Proc. Natl. Acad. Sci. USA
Vol. 83, pp. 4243–4247, June 1986
Biochemistry

Structure of a bovine gene for P-450c21 (steroid 21-hydroxylase) defines a novel cytochrome P-450 gene family

(adrenal gland/steroid hormone/molecular evolution/DNA sequence)

BON-CHU CHUNG, KARLA J. MATTESON, AND WALTER L. MILLER*

Department of Pediatrics and Metabolic Research Unit, University of California, San Francisco, CA 94143

Communicated by Bruce M. Alberts, February 27, 1986

ABSTRACT **P-450c21, a cytochrome P-450 enzyme [steroid 21-monooxygenase (steroid 21-hydroxylase), EC 1.14.99.10], mediates the 21-hydroxylation of glucocorticoid and mineralocorticoid hormones in the adrenal gland. The complete sequence of a bovine P-450c21 gene shows it is 3447 base pairs long and contains 10 exons. The intron/exon organization and encoded amino acid sequence indicate that P-450c21 represents a unique family of genes in the P-450 gene superfamily. Primer extension and S1 nuclease protection experiments identified several cap sites for initiation of transcription; the principal cap site produces mRNA with a 5′ untranslated region only 11 bases long. S1 nuclease protection experiments confirm that P-450c21 is actively expressed in the adrenal and the testis, an organ not known to secrete 21-hydroxylated steroids.**

21-Hydroxylation is a key enzymatic step in the synthesis of glucocorticoid and mineralocorticoid hormones. The 21-hydroxylation of progesterone to desoxycorticosterone or of 17-hydroxyprogesterone to 11-desoxycortisol is mediated by P-450c21, a cytochrome P-450 enzyme [steroid 21-monooxygenase (steroid 21-hydroxylase), EC 1.14.99.10] bound to endoplasmic reticulum. 21-Hydroxylation was the first enzymatic activity ascribed to any cytochrome P-450 (1). Cytochrome P-450s metabolize xenobiotic agents and endogenous steroid substrates. Although cytochrome P-450s are related in size, sequence, and spectral characteristics, they are encoded by several widely divergent families of genes belonging to a large gene superfamily (2). P-450c21 is of great medical interest as about 1 in 5000 persons has an autosomal recessive disorder of P-450c21 causing congenital adrenal hyperplasia, variously manifested by genital ambiguities, virilization, cardiovascular collapse, and death (3). We recently identified a bovine genomic DNA clone containing a P-450c21 gene and showed that cattle have two copies of this gene and synthesize two sizes of P-450c21 mRNA in the adrenal cortex (4). We now report the sequence of >6.5 kilobases (kb) of bovine genomic DNA containing this gene. Nuclease S1 protection experiments indicate this gene is active in the bovine adrenal cortex and, surprisingly, in the testis, which is not known to secrete 21-hydroxylated steroids. Analysis of the structure of this gene indicates that this P-450c21 gene is only distantly related to other P-450 gene families and hence represents a novel member of the P-450 gene superfamily.

MATERIALS AND METHODS

DNA sequencing was done by the dideoxy method using ^{35}S-labeled dNTPs and gradient gels (5).

The 27-base primer 3′ GACGATCGACCGCGAGTG-GACGATACC 5′ complementary to bases 45–71 was mechanically synthesized, purified, and labeled (4). For primer extension, 0.1 pmol of ^{32}P-labeled primer and 10 μg of bovine adrenocortical poly(A) RNA were hybridized for 2 hr at 65°C in 10 μl of 0.3 M NaCl/5 mM EDTA/10 mM Tris·HCl, pH 8.3, and then made to 50 μl, having final concentrations of 1 mM (each) dNTP, 10 mM Tris (pH 8.3), 10 mM dithiothreitol, 10 mM $MgCl_2$, 640 units of human placental ribonuclease inhibitor (Amersham) per ml, and 500 units of reverse transcriptase (Life Sciences, St. Petersburg, FL) per ml. After 30 min at 42°C, the reaction was stopped with 4 μl of 0.25 M EDTA.

S1 nuclease mapping was modified from Sudhof *et al.* (6). One picomole of ^{32}P-labeled 27-mer was used to initiate DNA synthesis on an M13 clone of a 412-base *Hinc*II/*Pst* I fragment of the P-450c21 gene corresponding to nucleotides −275 to +137 employing unlabeled dNTPs and the Klenow fragment of DNA polymerase I. The reaction proceeded 30 min at 25°C and then was heated to 65°C for 10 min. After *Bam*HI digestion, the labeled, 359-base oligonucleotide-initiated probe was gel-purified (4). About 10^4 cpm of 359-base probe and 50 μg of total RNA in 20 μl of 80% deionized formamide/0.4 M NaCl/1 mM EDTA/40 mM Pipes, pH 6.4, were denatured 10 min at 85°C and then hybridized 3 hr at 48°C. One hundred eighty microliters of 0.2 M NaCl, 30 mM NaOAc (pH 5.5), 1 mM $ZnSO_4$, 20 mg of denatured salmon sperm DNA per ml, and 100–300 units of S1 nuclease (Bethesda Research Laboratories) was added and incubated 1 hr at 37°C.

RESULTS AND DISCUSSION

Structure of the P-450c21 Gene. We manually synthesized 23- and 30-base oligonucleotides (4) corresponding to different regions of a 486-base bovine P-450c21 cDNA fragment (7) and used these to screen a bovine genomic DNA library. Clone λE11 was shown to contain P-450c21 sequences by initiating dideoxy sequencing directly on full-length double-stranded phage without subcloning into M13 (4). The bovine genomic DNA in λE11 was mapped and subcloned into pUC13 as a series of overlapping clones. These pUC13 clones were mapped further and appropriate fragments were subcloned into M13 for sequencing (Fig. 1). As we had no cDNA probe, coding regions and intron/exon boundaries were identified by maintenance of an open reading frame and by comparison to canonical splice-junction site sequences (8), since no exceptions to the canonical splicing rules have been found in any P-450 gene (9–15). Comparison of our sequence with the complete amino acid sequence of porcine

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: kb, kilobase(s); bp, base pair(s).

