Comparison of the Flanking Regions and Introns of the Mouse 2,3,7,8-Tetrachlorodibenzop-dioxin-inducible Cytochrome P₁-450 and P₃-450 Genes

(Received for publication, June 22, 1984)

Frank J. Gonzalez, Shiko Kimura, and Daniel W. Nebert‡

From the Laboratory of Developmental Pharmacology, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland 20205

The C57BL/6N inbred mouse cytochrome P₁-450 and P₃-450 genes, two genes in the same family and under control by the Ah receptor, have been completely sequenced. The transcription initiation sites were confirmed by primer extension studies. An additional 823 and 883 bp of the 5' upstream flanking regions of P₁-450 and P₃-450, respectively, and 1771 and 1251 bp of the 3' downstream flanking regions of P₁-450 and P₃-450, respectively, were sequenced and studied. P₁-450 exons total 2619 nucleotides, and the gene spans 6215 bp. P₃-450 exons total 1892 nucleotides, and the gene spans 6716 bp. Three interesting highly homologous regions of 11 or 12 bp, upstream between -280 and -530 from the cap site of both genes, are noted as possible candidates for binding by the inducer-Ah receptor complex (and/or other DNA-binding regulatory proteins). Several stretches of DNA upstream from the cap site, in several introns, and in the 3' flanking region of both genes have a high degree of homology with known core enhancer sequences. Other interesting stretches (DNA with Z-DNA-forming properties, DNA with recombinational potential, highly repetitive and middle repetitive sequences between 50 and 360 bp in length, and "simple" sequences presumably having no function in gene expression) exist throughout many of the introns and flanking regions in both the positive and negative strands of both genes.

The mouse 2,3,7,8-tetrachlorodibenzop-dioxin-inducible and rat phenobarbital-inducible P₁-450 genes were compared for the amino acid residue number at each exon-intron junction, the location in the coding triplet at which the exons are split, and homologies among introns and exons. It can be shown that these two genes probably diverged from a common ancestor more than 200 million years ago and that P₁-450 and P₃-450 split from each other about 65 million years ago.

The murine Ah locus governs the induction of several drug-metabolizing enzymes by polycyclic aromatic compounds such as 3-methylcholanthrene and TCDD† (reviewed in Refs. 6 and 7). These enzymes include microsomal cytochromes P₁-450 and P₃-450 (1), microsomal UDP-glucuronosyltransferase with 4-methylumbelliferosene as substrate (8), and the cytosolic NAD(P)H:menadione oxireductase (9). Each of these gene products appears to be controlled by the Ah receptor (7).

Full-length cDNA clones for both P₁-450 and P₃-450 have recently been isolated (10) and sequenced (11, 12). The complete cDNA nucleotide and deduced amino acid sequences exhibit 68% and 73% similarity, respectively. It was estimated that these two homologous genes of the same P-450 subfamily diverged from each other about 65 million years ago (12). The genomic clones for P₁-450 and P₃-450 have similar exon-intron patterns, the 2nd and 7th exons being much larger than the other 5 (10). The P₁-450 and P₃-450 genes have been localized to mouse chromosome 9 (13). The transcriptional activation of both genes by 3-methylcholanthrene and TCDD (14) has been rigorously correlated with the presence of the Ah-receptor complex gaining high affinity for nuclear chromatin material (14, 15). The P₁-450 induction process occurs developmentally several weeks earlier than the P₃-450 induction process (16, 17); both the P₁-450 and P₃-450 induction processes occur in liver, and P₁-450 but not P₃-450 induction occurs in C57BL/6N kidney (17). All of these lines of evidence taken together strongly suggest that these two genes will be found to lie in tandem and that the inducer-receptor complex controls the expression of both genes in an unknown manner with developmental and tissue specificity.

In this report we have examined the 5' and 3' flanking regions and all introns of both strands of both the P₁-450 and P₃-450 genes. We had anticipated, and in fact do find, numerous segments having Z-DNA-forming potential, regions of DNA with homology to previously published enhancer core sequences, and stretches of DNA homologous to highly and middle repetitive sequences. One strength of this study is that we are comparing these two homologous genes in the same P-450 subfamily that had been cloned from the same genomic DNA library from liver of the inbred C57BL/6N mouse strain. Any interesting regions that occur approximately the same.

