I (double lines above and below) is homologous to the 5' sequence of a human P2 processed pseudogene described here, and to an ovine P2 cDNA (Medd et al., 1993) from the site marked by A. Protein sequences are shown above exons II–V, and exon–intron boundaries are denoted by small arrows. In the proposed promoter region are three copies (in boxes) of the sequence GGGCGG and its complement. The doubly underlined sequence (nucleotides 14878–14883) is a polyadenylation signal (Proudfoot and Brownlee, 1976), although the exact position of polyadenylation is not known. The Alu repeats on the displayed and complementary DNA strands are denoted by (+) and (–) respectively. The *Pst*I and *Sac*I restriction enzyme sites used in the cloning of this region are shown at the extremities of the sequence. The partial human P2 cDNA sequence (Farrell and Nagley, 1987) corresponds to the coding sequence from bases 10860–14822. The cDNA is reported to have the additional sequences CGGCTCTCA and TCA at its 5' and 3' extremities, but they are not in the genomic sequence.

to link its two transmembrane  $\alpha$ -helices (Sebald and Hoppe, 1981). The presence of introns in segments of DNA coding for links between transmembrane  $\alpha$ -helices has been observed in genes for other intrinsic membrane proteins, including bovine and human rhodopsins (Nathans and Hogness, 1984), mouse

band III protein from the red cell membrane (Kopito et al., 1987) and ADP/ATP translocase (Cozens et al., 1989). It is consistent with the general view that exons often encode structural domains of proteins (Gilbert, 1978; Blake, 1979).

The nucleotide sequences adjacent to the 5' and 3' boundaries



**Figure 3** Structures of human P1 and P2 genes and the bovine P2 gene for the precursors of the c subunit of mitochondrial ATP synthase

In the human genes, exons I–V and introns A–D are represented by solid boxes and continuous lines respectively. The sizes of exons and introns are given in bp. Human P1 may have two promoters, one to initiate transcription from the 5' end of exon I, and the second close to the 5' boundary of exon II. The transcriptional initiation sites have not been determined experimentally. The known bovine P2 gene sequence does not extend into exon I (Dyer et al., 1989).

**Table 1** Exon sizes in genes for the c subunit of mitochondrial ATP synthase

Gene	Exon length (bp)				
	I*	II	III	IV	V
Human P1	(109)	48†	78	179	216
Human P2	(29)	70	78	194	259
Bovine P2	—	70	78	200	262

\* Parentheses indicate that the lengths of exons I have not been determined accurately, and that these are minimal estimates based upon cDNA sequences.

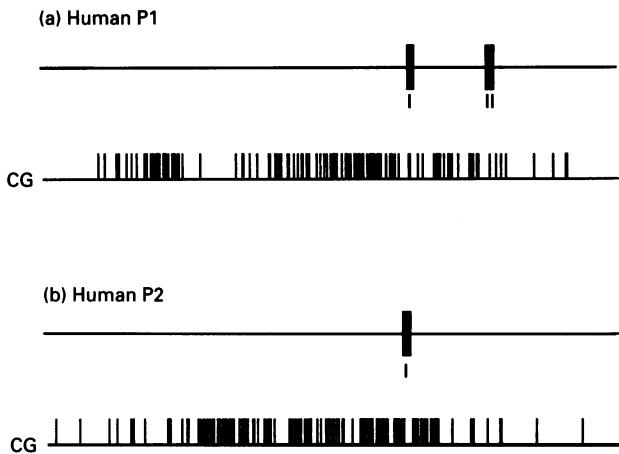
† It is suggested in the text that a sequence in intron A could promote transcription. If this is true in humans then exon II can also be 67 bp long.

of all of the introns in the human P1 and P2 genes, and also in the bovine P2 gene, are conserved (Table 2). They begin with the dinucleotide GT and end with AG, and so agree exactly with the consensus sequences adjacent to splice junctions (Breathnach and Chambon, 1981). Furthermore, the conservation extends for an additional 8–10 bp from the splice junctions in the sequences of the introns, and these extended sequences also agree with the consensus for sequences around splice sites (Mount, 1982). The classes of exon–intron boundary within homologous exons in both human genes (and also in the bovine P2 gene) are conserved (see Table 2). Extensive sequences are conserved within introns of human and bovine P2 genes (results not shown), indicating that they may be under evolutionary constraint.