*To whom reprint requests should be addressed at: Department of Pediatrics, S-677, University of California, San Francisco, CA 94143.

```
a
-3224 GGATCCCTTTATCAGCAGTATTACGTAACTTTCCCAGAAATGTGCAGTCATCAGGTCCGATGATGTCTTCATCCAGATACTGATTTCAAACAGACCAAAAGGCCGACTTTGGTGTTTCTTA
-3103 ATACTCTGAGGGAACTGCTGTTCTCGGGGTTAACCCTCCCCACATGCCTGCCTAGCTCTCCCCCATTATGGCAGCACCTCTTGGATCTTCAGGGCTAACCTAGAAATTCACTCTAGAGGAA
-2982 TGTGCCAAGGTCATTTTTGAGGCAAACTGACCTAACTCAGCAGCCTGCCATTCAGTCATTCAACAGGAATTGTGGAGCTCTCACTCCAGGCCATGCCCTGCTCCAGGTCCCCATCTTCTGT
-2861 GAATGGTGTGAGGTCAGGGACCTGCATGGTTCTGTGGGGCCTCTGACCAGTCTGGGTGCATGTGAGAAGTGGGCTGGGGAGATGACCTGTAAGTTGGACTCGGCAGGCTGAAAAGTCAAGC
-2740 AAAGGGAGTAGGAGTAGATGAGGGTATTTGAGGCAGAAGAAAGAACATAAAAGGCCCCAAGGCTGGAGGCAAGACAGCATTCTTGGGGGCATTGCATATCATTTATCTTGTTTCGGAATTA
-2619 GGATTTGGAGGTGGTCTACGAGAGAGGAGTTGCAAGTACAGGTGTCGACAAGCAGGTCAGCAGGTCCAGGTCACTCTCCATTCGCAACAGAGTCCCCTCCCATTCCATGAAACTCTGAGTA
-2498 GTTCTGACCTGGGAATTTTCTGGTGGCCCAGTGGCTATGACTTTGTGCTACCAATACAAGGGGTCCAGTTTGATCCCTGGTCAGGGAACTAATTCCTACATGCCACAAATAAGAGCCTCAA
-2377 CTGAAAACCAAAAAAAATCCTGCATGTGCAACTGCAGATCTAAGATTCCCAATGCTGCAACTAAGATCATGTGTAATCAAATAAGTTAATATTTTTTACAAAAGCGTTCTGACCTTCCTC
-2256 AGTGTTTCACAGGCTTCTTAGGGCTATTGTCTACATGGCCTGGTCATTGCCCATAAGATGTCTCCTGAAATCTGGTCAGCATGGTGGTGGCATAAAGCCACGTACGAGAGTCAAATTTTCA
-2135 GCTATAGTTGTGGCAAGACAACATCTAGGGCCTGAGGTCGAC........................approximately 200 bases ....................CAGC
-1892 TGGGATCTTGACCCAGGCTCTCTGGTGGCAGAGTCTTGCTCTGATTAAAGCTTTGAGGTCAGTTGGCAGAGAAAGGATGGGGTATTGGGGATGTGTGGGCTCAGGTCAGCTCTAAGCCCT
-1771 CACCCCACCTCCTAGGTGCCCCAGCCTCCCACCGAGTCCTGCCACCCCTCTTTAGCATGCTACTGCTACTGCTGCTAACTCACTTCAGTTGTGTCCGACACTGTGTGACCCCATAGACAGT
-1650 GGGCTGCCCTATCCCCCATCCCTGGGATTATCAGCCAAGAACATTGGAGTGGTTTGCCATTTCCTTCTCCAATGCAGGACACTGAAAAGCAAAAAGTGAAGTCGGTCAGTTGTGTCCAACTC
-1529 TTAGTGACCCCATGGACTGCAGCTACCAGGCTCCTCTGTCCATGGATTTTCCAGGCAAGAGTACTGGAGTGGGGTGCCATTGCCTTCTCCGCTCTTTAGCATGACCTCTCCTTTAAGTCC
-1408 TTCAACTGCCCCCTCACCTGCAACAGTGGCCCCTGATAGCCCTACTTAGTCCACTGGGACCCCACCCCTCATGGCAATAGCTCCCTGAAGAGGGCTGGCCCCCTGATCCCAGCTGTGGCTA
-1287 TCAAGTTCCTTCCTGAGGACTTTGGAATCCATGGGGGCATGTTTGATCACGCCATTACAGCCTTTTTGTCCTGGGGCTAGACCCGTATTCTGTGGAGTCTAACCTTGAAAGCAGTTCTGTG
-1166 CCGCCTGCACCTGGGAGCACAAGCAGCTTGACTGCAGACAGAGAGGAGGGAGGGACTTCCCTGGTGGTCCATTGGTTAACAATCCACCTGCCAATTAAAAAAAGGGGTTGGATCCCTGGT
-1045 CTGGGAAGATCACATGCCGTGAAGCAACCAGGCCTGTGGCCCCCACGACTGAGCCCGAGTTCTAGAGCCCATGAGCCTCAATTGCTCGAAGCCCACATGCCTGGGTGTCACAACCTAGAG
- 924 AGTACCCTCTGCTCGCCGCAACTTAGAGAAAGCGCAGGCCATGGCAATACAGGCCCAGTGCAGCCAAGATTAATAAACAAACGAAAGAGGAGGAACGAAGAGTGAGGAAGCAGCCATGTCT
- 803 CTAGGGCTCTTTAGTTCACAATCCCTGACTTTCCAAGGCCAAGAAGGTCCCTTCTACTTGCCTGCTCTGGATTCTGTGAGATATCCCTCTATTTGTAGAACAATTCTTACTTTTGGCTCGA
- 682 AACTAGCTCAAGTGGGTCTCTCTTCATTGCAAGCAAAGATCGTCAGGTAGAGGTAAGAGCTCCCCCAAAACTGGGAAACATTTAATTCCTCATCTTCAAGTTTTGAGAATAAAAAGAGGC
- 561 AATATTACTGATGACACTTGGACTATAGGTATAAACTGGATAACACATGCAAACCGGGACATGTGGTTCCCCTATCTGTAAGTTAAAAGAAAAACACACAGATCATTCTCATCTCATTTT
- 440 GTGCTTCAAAGGCCATCTGGCAGGGCATATTAGGCGTGGGTTGTCATTCTTCTCTCTGAGACAAACCCGTGGCAGGAAGCTGGTCCAAGGTGGTGGTGGGGCTGGCATGGGGTGGAGGTGG
- 319 GAGAGGGCTCATGGACACGTTTTCCTGAACGGTTGACGTCAGAAAAGCTGATTCTGGCTGTTGAAGAAGGTCAGGGTTGTGTTTCCTTTCCTTTCTTCTTGATAGATGAATGGAAAAGGC
- 198 CTGCTCTTTGATATTGGATTTTCCAAGGCCAAGAGATTCGATGCCATTCCAGGAATGTACTGTTTTGTGGCGGGACAGTGGCAAGGACCTGAGGGTAGGAGTCGAGGGAGGCTTCAGA
    1                                   *                                             *              Met Val Leu Ala Gly Leu Leu Leu Leu Leu