† The abbreviations used are: TCDD, 2,3,7,8-tetrachlorodibenzop-dioxin; bp, base pairs; Pu, purine; Py, pyrimidine.

‡ To whom correspondence should be addressed at: Building 10, Room 6C-101, National Institutes of Health, Bethesda, MD 20205.

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

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

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.
distance upstream from the cap site, downstream in the 3' flanking region, or in the same intron of both P1-450 and P3-450 will have a much higher probability of being functionally important (e.g., associated with transcriptional activation by TCDD) than if the interesting region exists in one gene but not the other. On the other hand, one or more interesting regions that differ between the two genes may reflect the distinct dissimilarities in developmental and tissue specificity mentioned above.

EXPERIMENTAL PROCEDURES

Isolation and Subcloning of the P1-450 and P3-450 Genes—A C57BL/6N liver genomic library was constructed with the lambdoid phage vector Charon 30 (18) as outlined in detail in Ref. 10. The regions that differ between the two genes may reflect the developmental and tissue specificity not the other. On the other hand, one or more interesting regions that differ between the two genes may reflect the distinct dissimilarities in developmental and tissue specificity mentioned above.

DNA Sequencing—All DNA sequence determinations were carried out by use of the M13 cloning and dideoxyribonucleotide sequencing strategies (19). Libraries of shotgun clones were produced and sequenced from each genomic subclone (20). Briefly described, DNA fragments (5 μg) were circularized by ligation and sonicated with four bursts (10 s each) at 100 W in 0.5 ml of 10 mM Tris-HCl buffer, pH 7.5. Fragments were concentrated by ethanol precipitation and re-paired with the Escherichia coli DNA polymerase large fragment. DNA fragments ranging from 500 to 1000 bp were isolated by agarose gel electrophoresis and ligated into the SmaI site of M13 mp11. Sequencing was carried out with the dideoxyribonucleotide reagent kits (P-L Biochemicals, Milwaukee, WI), [α-32P]dATP (400 Ci/mmol; Amersham Radiochemicals, Chicago, IL), and the DNA polymerase large fragment (Bethesda Research Laboratories, Bethesda, MD). Sequences were displayed on a salt gradient (21) and standard 50% urea-6% acrylamide gels (22). Sequence alignments were made with the Staden consensus program (23); any gaps that remained after the sequencing of many shotgun clones were filled in by use of the Staden consensus program (23); any gaps that remained after the sequencing of many shotgun clones were filled in by sequenced from restriction site-directed clones. Partial restriction maps, and directions of the primary fragments used to determine the complete nucleotide sequence of each gene, are displayed in Fig. 1. The BamHI and HindIII sites in the P1-450 and P3-450 genes were used to construct the subclones (with the exception of one HindIII site in exon 2 of P3-450). The possibility of a small stretch of DNA missed at these sites was ruled out, however, based on high resolution restriction mapping with HindIII sites in exon 2 of P3-450.

EXon-Intron Junctions—Sequences of the two genes and flanking regions are shown in Figs. 2 and 3. Analysis of both genes across the exon-intron junctions (Table I) reveals that all splice sites rigorously follow the Chambon rule (25). Consensus sequence of the acceptor splice sites is PyAG[N]. Consensus sequence of the donor splice sites is PuGTPu.

Primer Extension Analysis—We had concluded, from sequencing of the presumably full-length cDNA clones, that P1-450 mRNA was 2620 nucleotides (12) and P3-450 mRNA was 1894 nucleotides (11) in length. Sequencing of the genomic clones, plus primer extension studies (Fig. 4), confirmed that our cDNA clones (10) contained virtually every base in the cDNA. The 5'-most G and the 3'-most A in both P1-450 (12) and P3-450 (11) cDNA represent nucleotides from the poly(dG) and poly(dA) tracts, respectively, of the cloning vector (26). One additional base in P1-450 exon 7 (between position 6047 and 6048 in the gene sequence of Fig. 2) was originally reported (12) and is an error. This now stands corrected in Fig. 2. Hence, we are now certain that the total lengths of the P1-450 and P3-450 mRNAs are 2619 and 1892 bp, respectively.