There are probably more than  $10^5$  *Alu* repeats in the human genome, representing 5–6% of its DNA (Rinehart et al., 1981). They are usually about 300 bp long, and are dimeric structures

**Table 2** Introns in mammalian pre-proteolipid genes

Gene	Intron	Size (bp)	Class	Sequence	
				5' boundary	3' boundary
Human P1	A	498	—	gtg.cag.GTGACTTGGG	CCCTCTGCAG.act.gaa
Human P2	A	3407	—	gag.cag.GTAAGGCCTT	GTAATTCCAG.ctc.tcc
Human P1	B	915	0	gct.ctg.GTAAGGTGCC	GATTTTACAG.atc.cgc
				A L	I R
Human P2	B	2627	0	tcc.ttg.GTGAGTACCT	TTCCTGCTAG.gtc.aag
				S L	V K
Bovine P2	B	1069	0	tcc.ttg.GTGAGTACCC	TTCCGGCTAG.atc.agg
				S L	I R
Human P1	C	706	0	aaa.cag.GTAAGGGAGG	CTCTTCTAG.cct.tcc
				K Q	P S
Human P2	C	529	0	gat.gag.GTACCTTACA	TTTTTCACAG.agc.ctc
				D E	S L
Bovine P2	C	506	0	gat.gag.GTACCTTACA	TTCTTCACAG.agc.cac
				D E	S H
Human P1	D	320	2	gcc.ag.GTAAGTTGG	TCCCTCCCAG.g.aac
				A R	N
Human P2	D	3723	2	gcc.ag.GTAAGATAAG	CTTCTACCAG.g.aac
				A R	N
Bovine P2	D	1607	2	gcc.ag.GTAAGATGGG	CCCCTCCCAG.g.aac
				A R	N
Consensus sequence				cagGTAAGT	YYYYYYYYYYNCAGg



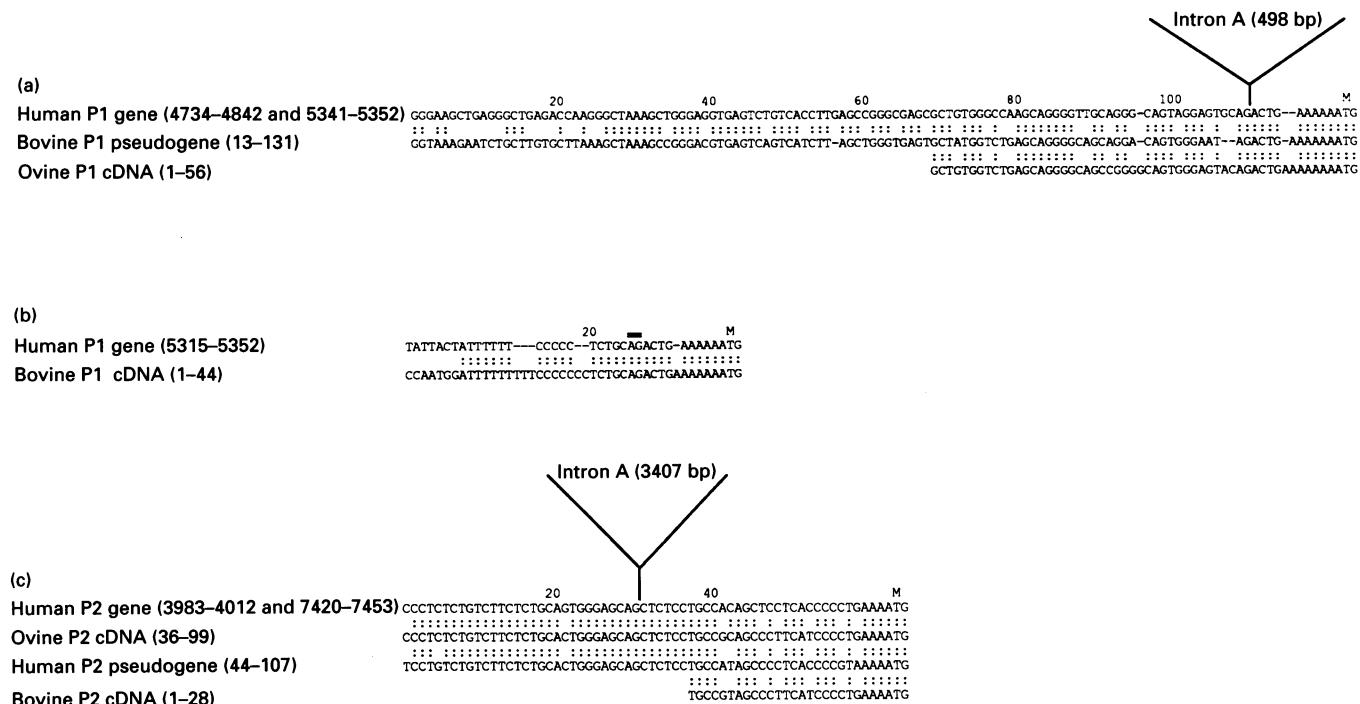
**Figure 4 Distribution of the dinucleotide CpG in the 5' regions of the human P1 and P2 genes**