-  68 ACAGCTCAAGTAGCCAGATGGCTGAAGGTGTGGAGAGCTATAAGTGCCGGGGCCCGGGGCCTCTGCCTG GGTCTCCACCC  ATG GTC CTC GCA GGC CTG CTG CTG CTG CTC
   11 Thr Leu Leu Ala Gly Ala His Leu Leu Trp Gly Arg Trp Lys Leu Arg Asn Leu His Leu Pro Pro Leu Val Pro Gly Phe Leu His Leu
+  42 ACC CTG CTA GCT GGC GCT CAC CTG CTA TGG GGC CGG TGG AAG CTC AGA AAC CTC CAC CTC CCA CCT CTG GTC CCC GGC TTC CTG CAC CTG
   41 Leu Gln Pro Asn Leu Pro Ile His Leu Leu Ser Leu Thr Gln Lys Leu Gly Pro Val Tyr Arg Leu Arg Leu Gly Leu Gln G
+ 132 CTG CAG CCC AAC CTC CCC ATC CAT CTG CTG AGC CTG ACT CAG AAA CTC GGG CCT GTC TAC AGG CTT CGC CTT GGC CTG CAA G GTGAGGGTCTGC
   68                                                                                      lu Val Val Val Leu Asn Ser Lys Arg
+ 226 CTCCTTCCTGGTCCACAACAGTGAGGAGGCTTGGGGGTGTGCTTTCCTGCTGACTTCCTGCTTTTCCTACCACTCTCCCCGCCAG AG GTG GTG GTG CTG AAC TCC AAG AGG
   77 Thr Ile Glu Glu Ala Met Ile Arg Lys Trp Val Asp Phe Ala Gly Arg Pro Gln Ile Pro Ser T
+ 337 ACC ATT GAG GAG GCT ATG ATC AGG AAG TGG GTG GAC TTT GCC GGC AGA CCC CAG ATA CCA TCC T GTAAGGGGTTGCTTGAGGGCAGAAGCAGGGCTTTGG
   98                                                                      yr Lys Leu Val Ser Gln Arg Cys Gln Asp Ile Ser
+ 437 GAAGCAGGGGAGGTTCAGGCCTGTGGCTTCCTTGGTCAAGTCGACCCTCCCACCCCCAACCCCTACTCGCAG AC AAG CTG GTG TCT CAG CGC TGC CAG GAC ATC TCT
  110 Leu Gly Asp Tyr Ser Leu Leu Trp Lys Ala His Lys Lys Leu Thr Arg Ser Ala Leu Leu Leu Gly Thr Arg Ser Ser Met Glu Pro Trp
+ 544 CTG GGG GAC TAC TCT CTG CTC TGG AAG GCC CAC AAG AAA CTC ACC CGC TCA GCC TTG CTG CTG GGC ACC CGC AGC TCC ATG GAG CCC TGG
  140 Val Asp Gln Leu Thr Gln Glu Phe Cys Glu
+ 634 GTG GAC CAG CTG ACC CAG GAG TTC TGC GAG GTGAGCCTGGACTCCCACAGTTAGCCCAGGTCAGCTTCACCCTCCTGGAAGGCCCTCTATTTTTTTTCGCTTATATTACC
+ 735 CAAAGCATGTGGCATCTAGTTCCCTCACCAAGGATCGAACCTGGGCCCCCTGCATTAGGAGGCAGAGAATTCAGCCTCTCGACCACCAGGGACATACCAGGTCCCTCCCTATTTAGTTCAT
  150                 Arg Met Arg Val Gln Ala Gly Ala Pro Val Thr Ile Gln Lys Glu Phe Ser Leu Leu Thr Cys Ser Ile Ile Cys Tyr
+ 856 GCTCTTTCTCTTCAG CGC ATG AGA GTC CAG GCC GGT GCC CCC GTG ACC ATC CAG AAG GAA TTC TCT CTG CTT ACC TGC AGC ATC ATC TGT TAC
  176 Leu Thr Phe Gly Asn Lys
+ 949 CTC ACT TTT GGA AAC AAG GTCAAGGCTCTCTTGCCCTCCCATATGCCTGCCCCCCTCCCCCCACCCCCCGCAACCCCTCCCTGATCCTCTCCTGGTTCTGCACTGAAAGGATGTC
  182                 Glu Asp Thr Leu Val His Ala Phe His Asp Cys Val Gln Asp Leu Met Lys Thr Trp Asp His Trp Ser Ile Gln Ile
+1064 CTTCCTTTCTGCAG GAG GAC ACC TTA GTA CAT GCC TTT CAC GAC TGT GTT CAG GAC TTG ATG AAG ACC TGG GAC CAC TGG TCC ATC CAG ATT
  208 Leu Asp Met Val Pro Phe Leu Arg
+1157 TTG GAC ATG GTT CCT TTT CTC AGG GTGAGGAGCGCAGCCCCGACATGCCTGGTCCTGGGAGAAAGGGTGGGGATGGGGAACAGGCTTCCTTGCCAGCAGCTACACCTTCTTAC
  216             Phe Phe Pro Asn Pro Gly Leu Trp Arg Leu Lys Gln Ala Ile Glu Asn Arg Asp His Met Val Glu Lys Gln Leu Thr Arg
+1270 TCTCTGCCCCAG TTC TTC CCC AAC CCT GGG CTC TGG AGG CTG AAG CAG GCC ATA GAG AAC AGG GAC CAC ATG GTG GAG AAG CAG CTG ACG CGC
  243 His Lys
+1363 CAC AAG GTGGGAATTGTGCGTGGACCGGGGTCCTCCTCGGTTGGCAGCGGTGACGCGTACTGGCCAGAGCCAGGCAGCTTCTGGGCCATCCCTGCGGTCCTGGACCCATGGCACCCTCT
  245                                                       Glu Ser Met Val Ala Gly Gln Trp Arg Asp Met Thr Asp Tyr Met Leu Gln
+1482 AAACACTCTCCCCAAGCCAGGCACTCAGGCCAGCTGGACCTCTCCTCCCCAG GAG AGC ATG GTG GCC GGC CAG TGG AGG GAT ATG ACG GAC TAC ATG CTC CAG
  262 Gly Val Gly Arg Gln Arg Val Glu Glu Gly Pro Gly Gln Leu Leu Glu Gly His Val His Met Ser Val Val Asp Leu Phe Ile Gly Gly