Primer extension analysis (Fig. 4) localized the sites of initiation of transcription. The distance from the Sau3A cleavage site of P1-450 to the putative initiation site covers 100 nucleotides, with a second initiation site 2 nucleotides downstream. The distance from the Rsal cleavage site to the putative initiation site of P3-450 covers 75 nucleotides, with...
Mouse P1-450 and P3-450 Genes

The TATA boxes can be seen in both the cap site and the initiation site. The TATA boxes are numbered starting with +1. The cap site denotes possible cap sites. In conclusion, primer extension studies confirm that the cap site is defined as the putative site of transcription initiation; negative numbers represent the upstream 5' flanking regions. Nucleotides upstream from the transcription initiation site have the canonical sequence 5'-GGTCAATCT-3'.

Both genes thus appear to have degenerate CCAAT boxes.

In each case two potential cap sites were detected in both exons and introns of some genes. Small stretches of homology to another mRNA in this primer extension assay. Both genes have degenerate CCAAT boxes.

Fig. 2. Nucleotide sequence of the mouse P1-450 gene and flanking regions. Nucleotides are numbered relative to position 1, defined as the putative site of transcription initiation. Negative numbers represent the upstream 5' flanking regions. Nucleotides in the introns and flanking regions are illustrated in groups of 10, whereas nucleotides upstream are shown without spacing. The coding region starts with ATG, and the residues are shown in the single-letter amino acid code. The TATA box, initiation codon, codon for the cysteine believed to function in the enzyme active-site (12), and termination codon and the putative poly(A) addition signal appear in black boxes. Following the 9th (final) exon, the 3' flanking region is numbered starting with +1. The three black boxes upstream from the cap site denote possible 5' flanking regions with striking homology between the P1-450 and P3-450 genes (see text). The grey boxes in the flanking regions and introns denote regions having more than 75% homology with proven core enhancer elements.

Fig. 3. Homology of flanking regions between mouse P1-450 and P3-450 genes (Figs. 2 and 3) sequenced immediately upstream from the transcription initiation site. The TATA boxes can be seen in both P1-450 and P3-450 genes. The CCAAT box (28) is less well-conserved, is typically 65 to 90 bp upstream from the cap site, and has the canonical sequence 5'-GO_C TCAATCTT-3' (29). P1-450 has the sequence GGTCCACCACT at about -80, and P3-450 has the sequence GATATCAAGT at about -70. Both genes thus appear to have degenerate CCAAT boxes.

In the 5' flanking regions of some genes, small stretches of
DNA are under consideration for being the binding site for inducer-receptor complexes or DNA-binding proteins. Examples include the glucocorticoid (30) and estrogen (31) receptor complexes and the TGGCA-binding protein (32). We therefore looked for such areas of homology in the 5' flanking region that might explain why these two genes are under control of the same Ah receptor. Three stretches of DNA (shown in Figs. 2 and 3 in black boxes) are highly suspect: (i) 11 out of 12 bases match in the box around -380 in P$_1$-450 and around -280 in P$_3$-450; (ii) 10 of 11 bases match in the box around -460 in P$_1$-450 and around -415 in P$_3$-450; and (iii) 11 of 12 bases match in the box around -530 in P$_1$-450 and around -455 in P$_3$-450. The difference between P$_1$-450 and P$_3$-450 with respect to the distance of each of the three boxes from the cap site should be noted: about 100, 45, and 45 bases, respectively. These three interesting homologous regions, although somewhat short, thus exhibit striking similarity and each of the three are located in approximately the same respective distances from the cap site. Possible regulatory functions for these three regions will be studied.

The fibrinogen genes were recently reported to have conserved stretches in their 5' flanking regions (33). It is of interest that P$_1$-450 has a region (-514 to -549) with 28 of 36 bases, 78% similar, and that P$_3$-450 has a region (-449 to
forming Potential, Recombinational Properties, and Repetitive DNA having greater than rine sarcoma virus 72-bp repeat enhancer in the and introns of both genes where conserved stretches of DNA are homologous to core enhancer elements, enhancer elements.