The vertical lines mark each CpG in (a) nucleotides 2250–6250 of the human P1 gene, and (b) nucleotides 1500–5500 of the human P2 gene. The horizontal and solid lines indicate non-coding regions and exons respectively.

which have apparently formed from internal deletions and dimerizations of 7SL RNA (Ullu and Tschudi, 1984). The two segments of human genomic sequence encompassing the P1 and P2 genes (Figures 1 and 2) contain 10 and 14 examples respectively, some in introns and others in flanking sequences. In each DNA sequence four of the repeats are clustered in pairs. *Alu* repeat 2 in intron A of the human P2 gene is exceptional. It is 102 bp long and contains only the 3' monomeric unit. The B1 family of repeated DNA sequences in rodents have similar structures (Rogers, 1985).

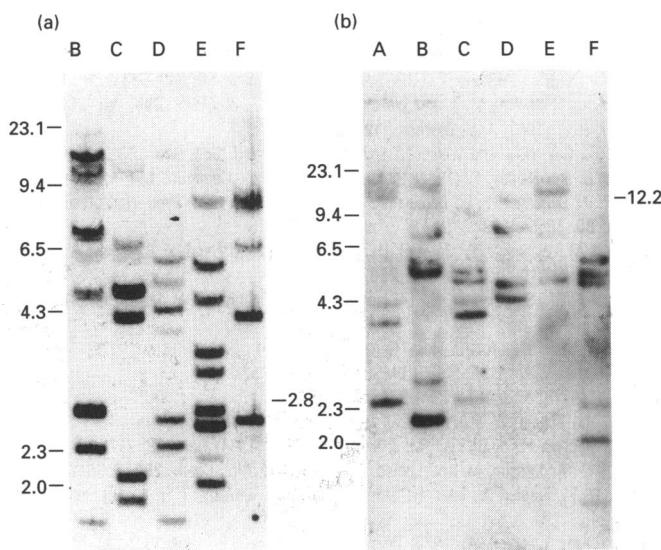
#### Transcription of P1 and P2 genes

Within their 5' regions and extending over exons I, the human P1 and P2 genes have CpG-rich islands (Bird, 1986) of 2 and 1.5 kb long respectively (Figure 4). Transcription probably initiates in these islands, but the transcriptional start sites for neither gene have been determined experimentally. However, the 5' sequences determined in the P1 and P2 cDNAs in cows (Gay and Walker, 1985) and sheep (Medd et al., 1993) help to pin-point these sites. Since processed pseudogenes are believed to have arisen by a process that involved reverse transcription of mRNAs (Rogers, 1985; Weiner et al., 1986), further clues are to be found in the



**Figure 5 Comparisons of DNA sequences in the 5' non-coding regions of the human P1 and P2 genes, in the bovine and ovine cDNAs and in related pseudogenes**