+1585 GGC GTA GGC AGG CAA AGA GTA GAA GAG GGC CCG GGA CAG CTC CTG GAA GGA CAC GTG CAC ATG TCT GTG GTG GAC CTT TTC ATC GGC GGC
  292 Thr Glu Thr Thr Ala Ser Thr Leu Ser Trp Ala Val Ala Phe Leu Leu His His Pro Glu
+1675 ACT GAA ACC ACG GCG AGC ACC CTC TCC TGG GCT GTG GCG TTC CTA CTT CAC CAC CCC GAG GTATGCCCTGGAGCAGGCAAAGGATCCTTCCTGGCAGCCT
+1776 GGCCAGGGTGGCGGGAACCTAGCTCACTGGTGGCGATGCAGGCTGTGTTGTGGGGTGTAGGACTGGGTGGATCAGTGGGGTTTGCCTGGGTTCTTCTACTCGGCACATAGTTTCATGTGAG
+1897 TTCCTCCAGGGAATGGTGGAAGTTTGGGCTTCCCTAGTGGCTCAGATGGTAAAGCATCTGCCTGTAATGCAGGAGACCTAGGTTGGACCCCTGGCTGGGAAAGATGCCCTGAAGAAGAGAA
+2018 TGGCAACCCATTCTAGTACTCTTGCCTGGCGAATTCCATGGACAGAGGAGTCTGGAGGGTTACAGTCCATGAGTTTGCAAAGAGTGGACATGACTGAGTGACTGACTCACACACACACAC
  312                                                                                                           Ile Gln Arg Arg
+2149 ACACACACACACACACACACACACACATATACATAGAGACACACATGGTGGAGGTTGGCAGCTGGGGGCTGGCCAGTGACACGATTGGGGCTCCATTTCCCAG ATT CAG CGC CGC
  316 Leu Gln Glu Glu Leu Asp Arg Glu Leu Gly Pro Gly Ala Ser Cys Ser Arg Val Thr Tyr Lys Asp Arg Ala Arg Leu Pro Leu Leu Asn
+2254 CTT CAG GAG GAG TTG GAT CGC GAG CTG GGC CCC GGA GCC TCA TGC TCC CGA GTC ACT TAC AAG GAC CGC GCT CGG CTG CCT CTG CTC AAC
  346 Ala Thr Ile Ala Glu Val Leu Arg Leu Arg Pro Val Val Pro Leu Ala Leu Pro His Arg Thr Thr Arg Pro Ser Se
+2344 GCC ACC ATC GCG GAG GTG CTG CGC CTG CGG CCG GTC GTG CCC CTG GCC TTG CCT CAC CGC ACC ACG CGG CCT AGC AG GTGACTTCCGAGGCCTAGG
  371                                                                 r Ile Phe Gly Tyr Asp Ile Pro Glu Gly Met Val Val Ile Pro Asn
+2440 GATGAACGGCCAAGGCCCCAGCGGCTCGCGGGCCTCTAACTTAACCCTAACCCTCAG C ATC TTC GGC TAC GAC ATC CCT GAG GGC ATG GTC GTT ATC CCC AAC
  387 Leu Gln Gly Ala His Leu Asp Glu Thr  Val  Trp  Glu  Gln  Pro  His  Glu Phe  Arg Pro  A
+2543 CTC CAA GGT GCC CAC CTG GAC GAG ACC  GTC  TGG  GAG  CAG  CCG  CAC  GAG TTC  CGG CCG  GCTGTGTGGCGGCCGGGTGCCCCGCGATCGGTCC
  406                 sp Arg Phe Leu Glu Pro Gly Ala Asn Pro  Ser  Ala  Leu Ala  Phe Gly Cys Gly Ala Arg Val Cys Leu Gly Glu
+2653 TTCCGCCCGCCCCGCAG AC CGC TTC CTG GAG CCG GGC GCG AAC CCC  AGC  GCG  CTG GCA  TTT GGC TGC GGC GCG CGC GTG TGC CTG GGC GAG
  431 Ser Leu Ala Arg Leu Glu Leu Phe Val Val Leu Leu Arg Leu Leu Gln Ala Phe Thr Leu Leu Pro Pro Pro Val Gly Ala Leu Pro Ser
+2723 TCG CTC GCG CGC CTG GAG CTC TTC GTG GTG CTC CTC CGA CTG CTG CAG GCC TTC ACG CTG CTG CCC CCC CCC GTG GGC GCC CTG CCG TCG
  461 Leu Gln Pro Asp Pro Tyr Cys Gly Val Asn Leu Lys Val Gln Pro Phe Gln Val Arg Leu Gln Pro Arg Gly Val Glu Ala Gly Ala Trp
+2813 CTG CAG CCG GAT CCC TAC TGC GGT GTC AAC CTC AAG GTC CAG CCT TTC CAG GTG CGG CTG CAG CCC CGG GGC GTG GAG GCC GGG GCC TGG
  491 Glu Ser Ala Ser Ala Gln OP
+2903 GAG AGC GCC AGT GCC CAG TGA TGGGGGAGGAGTGGGTGGAGGACTCGGCCAGTCACGTGCCTCGGTGTCTCCTTTTATTGCTCCCGTACAAACCCTTGCCCTCCCCCCTGTAA
+3017 ACATGGTGCTGTGAGATCGCGGGTGGAGAAGGCTTCCCGCTGGTGGCTTGATGGTCAAGGTGGCCCCTGGCTCTTCTCTCGGCGCAACCCCTCAGTGCTCGGCAGTCATGGTGGGGCGCTGG
+3138 AGGGGTGAGCGGCGGCTCAGCCTCCCCTGCCCAGGGGCCGGTAGCTTCTTGGTCTCAGCTTCATTTCCGTGAAGGGCACTGAGAAATCGAAGCCCTTCCAGTAGTACCAGCTCACTCCCTG
+3259 GGAAGGGGGTTGTTGGGAGAGACTATTAGACGTTCAGCTCCCACCTGCCCTCCCCTCTGGAGATGGTTCCAGAAATCAACCAGCCATTCTCCCTGACTCCCCTTTTCCTTTTAAACCCTC
+3380 AACTGCAAAGGACTGAAGCAGAATCCTGAGCCGTCCCTTTACATCCAATAAATCATCTCCTCCTCAAG ATAACTTGGTGAGATTGACATTGTTCCATTTCACAGGCGAGAGCCCTGAAGCG
+3501 CTGAGCCATTCAGTAACTTATCACACTGTGTATCACAGACACCCAGGCAGGTGGAGCCCCAAGCACCCCCAGCT
```