Sequences—Figs. (38). P3-450 has a (TG) tract around -650 (Fig. 3), but P1-450 has nothing like this in its 5' flanking region (Fig. 2). In the first intron of both genes, there are interesting similarities. P1-450 has (TG5),TA(TG5),C(TG5) centered around 1210; P3-450 has (TG5),TA(TG5),TGTAGA(TG5) centered around 240. These regions not only have Z-DNA-forming properties, but they also display strong homology to the immunoglobulin recombinational signals.

DNA Having Possible Core Enhancer Elements, Z-DNA-forming Potential, Recombinational Properties, and Repetitive Sequences—Figs. 2 and 3 also depict in the flanking regions and introns of both genes (grey boxes) several stretches of DNA having greater than 75% similarity to the Harvey murine sarcoma virus 72-bp repeat enhancer in the U3 region of the long terminal repeat (34) and several other putative core enhancer elements (35-37). Although each of these illustrated stretches of DNA are homologous to core enhancer elements, it goes without saying that biological function must be demonstrated before any of these regions can be proved to have a role in expression among any of these genes.

Although each of these illustrated stretches of DNA are homologous to core enhancer elements, it goes without saying that biological function must be demonstrated before any of these regions can be proved to have a role in the expression of either gene.

DNA having Z-DNA-forming potential consists of (P1Py)n where n ≥ 4 (38). P2-450 has a (GT)n repeat around 250, but P1-450 has nothing like this in its 5' flanking region (Fig. 2). In the first intron of both genes, there are interesting similarities. P1-450 has (TG)n,TA(TG)n,C(TG)n centered around 2160; P3-450 has (TG)n,TA(TG)n,TGTGA(TG)n,TA centered around 240. These regions not only have Z-DNA-forming properties, but they also display strong homology to the immunoglobulin recombinational signals (39, 40).
portions of introns were found when the two genes were compared. Hence, we find no evidence for gene conversion and/or unequal crossing-over between these two genes. Most likely, these two genes thus arose from duplication.

"Simple sequences" are stretches of DNA that consist of one or several tandemly repeated sequences, e.g., (AT)_m (AG)_n, (GT)(CA)_m, (CTT)_n, and (GAG)_n (41). These simple sequences are repetitive, interspersed throughout many eukaryotic genomes, and are believed to arise by slippage replication and unequal crossing-over and to have no general function with regard to gene expression (41). Many of these sequences can be seen in the introns and flanking regions of the P1-450 and P3-450 genes (Figs. 2 and 3).

Comparison of Exon and Intron Conservation—The homologous two genes in this P-450 family have similar exon-intron patterns, with the 2nd and 7th exons much larger than the other 5. The 727-bp difference in size between P1-450 and P3-450 cDNAs can be mostly accounted for by the larger 3' seventh exon of P1-450 (Table II). Variability among the introns is noteworthy. P1-450 intron 1 is more than twice as large as P3-450 intron 1, P3-450 intron 3 is more than seven times longer than P1-450 intron 3, and P3-450 intron 6 is more than 10 times longer than P1-450 intron 6. The degrees of nucleotide similarity among the introns range between 20 and 42%; these values were consistent with the expected amount of nucleotide divergence among introns during about 65 million years.

Although the first 6 exons are similar in size, it is interesting that exons 4, 5, and 6 are exactly the same size in the two genes: 90, 124, and 87 bp, respectively. Besides exon 2 (81% similarity), exons 4, 5, and 6 are quite conserved (73%, 85%, and 78% similarity, respectively). Exons 2, 4, 5, and 6 and the 5' end of exon 7 thus are good candidates for encoding highly conserved protein domains—perhaps those of importance to catalytic activities of these two membrane-bound multicomponent enzymes. Other possible reasons for conservation of these protein subunits include (i) type of attachment and configuration in the endoplasmic reticulum, (ii) participation in the heme-binding enzyme active site, and (iii) area of attachment of the NADPH-P-450 oxidoreductase. The so-called "conserved tridecapeptide" (42) is encoded by exon 5 in both P1-450 and P3-450, and the conserved cysteinyl fragments in the NH2-terminal and COOH-terminal portions of the proteins (11, 12) are encoded by exons 2 and 7, respectively, in both P1-450 and P3-450. In the phenobarbital-inducible P450 genes, the conserved tridecapeptide is encoded by exon 7 and the NH2-terminal and COOH-terminal conserved cysteinyl fragments are encoded by exons 3 and 9, respectively (43).