The positions of the sequences in the determined sequences are given in parentheses on the left. Identities are denoted by colons (:). The positions of the translational initiator methionines are denoted by M. In (a), part of the human genomic sequence is aligned with the 5' region of a bovine P1 processed pseudogene immediately following its 5' flanking direct repeat sequence (Dyer et al., 1989), and with the 5' untranslated region of an ovine liver cDNA for P1 (Medd et al., 1993). The position of intron A is shown. In (b), a sequence in the 5' untranslated region of a bovine P1 cDNA is aligned with a different sequence in the human gene that is found adjacent to the 5' boundary of exon II (see Figure 1). The dinucleotide AG with a bar above it could be used as a 3' splice site in a putative human pre-mRNA initiated upstream of exon I. This would result in an mRNA similar in structure to the ovine mRNA. If, as proposed, a second promoter is found in the sequence preceding exon II, then the 3' region of intron A (P1 gene) codes for the 5' untranslated region of a human mRNA that is related to the bovine P1 mRNA from heart. In (c), the human P2 gene is compared with the 5' untranslated region of an ovine liver cDNA. The latter contains a run of T and C residues at nucleotides 1–35 which is not related to either the human genomic sequence or the human P2 processed pseudogene. It is possible either that this TC-rich sequence is a cloning artefact, or that the ovine sequence is unrelated over this stretch. The remainder of the 5' untranslated region of the sheep P2 mRNA is aligned with the human genomic sequence. A sequence from a human P2 pseudogene (see Figure 7) immediately downstream from its 5' flanking repetitive sequence is also shown, as is the entire 5' untranslated region present in a bovine heart P2 cDNA (Gay and Walker, 1985). In the human genomic sequence the position of intron A is indicated. These proposals concerning the transcription of the P1 and P2 genes have not yet been tested by transcriptional mapping studies.



**Figure 6** Hybridization of human DNA with specific DNA probes for the pre-proteolipid genes P1 and P2

The probes are nucleotides 404–558 and 406–615 of the bovine cDNAs for P1 and P2 respectively (Gay and Walker, 1985). Human placental DNA (20 µg) was digested with the restriction enzymes *Bam*HI (lane A), *Eco*RI (lanes B), *Hind*III (lanes C), *Nco*I (lanes D), *Sac*I (lanes E) and *Xba*I (lanes F). The fragments were fractionated by electrophoresis in a 0.6% agarose gel and then were hybridized on nitrocellulose filters to prime-cut probes for P1 (a) and P2 (b). The filters were washed in 0.2 × SSC at 65 °C and then autoradiographed at –70 °C for 72 h. In (a), lane B, an *Eco*RI fragment of 2.8 kb is observed; subsequently the DNA sequence of human P1 was found to contain an *Eco*RI fragment of this size. In (b) lane E, a *Sac*I fragment of 12.2 kb is indicated; a fragment of the same size was sequenced from the DNA of λAT5P2.1. Marker and fragment sizes are in kb.