FIG. 1. (Figure continues on opposite page.)

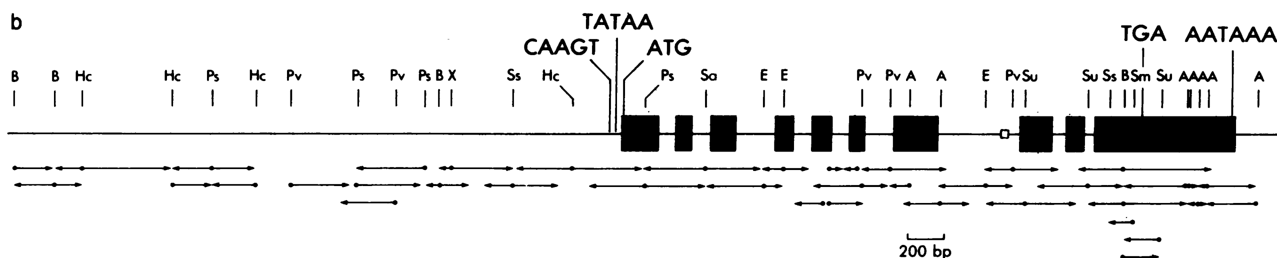

FIG. 1. (*a*) Sequence of the bovine P-450c21 gene. Base numbering (indicated with plus and minus signs) begins with the guanosine residue (asterisk) representing the principal RNA transcriptional initiation site. Two minor transcriptional initiation sites at −35 and −72 are also indicated by asterisks (see Fig. 2). Where two numbers appear in the left-hand margin, the upper number refers to the amino acid sequence. Spaces between introns and exons and between codons are for clarity only and do not represent absent bases. The TATAA and CAAGTA regions upstream from the principal transcriptional start site and the AATAAA sequence 22 bp from the poly(A) site are underlined. (*b*) Restriction map of the bovine P-450c21 gene and sequencing strategy. The solid boxes represent exons, and the small open box represents a region of repeated DNA (see text). Only the restriction sites used in sequencing are shown: A, *Alu* I; B, *Bam*HI; E, *Eco*RI; Hc, *Hin*cII; K, *Kpn* I; Ps, *Pst* I; Pv, *Pvu* II; Sa, *Sal* I; Sm, *Sma* I; Ss, *Sst* I; Su, *Sau*3A; X, *Xba* I. Arrows indicate the direction and extent of dideoxy sequencing. Arrows not originating at restriction sites represent sequencing initiated by the two unique oligonucleotides (4) used to identify λE11 or by the 27-mer used for primer extension. Each sequencing reaction was performed at least twice, generally on different clones of a given template. Rare areas of ambiguity were resolved by Maxam–Gilbert sequencing. bp, Base pairs.

P-450c21 (John Shively, personal communication) confirms our sequence.

Excluding the flanking DNA, the gene contains 3447 bp divided into 10 exons; other sequenced P-450 genes are all 6–23 kb long and contain 7 or 9 exons (9–15). Unlike other P-450 genes, exon 1 encodes a 5′ untranslated segment of mRNA only 11 bases long; 5′ untranslated regions of this size are unusual but have been reported as short as 7 bases (17). The nine introns, designated A through I, vary from 48 to 507 bp. Introns A and B contain perfect 12-bp inverted repeats at loci 253–264 and 487–498. Intron I is very short (48 bp), maintains an open reading frame, and lacks a consensus lariat junction sequence (18) and hence resembles exonic DNA. However, it has perfect splice sites, and no P-450 sequenced to date, including porcine P-450c21 (John Shively, personal communication), has amino acids corresponding to this region; furthermore, a bovine P-450c21 cDNA fragment lacks this region (Michael Waterman, personal communication): hence, the sequence designated intron I appears to be an intron. Intron G contains a 48-bp region of alternating pyrimidines and purines, including the sequence $(CA)_{20}$; this region may be able to form Z-DNA (19) and might thus be involved in the regulated expression of P-450c21. Hybridization of labeled total bovine genomic DNA to restriction fragments of the gene show that this intron is the only one containing highly repetitive DNA (ref. 4 and other data not shown). The rat P-450b and P-450e genes induced by phenobarbital have similar $(CA)_n$ sequences in their 5′ flanking DNA but not in introns (13, 15), whereas rat methylcholanthrene-inducible P-450c (9) and mouse dioxin-inducible P_1-450 (11) have similar structures in introns.

Analysis of the 5′ flanking region showed no glucocorticoid-regulatory elements, indicating it is unlikely that the P-450c21 gene is regulated by feedback of the glucocorticoid hormones to whose synthesis it contributes. Similarly, by comparison to the sequence of the murine cDNA (20), no evidence was found for the 3′ end of an adjacent C′4 gene encoding the fourth component of complement, even though the two human and mouse P-450c21 genes lie <4 kb 3′ to the C′4 genes (21, 22). Bases −444 to −415 are closely homologous (18 of 29 bases) to the putative cAMP regulatory site of several eukaryotic genes (23), but the significance of this is not clear, as cAMP regulation of P-450c21 is probable (24) but not established.

The gene sequence predicts synthesis of a 496 amino acid protein containing a leucine-rich hydrophobic amino-terminal region reminiscent of leader peptides of secreted proteins. As P-450c21 is a membrane-bound enzyme (25), this region is probably involved in affixing P-450c21 to the smooth endoplasmic reticulum. Amino acids 1–16 and 30–47 are hydrophobic regions lacking charged amino acids, each