Comparison of P1-450 and P3-450 with Rat P-450—It should be noted that the locations in the coding triplet at which the exons are divided are identical between P1-450 and P3-450 (Table III). The residue number at the end of each exon varies by no more than 3 amino acids between P1-450 and P3-450. Differences between the TCDD-inducible and the phenobarbital-inducible P450 gene families are striking: (i) the former has 7 exons and the latter has 9 (43, 44); (ii) the former genes span less than 7 kb while the latter genes span...
**TABLE IV**

List of potentially interesting homologous regions between mouse P₁-450 or P₃-450 and sequences published in the genetic sequence data bank

This computer program SRCHN (47) compares a given nucleic acid sequence with all sequences in the Genbank Data Bank as of June 1, 1984, using the algorithm of Wilbur and Lipman. It must be recognized in gene data bank searches that repetitive tracts such as (AG)ₙ, (CT)ₙ, (CCT)ₙ, (GC)ₙ, (CAA)ₙ, and (GAA)ₙ, may yield high similarity scores of unknown biological importance and are omitted from further consideration in this table. Besides those listed, other P₁-450 and P₃-450 exons exhibited significant similarity with rat P-450e exons, but more homology was seen at the protein level. At the nucleotide level, substantial divergence has occurred and therefore the significance of match-up does not show up in this search program.

| Gene   | Portion of gene | Nucleotide position according to Figs. 2 and 3 | No. of bases matched | Similarity | Name of matching sequence | Significance of match (standard deviations above the mean) |
|--------|-----------------|------------------------------------------------|---------------------|------------|---------------------------|----------------------------------------------------------|
| P₁-450 | 5' Flanking region | -131 to -106 | 23/26 | 88% | Human oncogene c-fos, complete gene and 5' flanking region | 5.3 |
|        |                 | -190 to -189 | 18/22 | 82% | Adenovirus 5 early E1 transforming region | 4.6 |
|        |                 | -132 to -177 | 36/56 | 64% | Human oncogene c-myc, 5' end | 4.5 |
|        |                 | -192 to -176 | 16/18 | 89% | Mouse mammary tumor virus, long terminal repeat | 3.9 |
|        | Intron 1 | 851 to 986 | 115/136 | 85% | Mouse α-fetoprotein, 5' end and flanking region | 47 |
|        |                 | 881 to 953 | 62/72 | 86% | Mouse H-2 Tla complex, pseudogene | 27 |
|        |                 | 886 to 965 | 65/80 | 81% | Mouse H-2 Class II Ia antigen Aα gene | 21 |
|        |                 | 856 to 905 | 42/50 | 84% | Mouse H-2Lα gene | 13 |
|        |                 | 837 to 978 | 89/142 | 63% | Human oncogene c-myc | 11 |