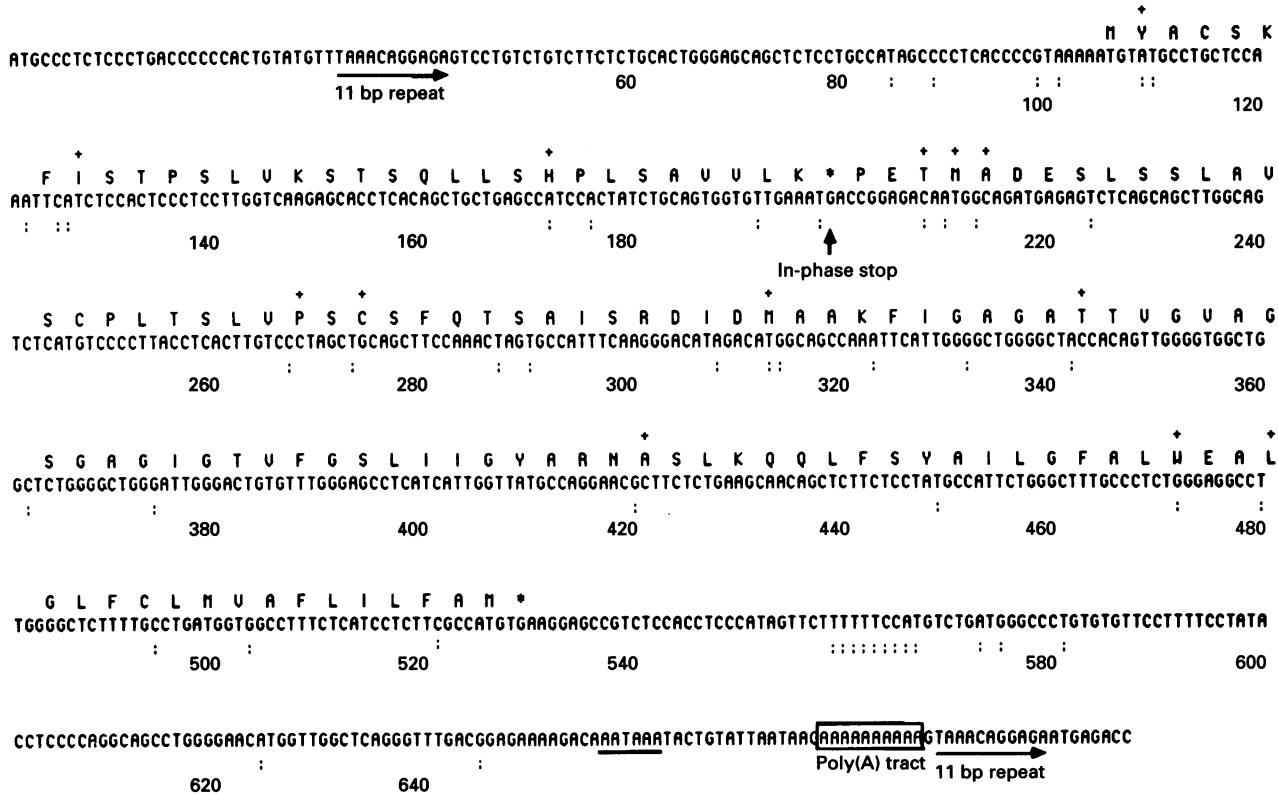
sequences immediately downstream of the 5' flanking repeated sequences of a human P2 and a bovine P1 processed pseudogene (Dyer et al., 1989; see Figures 5a and 5b).

In the human P1 gene, the available information (Figure 5a) indicates the presence of two independent transcriptional initiation sites. These alternative promoters could be used to regulate expression of the gene in various tissues. The transcription of the human P2 gene appears to be simpler. All of the available information (Figure 5c) is consistent with a single transcriptional initiation site in the vicinity of nucleotide 3984.

The 3' limits of transcription of the human P1 and P2 genes are more readily discerned. Human P1 has the uncommon polyadenylation signal, ATTAAA (Berget, 1984; Martini et al., 1986), which is also used in the bovine P1 gene (Gay and Walker, 1985). The more usual polyadenylation signal, AATAAA (Proudfoot and Brownlee, 1976), is found 122 bp and 125 bp respectively after the termination codons in both the bovine and human P2 genes. Poly(A) addition to the human transcripts probably occurs within 11–13 nucleotides, to the 3' side of these sequences.

#### Number of human genes for P1 and P2

Previous studies of bovine cDNAs (Gay and Walker, 1985), together with the work presented in this paper, have shown that both the human and bovine genomes contain at least two expressed genes for the dicyclohexylcarbodi-imide-reactive proteolipid subunit of mitochondrial ATP synthase. In addition, numerous spliced pseudogenes have been detected in both animals, and these observations are consistent with the complex Southern blots obtained with digests of both bovine and human



**Figure 7** Sequence of a human processed pseudogene for the mitochondrial pre-proteolipid P2

Colons and crosses indicate the 50 differences in nucleotide sequence and 13 differences in protein sequence respectively between the pseudogene and the coding regions and protein sequence of human P2. The position of an in-phase stop codon is indicated by a vertical arrow. The underlined sequence is a poly(A) addition signal (Proudfoot and Brownlee, 1976; Gay and Walker, 1985). The following poly(A) tract is boxed. The direct 11 bp repeated sequences which flank the pseudogene are indicated by horizontal arrows.