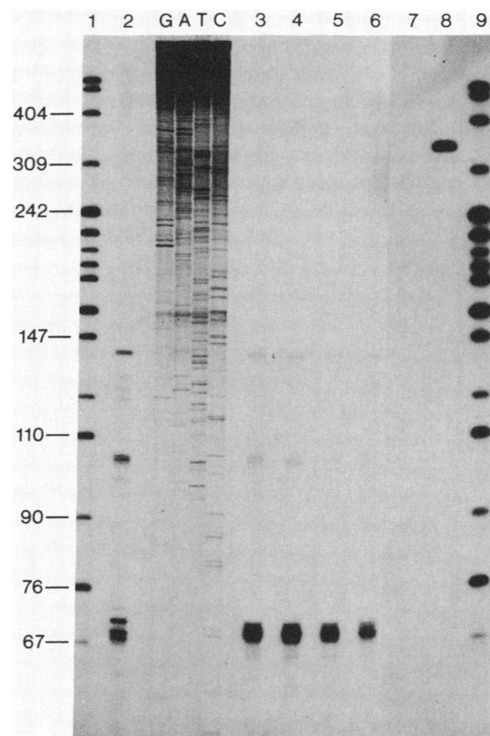

FIG. 2. Location of the transcriptional start site(s) by primer extension and S1 nuclease protection. A 5% acrylamide/7 M urea gel contains the following. Lanes 1 and 9, ^{32}P-labeled *Hpa* II-cut pBR322 size markers (indicated in bp). Lane 2, primer-extended cDNA initiated from the 27-mer on a bovine adrenal RNA template. Lanes G, A, T, and C, dideoxy sequencing ladder of the 5′ end of the gene initiated by the 27-mer on an M13 clone of the 412-base *Hin*cII/*Pst* I fragment. Lanes 3–8, S1 nuclease protection of the ^{32}P-labeled transcript of this same M13 clone hybridized to bovine RNA. Lanes 3–5, 50 μg of total bovine adrenal RNA and 500, 750, and 1500 units of S1 nuclease per ml, respectively. Lanes 6 and 7, 50 μg of total bovine testis and brain RNA, respectively, with 750 units of S1 nuclease. Lane 8, the 359-base probe [corresponding to bases 71 (3′ end of 27-mer) to −275, plus 13 bp of the M13 polylinker site] mixed with 50 μg of tRNA and no S1 nuclease. All samples were electrophoresed on a single gel; the photograph of lanes 7–9 is from a longer autoradiographic exposure of the gel.

```
bP450c21
rP450b
rP450c                                                                                         mpsvyg
bP450scc     mlarglplrsalvkacppilstvgegwghhrvgtgegagistktprpyseipspgdngwlnlyhfwrekgsqrihfrhienfqkygpiy

c21       MVLAGLLLLLTLLAGAHLLWG----RWKLRNLHLPPLVPGFLH-------LLQPNLPIHLLSLT----QKLGPVYRLRLGLQ*EVVVLNSKR
b         MepsiLLLLalLvgfllLLvrg---hpKsRgnf-PPgprplpllgn----LLQldrgglLnSfmg*lreKyGdVftvhLGpr-pVVmLcgtd
c      fpaftsAteLLLavttfclgfwvvrvtrtWvpkgLksPPgpwGlpfighvltlgknPhLsltkLS------QqyGdVlqiRiGst-pVVVLsgln
scc    reklgnlesvyiihpedvAHLf---kfegsyperydiPPwlayhry-------yqkPigvlfkkSgTw--------------kk-drVVLNtev

c21    TIEEAMIRKWVDFAGRPQIPS*--YKLVSQRCQDISLGDYSLLWKAHKKLTRSALLLGTRS---------SMEPWVDQLTQEFCE*RMRVQA-GA
b      TIkEAlvgqaeDFsGRgtIav---iepifkey*-gvifangerWKAlrrfslatmrdfgmg-------krSvEerVqeeaQclvE-elRksq*GA
c      TIkqAlvkqgdDFkGRPdlyS---ftLiangqsmtfnpDsgpLWaArrrLaqnALksfsiasdptlasscylEehVskeaeylis-kfqklm-ae
scc    mapEAiknfipllnpvsQdfvsllhKrikQqgsgkfvGDikedlfhfafesitnvmfGeRl--------gmlEetVnpeaQkF------------

c21    PVTIQKEFSL--LTCSIICYLTFGNK*EDTLVHAFHDCVQDLMKTW-DHWSIQILDMVPFLR*FFP-NPGLWRLKQAIENRDHMVEKQLTRHK*E
b      PldptflFqc--iTanIICtivFGer-fDytdrqFlrllelfyrTfsllsSfseq*vfeFfsgFlkyfPGahRqislnlqeildyighivekhra
c      vghfdpfkyLvvsvanvICaicFGrr-yDhddqellsiVmlsnefgevtgSgypaDfiPiLR-ylP-NssLdafKdlnkkfysfmkKlikeHy-r
scc    --------------------------iDavykmFHtsVplLnvppelyrlfrtktwrdhvaawd------------------------------t

c21    SMVA-GQWRDMTD-YMLQGVGRQRVEEGPGQLLEGHVHMSVVDLFIG---GTETTASTLSWAVAFLLHHPE--*IQRRLQEEL-DRELGPGASCS
b      tldp-napRDfiDtYlLrmek*-eksnhhtefhhenlmiSllsLFfa---GTETsstTLrygfllmLkyPhva*--ekvQkEi-DqviG---Shr
c      tfek*GhiRDiTDsliehcqdRrldEnanvQLsddkVitiVfDLFga*gfdTiTTAiswSlmylvtnpr-----IQRkiQEEL*DtviG---rdr
scc    ifnk-aekyteifvqdLrrKtefRnypGilycLlksekMlleDvkan---iTEmlAggvnttsmtLqwHlyemarslnvQEmLreevLnarrqae

c21    RVTYKDRARLPLLNATIAEVLRLRPVVPLALPHR-TTRPSS*IFGYDIPE-GMVVIPNLQGAHLDE-TVWEQPHEFRP*DRFLEPGAN------P
b      lpTldDRskmPytdAvIhEiqRfsdlVPigvPHR-vTkdtm-frGYllPk*nteVyPirssAlhyp-qyfdhPdsFnP-ehFLdangalkk--se
c      qprlsDRpqLPyLeAfIlEtfRhssfVPftiPHs*TiRdtS-lnGfyIPk-GhcVfvNqwqvnhDq*elWgdPnEFRP-eRFLtssgtldkhlse
scc    gdisKmlqmvPLLkAsIkEtLRLhPisvtlqr---ypesdlvlqdYlIPa-ktlVqvaiyamgrDp-affssPdkFdP-tRwLskdkdlihf---

c21    SALAFGC-GARVCLGESLARLELFVVLLRLLQAFTLLPPPVG---ALPSLQPDPYCGVNLKVQPF-QVRLQPRGVEAGAWESASAQ
b      afmpFst*GkRiCLGEgiARnELFlffttiLQnFsvsshlap---kdidLtPke-sGigkippty-QicfsaR
c      kvilFGl-GkRkCiGEtigRLEvFlfLaiLLQqmefnvsPge---kvdmtpaygltlkharcehF-QVqmrssGpqhlqa
scc    rnLgFGw-GvRqCvGrriAeLEmtlfLihiLenFkvemqhiGdvdtifnLiltPdkpifLvfrPFnQdppQa
```

FIG. 3. Relationship of P-450c21 to other P-450 sequences. Amino acid sequences of bovine P-450c21, rat P-450b (19), rat P-450c (9), and bovine P-450scc (39) are aligned to show regions of homology. The alignments and insertions of gaps were done manually as described (40). Asterisks in the sequences indicate the locations of intervening sequences; dashes indicate gaps.

sufficiently long to anchor the protein to the membrane (26). Amino acids 35–52 are 50% identical to the membrane-spanning region of the bovine (27) or human (28) low density lipoprotein receptor; amino acids 35–91 are 32% identical to amino acids 1324–1381 of human complement component C3a (29); the significance of these similarities is unknown. The overall sequence is 85% identical to that for porcine P-450c21, determined by amino acid sequencing (John Shively, personal communication). Of the 76 residues differing between the two species, 13 are at the carboxyl terminus, which is poorly conserved among other closely related forms of cytochrome P-450 (9–15).