|        |                 | 844 to 957 | 67/114 | 59% | Human prothrombin gene | 11 |
|        |                 | 851 to 986 | 84/136 | 62% | Human embryonic c-globin gene | 9.7 |
|        | Intron 1 (reverse strand) | 982 to 857 | 104/126 | 83% | Mouse H-2 Tla complex, pseudogene | 43 |
|        |                 | 983 to 846 | 115/138 | 83% | Mouse B1 ubiquitous repeat | 42 |
|        |                 | 981 to 840 | 111/142 | 78% | Rat repetitive sequence cluster, Alu-like (B1) | 28 |
|        |                 | 984 to 870 | 93/115 | 81% | Chinese hamster Alu-equivalent gene | 28 |
|        |                 | 1029 to 842 | 118/188 | 63% | Human Alu family | 23 |
|        |                 | 965 to 915 | 42/51 | 82% | Mouse and hamster 4.5 S RNA associated with poly(A)-containing RNAs | 22 |
|        |                 | 981 to 786 | 96/196 | 49% | Mouse U6 small nuclear RNA pseudogene | 17 |
|        |                 | 992 to 850 | 90/143 | 63% | Ape (chimpanzee) Alu-type DNA | 16 |
|        |                 | 982 to 888 | 59/95 | 62% | Human 7SL RNA | 13 |
|        | Exon 2 | 2734 to 2781 | 35/48 | 73% | Rat phenobarbital-inducible P-450, Exon 2 | 8.0 |
|        |                 | 3072 to 3124 | 35/53 | 66% | Rat phenobarbital-inducible P-450, Exon 4 | 7.3 |
|        | Intron 2 | 3631 to 3805 | 132/175 | 75% | Mouse PR1 repetitive sequence and Type 2 Alu repeat | 56 |
|        |                 | 3647 to 3805 | 122/159 | 77% | Mouse H-2 Tla complex, pseudogene | 54 |
|        |                 | 3723 to 3804 | 31/82 | 62% | Mouse immunoglobulin κ light-chain, J and C regions | 8.9 |
|        | Intron 2 (reverse strand) | 3824 to 3624 | 163/201 | 81% | Mouse B2 repeat sequence | 62 |
|        |                 | 3811 to 3624 | 154/188 | 82% | Mouse B1 repetitive elements | 61 |
|        |                 | 3826 to 3853 | 139/173 | 80% | Rat growth hormone gene | 56 |
|        |                 | 3811 to 3637 | 139/175 | 78% | Mouse Type 2 Alu sequence | 50 |
|        |                 | 3790 to 3655 | 113/136 | 83% | Mouse H-2α gene mRNA | 48 |
|        |                 | 3811 to 3658 | 131/154 | 85% | Mouse U6 small nuclear RNA pseudogene | 46 |
|        | 3' Flanking region | +527 to +557 | 24/31 | 77% | Chicken myosin heavy chain, translational control RNA | 11 |
|        |                 | +1740 to +1768 | 26/29 | 90% | Mouse U1 small nuclear RNA gene | 7.5 |
| P₃-450 | 5' Flanking | -734 to -616 | 75/115 | 65% | Mouse immunoglobulin heavy-chain, V region | 21 |
|        |                 | -530 to -443 | 73/88 | 83% | Human Alu Type I in δ-globin 5' flanking region | 20 |
|        | Exon 2 | 1370 to 1417 | 35/48 | 73% | Rat phenobarbital-inducible P-450, Exon 2 | 8.5 |
TABLE IV—continued.