DNA (see Figure 6). During the course of the cloning and sequencing experiments described above, the complete sequence of a P2 pseudogene (Figure 7) was determined from the overlapping recombinants  $\lambda$ HP2.8 and  $\lambda$ HP2.13. Several features of this sequence support the view that it arose by reverse transcription of the P2 mRNA, followed by recombination into the human genome. For example, the sequence is flanked by two direct 11-nucleotide repeats, and the direct repeat at the 3' end of the pseudogene is preceded by a potential polyadenylation signal and the sequence A<sub>10</sub>. Also, the pseudogene sequence differs in 50 nucleotides from the human P2 cDNA sequence deduced from the gene. This causes 13 substitutions in the amino acid sequence and introduces an in-phase stop codon. As described in the following paper (Medd et al., 1993), an intronless P2 pseudogene in the sheep genome is transcribed, and an intronless human gene encoding phosphoglycerate kinase has been shown to express the protein, but only in testis (McCarrey and Thomas, 1987). Therefore it is conceivable that some of the other processed P1 and P2 sequences in the human genome may not be pseudogenes, as we have tended to assume, but may be functional retroelements also.

The work described in this paper has a direct bearing on the fatal disease, ceroid lipofuscinosis, found in man and other mammals. In the juvenile and late-infantile forms of the human disease, and in the sheep disease (Fearnley et al., 1990), the affected individuals accumulate large amounts of the c subunit of mitochondrial ATP synthase in lysosomes. The accumulated material appears to be chemically identical to the protein normally found in mitochondria (Palmer et al., 1992). In diseased sheep the P1 and P2 cDNAs are identical in sequence to those from normal animals, and the amounts of mRNAs for both P1 and P2 are unaffected in the diseased animals (Medd et al., 1993). Therefore the disease appears not to involve mutation of the coding sequences of the P1 and P2 genes. Similar investigations have not been conducted in humans, but the gene for the juvenile form of human ceroid lipofuscinosis maps to the long arm of chromosome 16 (Gardiner, 1992), whereas the human P1 and P2 genes are on human chromosomes 17 and 12 respectively (M. R. Dyer and J. E. Walker, unpublished work).

We thank Dr. T. H. Rabbits (of this laboratory) for supplying us with samples of human genomic libraries. M. R. D. was supported by an M.R.C. research studentship and an M.R.C. research training Fellowship.