Expression of P-450c21. Cattle (4), mice (22, 30), and human beings (21, 31) have two P-450c21 genes. The human and murine genes lie in the major histocompatibility loci in tight linkage to genes for the fourth component of complement. In man the array is C4A–21A–C4B–21B (21), whereas in mice the array is C4(Slp)–21A–C4–21B (30, 32), with all four genes having a left-to-right transcriptional orientation in both species. Deletion of the human P-450c21 B gene eliminates all 21-hydroxylase activity, suggesting the A gene is normally nonfunctional (31); by contrast, gene-specific oligonucleotide hybridizations indicate only the mouse A gene is functional (32). As different members of the pair of P-450c21 genes appear to be inactivated in mouse and man, such inactivation must have occurred after mammalian speciation about 85 million years ago. Therefore, both P-450c21 genes might remain functional in cattle; this might be consistent with finding two distinct sizes of P-450c21 mRNA in bovine adrenal cortex (4) but only a single size of P-450c21 mRNA in mouse (22) and human (unpublished) adrenal cortex.

To determine if our bovine P-450c21 gene was transcriptionally active and to locate the site(s) of transcriptional initiation (cap site), we performed primer extension and nuclease S1 protection experiments. The two experiments are entirely consistent in identifying cap sites (Fig. 2). Most P-450c21 mRNA molecules begin at one of three guanosine residues 10–12 bases 5′ to the ATG initiation codon; some of the mRNA begins at 46 or 83 bases 5′ to the ATG, and rare mRNA molecules may be initiated at other loci (Fig. 2 and other gels at higher resolution). The sequence TATAA is found 30 bases from the principal cap site, typical of most eukaryotic genes. The sequence GCTCAAGTA is found 65 bp from the cap site and closely resembles the consensus sequence 5′ GG$\frac{C}{T}$CAATCT 3′ found 65–90 bp 5′ to the cap site of many eukaryotic genes (33). The sequence GGGCAAGGA found 108 bp from the cap site resembles the consensus "CAAT box" more distantly. These "CAAT" sequences may function as weak promoters initiating transcription of the minor P-450c21 mRNA species originating at bases −35 and −72. Bovine adrenal RNA did not protect the *Sst* I/*Hinc*II fragment from bases −275 to −623 from S1 nuclease digestion, demonstrating that the CAAT and TATAA regions of −562 and −532 do not function as a promoter.

21-Hydroxylase activity is found in a wide variety of extraadrenal tissues (34), but the adult testis is not known to synthesize 21-hydroxylated steroids (35). The nature of extra-adrenal 21-hydroxylase is controversial; its activity appears to remain when adrenal 21-hydroxylation is impaired by congenital adrenal hyperplasia (36) and, unlike adrenal 21-hydroxylase, it appears to be regulated by estrogen (34). Furthermore, rabbit hepatic P-450 isozyme 1 readily 21-hydroxylates progesterone (37) but is structurally no more

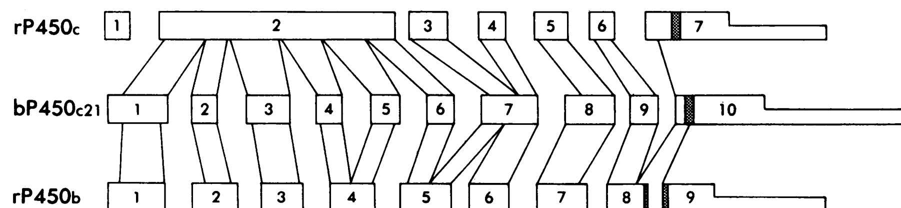

FIG. 4. Exon structures of selected P-450 genes. Exons of rat phenobarbital-inducible P-450b (15), rat 3-methylcholanthrene-inducible P-450c (9), and bovine P-450c21 are drawn to scale, but the spaces between exons do not correlate with the lengths of the corresponding introns. Lines connecting exons of different P-450 genes show homologous regions of exons. The stippled area represents the heme-binding site. The narrow regions of the 3′ terminal exons indicate the portion of that exon encoding the 3′ untranslated regions of the corresponding mRNAs.

related to P-450c21 than are other hepatic P-450s (38). To determine if the P-450c21 gene is expressed in nonadrenal tissues, we performed S1 nuclease protection experiments with several bovine RNAs. The result with testicular RNA (Fig. 2, lane 6) is indistinguishable from that obtained with bovine adrenal RNA, indicating that this P-450c21 gene is expressed in adrenal cortex and testis. By contrast, bovine brain RNA (Fig. 2, lane 7) and liver, kidney, and ovary RNA (not shown) did not protect the labeled P-450c21 gene fragment from S1 nuclease digestion.

Evolution of the P-450c21 Gene. The P-450c21 gene has 10 exons spanning 3.5 kb, distinguishing it from rabbit and rat P-450 genes induced by phenobarbital, which have 9 exons spanning 14–23 kb (13–15), and the rat and mouse P-450 genes induced by 3-methylcholanthrene or dioxin, which have 7 exons spanning 6–7 kb (9–12). To examine potential relationships among various P-450 sequences, we aligned (Fig. 3) the amino acid sequences of bovine P-450c21 with rat hepatic P-450b (induced by phenobarbital) (15), rat hepatic P-450c (induced by 3-methylcholanthrene) (9), and bovine adrenal P-450scc (39), a mitochondrial P-450 also engaged in steroid hormone synthesis (16). These amino acid alignments in turn indicate homologies among the various exons of the three genes (Fig. 4). Inspection of these exon domains suggests that a precursor to these genes was divided into at least 12 exons. Based upon similar analysis of dioxin-inducible mouse P-450 genes and the rat P-450b and P-450e genes, Gonzales *et al.* (11) concluded that an ancestral P-450 gene had a minimum of 14 exons.