| Gene            | Portion of gene | Nucleotide position according to Figs 2 and 3 | No. of bases matched | Similarity | Name of matching sequence                               | Significance of match (standard deviations above the mean) |
|-----------------|-----------------|-----------------------------------------------|----------------------|------------|----------------------------------------------------------|----------------------------------------------------------|
| Intron 1 (reverse strand) | 244 to 207     | 34/38                                         | 89                   | Rat repetitive sequence cluster                     | 15                                                      |
|                 | 430 to 390      | 45/50                                         | 90                   | Mouse B2 repeat sequence                            | 12                                                      |
|                 | 446 to 390      | 42/57                                         | 74                   | Mouse B1 repetitive elements                         | 11                                                      |
| Intron 2        | 2292 to 2318    | 22/27                                         | 81                   | Human Alu Type 1 and Type R, β-globin 5' flanking region | 6.6                                                     |
|                 | 2009 to 2114    | 63/106                                        | 59                   | Mouse H-2 Tla complex, pseudogene                    | 5.9                                                     |
| Intron 2 (reverse strand) | 2545 to 2479  | 41/67                                         | 61                   | Rat brain-specific identifier sequence RNA           | 9.5                                                     |
| Intron 3        | 3135 to 3223    | 70/89                                         | 79                   | Mouse pro-opiomelanocortin pseudogene                | 29                                                      |
|                 | 3130 to 3186    | 44/57                                         | 77                   | Sea urchin histone complex (GACA tract)              | 17                                                      |
|                 | 2981 to 3081    | 64/101                                        | 63                   | Human Alu Type T sequence                            | 7.2                                                     |
| Intron 3 (reverse strand) | 3255 to 3148  | 85/117                                        | 73                   | Mouse PR1 repetitive sequence and Type 2 Alu repeat  | 29                                                      |
|                 | 3124 to 2925    | 122/200                                       | 62                   | Mouse B1 ubiquitous repeat                           | 18                                                      |
|                 | 3062 to 2932    | 94/131                                        | 72                   | Mouse H-2 Tla complex, pseudogene                    | 15                                                      |
|                 | 3087 to 2925    | 93/163                                        | 57                   | Human Alu repeat B                                   | 15                                                      |
| Exon 5          | 3685 to 3762    | 51/78                                         | 65                   | Rat phenobarbital-inducible P-450, Exon 7            | 8.5                                                     |
| Intron 6        | 5738 to 5887    | 82/150                                        | 55                   | Rat brain-specific identifier sequence RNA           | 29                                                      |
|                 | 4846 to 4895    | 37/50                                         | 74                   | Human 7SL RNA                                        | 19                                                      |
|                 | 4846 to 4899    | 42/54                                         | 78                   | SV40 viable mutant IN149 DNA                         | 14                                                      |
|                 | 4846 to 4899    | 42/54                                         | 78                   | Human Alu Type T sequence                            | 14                                                      |
| Intron 6 (reverse strand) | 4904 to 4852  | 42/53                                         | 79                   | Human prothrombin gene, partial                      | 14                                                      |
|                 | 5593 to 5760    | 87/134                                        | 65                   | Rat seminal vesicle secretion IV gene                | 14                                                      |
|                 | 4924 to 4852    | 48/73                                         | 66                   | Mouse α-fetoprotein, 5' end and flanking region       | 11                                                      |
|                 | 4904 to 4852    | 39/53                                         | 74                   | Human c-myec oncogene                                 | 10                                                      |
|                 | 4905 to 4852    | 42/54                                         | 78                   | Human embryonic β-globin gene and two Alu family sequences | 10                                                      |
|                 | 5627 to 5473    | 84/155                                        | 54                   | Mouse H-2 Tla complex, pseudogene                    | 10                                                      |
| Exon 7          | 6205 to 6298    | 59/94                                         | 63                   | Rat phenobarbital-inducible P-450, Exons 8 and 9     | 9.6                                                     |
|                 | 6338 to 6397    | 47/60                                         | 78                   | Human immunoglobulin ε gene, C region                | 8.9                                                     |
| 3' Flanking region | +842 to +982   | 109/141                                       | 77                   | Rat growth hormone gene                               | 42                                                      |
|                 | +854 to +1026   | 122/173                                       | 71                   | Chinese hamster Alu-equivalent type 2 repeat.        | 39                                                      |
|                 | +849 to +1106   | 149/258                                       | 58                   | Mouse B1 repetitive elements                          | 38                                                      |
|                 | +848 to +982    | 104/135                                       | 77                   | Mouse B2 repeat sequence                             | 38                                                      |
|                 | +831 to +1001   | 118/171                                       | 69                   | Mouse U6 small nuclear RNA                           | 35                                                      |
|                 | +854 to +950    | 74/96                                         | 77                   | Mouse H-2β, 3' flanking region                       | 23                                                      |
|                 | +598 to +712    | 72/115                                        | 63                   | Rat repetitive sequence cluster                      | 21                                                      |
|                 | +830 to +919    | 61/90                                         | 68                   | Rat brain-specific identifier sequence RNA            | 15                                                      |
| 3' Flanking region (reverse strand) | +988 to +836  | 105/153                                       | 69                   | Mouse H-2 Tla complex, pseudogene                    | 29                                                      |
|                 | +1073 to +853   | 148/221                                       | 67                   | Mouse H-2β Eα chain gene, exons 2-4                  | 28                                                      |
|                 | +1100 to +855   | 132/246                                       | 54                   | Mouse PR1 repetitive sequence                        | 25                                                      |
|                 | +1063 to +704   | 162/360                                       | 45                   | Mouse immunoglobulin K germ-line J-C region          | 17                                                      |