## REFERENCES

- Bachman, N. J., Lomax, M. I. and Grossman, L. I. (1987) *Gene* **55**, 219–229  
 Benton, W. D. and Davis, R. W. (1977) *Science* **196**, 180–182  
 Berget, S. M. (1984) *Nature (London)* **309**, 179–182  
 Biggin, M. D., Gibson, T. J. and Hong, G. F. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 3963–3965  
 Bird, A. P. (1986) *Nature (London)* **321**, 209–213  
 Blake, C. C. F. (1979) *Nature (London)* **277**, 598  
 Breathnach, R. and Chambon, P. (1981) *Annu. Rev. Biochem.* **50**, 349–383  
 Cozens, A. L., Runswick, M. J. and Walker, J. E. (1989) *J. Mol. Biol.* **206**, 261–280  
 Deininger, P. L. (1983) *Anal. Biochem.* **129**, 216–223  
 Dyer, M. R., Gay, N. J. and Walker, J. E. (1989) *Biochem. J.* **260**, 249–258  
 Farrell, L. B. and Nagley, P. (1987) *Biochem. Biophys. Res. Commun.* **144**, 1257–1264  
 Farrell, P. J., Deininger, P. L., Bankier, A. and Barrell, B. G. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 1565–1569  
 Fearnley, I. M., Walker, J. E., Jolly, R. D., Martinus, R. D., Kirkland, K. B., Shaw, G. J. and Palmer, D. N. (1990) *Biochem. J.* **268**, 751–758  
 Forster, A., Huck, S., Ghanem, N., LeFranc, M. P. and Rabbits, T. H. (1987) *EMBO J.* **6**, 1945–1950  
 Gardiner, R. M. (1992) *Am. J. Hum. Genet.* **15**, 539–541  
 Gay, N. J. and Walker, J. E. (1985) *EMBO J.* **4**, 3519–3524  
 Gilbert, W. (1978) *Nature (London)* **271**, 501  
 Jackl, G. and Sebald, W. (1975) *Eur. J. Biochem.* **54**, 97–106  
 Karn, J., Matthes, H. W. D., Gait, M. J. and Brenner, S. (1984) *Gene* **32**, 217–224  
 Kopito, R. R., Andersson, M. and Lodish, H. F. (1987) *J. Biol. Chem.* **262**, 8035–8040  
 LeFranc, M. P., Forster, A., Baer, R., Stinson, M. A. and Rabbits, T. H. (1986) *Cell* **45**, 237–246  
 Macino, G. and Tzagoloff, A. (1979) *Proc. Natl. Acad. Sci. U.S.A.* **76**, 131–135  
 Maniatis, T., Fritsch, E. F. and Sambrook, J. (1982) in *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Press, Cold Spring Harbor, New York  
 Martini, G., Toniolo, D., Vulliamy, T., Luzzatto, L., Dono, R., Viglietto, G., Paonessa, G., D'Urso, M. D. and Persico, M. G. (1986) *EMBO J.* **5**, 1849–1855  
 McCarrey, J. R. and Thomas, K. (1987) *Nature (London)* **326**, 501–505  
 Medd, S. M., Walker, J. E. and Jolly, R. D. (1993) *Biochem. J.* **293**, 65–73  
 Messing, J. (1983) *Methods Enzymol.* **101**, 20–78  
 Mills, D. R. and Kramer, F. R. (1979) *Proc. Natl. Acad. Sci. U.S.A.* **76**, 2232–2235  
 Mizusawa, S., Nishimura, S. and Seela, F. (1986) *Nucleic Acids Res.* **14**, 1319–1324  
 Mount, S. M. (1982) *Nucleic Acids Res.* **10**, 459–472  
 Naora, H. and Deacon, N. J. (1982) *Proc. Natl. Acad. Sci. U.S.A.* **79**, 6196–6200  
 Nathans, J. and Hogness, D. S. (1984) *Proc. Natl. Acad. Sci. U.S.A.* **81**, 4852–4855  
 Ohta, S., Yohda, M., Ishizuka, M., Hirata, H., Hamamoto, T., Otawara-Hamamoto, Y., Matsuda, K. and Kagawa, Y. (1988) *Biochim. Biophys. Acta* **933**, 141–155  
 Palmer, D. N., Fearnley, I. M., Walker, J. E., Hall, N. A., Lake, B. D., Wolfe, L. S., Haltia, M., Martinus, R. D. and Jolly, R. D. (1992) *Am. J. Med. Genet.* **41**, 561–567  
 Proudfoot, N. J. and Brownlee, G. G. (1976) *Nature (London)* **263**, 211–214  
 Rinehart, F. P., Ritch, T. G., Deininger, P. L. and Schmid, C. W. (1981) *Biochemistry* **20**, 3003–3010  
 Rogers, J. H. (1985) *Int. Rev. Cytol.* **93**, 187–279  
 Sanger, F., Nicklen, S. and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463–5467  
 Sebald, W. and Hoppe, J. (1981) *Curr. Top. Bioenerget.* **12**, 2–64  
 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. and Hollaender, A., eds.), pp. 67–114, Plenum Publishing Corporation, New York and London  
 Turner, G., Imam, G. and Kuntzel, H. (1979) *Eur. J. Biochem.* **97**, 565–571  
 Ullu, E. and Tschudi, C. (1984) *Nature (London)* **312**, 171–177  
 Walker, J. E., Gay, N. J., Powell, S. J., Kostina, M. and Dyer, M. R. (1987) *Biochemistry* **26**, 8613–8619  
 Walker, J. E., Lutter, R., Dupuis, A. and Runswick, M. J. (1991) *Biochemistry* **30**, 5369–5378  
 Weiner, A. M., Deininger, P. L. and Estratiadis, A. (1986) *Annu. Rev. Biochem.* **55**, 631–661  
 Wilbur, W. J. and Lipman, D. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 726–730