Of the 496 amino acids in P-450c21, 129 are identical in P-450b and 139 are identical in P-450c, further emphasizing that the families of P-450 genes induced by 3-methylcholanthrene and phenobarbital are about equally distant in evolution from the P-450c21 gene. The sequence of P-450scc was much less related, having only 88 identical amino acids, including none in its first 115 residues. As P-450scc is a mitochondrial enzyme, this great difference may reflect on ancient evolutionary divergence between nuclear genes encoding mitochondrial and cytoplasmic P-450s as well as the presence of a 39 amino acid mitochondrial leader sequence in P-450scc. Thus, gene structure and amino acid sequence indicate that P-450c21 represents a separate family of P-450 genes, only distantly related to other members of the P-450 gene superfamily.

We thank John Shively for communicating his complete porcine P-450c21 amino acid sequence prior to submission for publication, Michael Waterman for his partial cDNA sequence, and Carol Dahlstrom for typing. This work was supported by March of Dimes Grant 6-396 and National Institutes of Health Grant HD16047.

1. Cooper, D. Y., Levin, S., Narasimhulu, S., Rosenthal, O. & Estabrook, R. W. (1965) *Science* **147**, 400–402.
2. Nebert, D. W. & Gonzalez, F. J. (1985) *Trends Pharmacol. Sci.* **6**, 160–165.
3. Rosenfield, R. L. & Miller, W. L. (1983) in *Hirsutism and Virilism*, eds. Mahesh, V. B. & Greenblatt, R. B. (Wright-PSG, Boston), pp. 87–119.
4. Chung, B., Matteson, K. J. & Miller, W. L. (1985) *DNA* **4**, 211–219.
5. Biggin, M. D., Gibson, P. H. & Hong, G. F. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 3963–3967.
6. Sudhof, T. C., Goldstein, J. L., Brown, M. S. & Russell, D. W. (1985) *Science* **228**, 815–822.
7. White, P. C., New, M. I. & Dupont, B. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 1986–1990.
8. Breathnach, R. & Chambon, P. (1981) *Annu. Rev. Biochem.* **50**, 349–383.
9. Sogawa, K., Gotoh, O., Kawajiri, K. & Fujii-Kuriyama, Y. (1984) *Proc. Natl. Acad. Sci. USA* **81**. 5066–5070.
10. Sogawa, K., Gotoh, O., Kawajiri, K., Harada, T. & Fujii-Kuriyama, Y. (1985) *J. Biol. Chem.* **260**, 5026–5032.
11. Gonzalez, F. J., Kimura, S. & Nebert, D. W. (1985) *J. Biol. Chem.* **260**, 5040–5049.
12. Kimura, S. & Nebert, D. W. (1985) *DNA* **4**, 368–375.
13. Mizukami, Y., Sogawa, K., Suwa, Y., Muramatsu, M. & Fujii-Kuriyama, Y. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 3958–3962.
14. Atchison, M. & Adesnick, M. (1983) *J. Biol. Chem.* **258**, 11285–11295.
15. Suwa, Y., Mizukami, Y., Sogawa, K. & Fujii-Kuriyama, Y. (1985) *J. Biol. Chem.* **260**, 7980–7984.
16. Matteson, K. J., Chung, B., Urdea, M. S. & Miller, W. L. (1986) *Endocrinology* **118**, 1296–1305.
17. Jameson, L., Chin, W. W., Hollenberg, A. N., Chang, A. S. & Habener, J. F. (1984) *J. Biol. Chem.* **259**, 15474–15480.
18. Keller, W. (1984) *Cell* **39**, 423–425.
19. Wang, A. H.-J., Quigley, G. J., Kolpak, F. J., Crawford, J. L., van Boom, J. M., van der Marel, G. & Rich, A. (1979) *Nature (London)* **282**, 680–686.
20. Sepich, D. S., Noonan, D. J. & Ogata, R. T. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 5895–5899.
21. Carroll, M. C., Campbell, R. D. & Porter, R. R. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 521–525.
22. Amor, M., Tosi, M., Duponchel, C., Steinmetz, M. & Meo, T. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 4453–4457.
23. Nagamine, Y. & Reich, E. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 4606–4610.
24. Boggaram, V., Simpson, E. R. & Waterman, M. R. (1984) *Arch. Biochem. Biophys.* **231**, 271–279.
25. Kominami, S., Ochi, H., Kobayashi, Y. & Takemori, S. (1980) *J. Biol. Chem.* **255**, 3386–3394.
26. Adams, G. A. & Rose, J. K. (1985) *Cell* **41**, 1007–1015.
27. Russell, D. W., Yamamoto, T., Schneider, W. J., Slaughter, C. J., Brown, M. S. & Goldstein, J. L. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 7501–7505.
28. Yamamoto, T., Davis, C. G., Brown, M. S., Schneider, W. J., Casey, M. L., Goldstein, J. L. & Russell, D. W. (1984) *Cell* **39**, 27–38.
29. de Bruijn, M. H. L. & Fey, G. F. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 708–712.
30. White, P. C., Chaplin, D. D., Weis, J. H., Dupont, B., New, M. I. & Seidman, J. G. (1984) *Nature (London)* **312**, 465–467.
31. White, P. C., New, M. I. & Dupont, B. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 7505–7509.
32. Parker, K. L., Chaplin, D. D., Wong, M., Seidman, J. G., Smith, J. A. & Schimmer, B. P. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 7860–7869.
33. Benoist, C., O'Hare, K., Breathnach, R. & Chambon, P. (1980) *Nucleic Acids Res.* **8**, 127–142.
34. Casey, M. L. & MacDonald, P. C. (1982) *Endocr. Rev.* **3**, 396–403.
35. Finkelstein, M. & Schaefer, J. M. (1979) *Physiol. Rev.* **59**, 353–406.
36. Winkel, C. A., Casey, M. L., Worley, R. L., Madden, J. D. & MacDonald, P. C. (1983) *J. Clin. Endocrinol. Metab.* **56**, 104–107.
37. Dieter, H. H., Muller-Eberhard, U. & Johnson, E. F. (1982) *Science* **217**, 741–743.
38. Tukey, R. H., Okino, S., Barnes, H., Griffin, K. J. & Johnson, E. F. (1985) *J. Biol. Chem.* **260**, 13347–13354.
39. Morohashi, K., Fujii-Kuriyama, Y., Okada, Y., Sogawa, K., Hirose, T., Inayama, S. & Omura, T. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 4647–4651.
40. Miller, W. L., Coit, D., Baxter, J. D. & Martial, J. A. (1981) *DNA* **1**, 37–50.