about 14 to 23 kb (43, 44); (iii) the former has a 5' nontranslating exon and the latter does not (Table III); and (iv) exons 4, 5, and 6 of the former are highly conserved, whereas exons 7 and 8 of the latter are extraordinarily conserved (44). The 2nd exon of P1-450 and P3-450 ends with an AAG encoding Lys-276 (Table III). The 5th exon of P-450 ends with an AAG encoding Lys-274. By these types of comparisons, plus homologies found between several exons at the nucleotide

3 Y. Fujii-Kuriyama, personal communication.
level (Table IV) and among the amino acids encoded by distinct exons of the two P-450 gene families, we conclude that P1-450 and P3-450 exon 2 are similar to P-450c exons 1 through 5, P1-450 and P3-450 exon 3 is similar to P-450c exon 6, P1-450 and P3-450 exons 4 and 5 are similar to P-450c exon 7, P1-450 and P3-450 exon 6 is similar to P-450c exon 8, and P1-450 and P3-450 exon 7 is similar to P-450c exon 9. No inversions of genetic material have been detected during this evolutionary process. We therefore conclude that the ancestral P-450 gene had a minimum of 14 exons (46).

No significant homology with other published P-450 nucleotide sequences was found in any flanking region, intron, or on the negative DNA strand (Table IV). These data also indicate that there exists no evidence for inversion during evolution of these two P-450 gene families. Comparison of the two TCDD-inducible P-450 proteins from mouse and rat (12) results in an estimate of 1% divergence every 2.4 million years. We thus conclude that the P1-450 and P3-450 genes diverged from each other about 65 million years ago and that the TCDD-inducible and phenobarbital-inducible subfamilies separated from each other more than 200 million years ago. The data in Tables III and IV confirm further that these two gene families arose from a common ancestor.

Intermediate Repetitive Sequences—During the computer program gene search, several large stretches of DNA (50 to 360 bp) in the P1-450 and P3-450 introns were homologous to previously reported sequences (Table IV) at occurrence rates between 10 and 56 standard deviations above the mean. Such stretches appeared on both the positive strand and the reverse strand. These appear to represent highly repetitive and middle repetitive sequences, such as the PR1 and Alu types, many of which are found in the introns and flanking regions of other reported genes. None was found in any exon. Hence, P1-450 introns 1 and 2 and the P3-450 5’ flanking region, introns 1, 2, 3, and 6, plus the P3-450 3’ flanking region, all have repetitive elements of one or another family. Such insertions of these repetitive elements, in addition to duplication and unequal crossing-over of simple sequences which presumably have no function, undoubtedly account for the large differences in lengths of the P1-450 and P3-450 corresponding introns. These data show quite dramatically, during the approximately 65 million years that these genes have diverged, how the exon sequences and lengths are highly conserved yet the intron sequences and lengths clearly are not.

Conclusions—We have sequenced all introns and exons, plus a portion of the 5’ and 3’ flanking regions, of both the P1-450 and P3-450 genes in the C57BL/6N inbred mouse strain. Regions of possible regulatory importance have been pointed out. This study is an important prelude to experiments designed to prove a biological function for any of these interesting regions within the gene and/or flanking regions.

Acknowledgments—We thank our colleagues, especially David J. Lipman and Gerald R. Crabtree, for helpful discussions and reading of the manuscript. The valuable technical assistance of Mary Lynn Lipman and Gerald R. Crabtree, for helpful discussions and reading of the manuscript. We thank Ingrid E. Jordan for expert secretarial help.

Addendum—After this manuscript had been accepted, the sequence of the rat P-450c gene and flanking regions was reported (48). The following table illustrates the percentage of similarities and lengths of the exons and introns between mouse P1-450 and rat P-450c.

| Region          | Mouse P1-450 Length | Rat P-450c Length | Similarity |
|-----------------|---------------------|-------------------|------------|
| 5’ Flanking     | 823                 | (564)             | 75         |
| Exon 1          | 87                  | 87                | 86         |
| Intron 1        | 2380                | 2464              | 79         |
| Exon 2          | 851                 | 851               | 91         |
| Intron 2        | 748                 | 464               | 80         |
| Exon 3          | 127                 | 127               | 96         |
| Intron 3        | 87                  | 84                | 89         |
| Exon 4          | 90                  | 90                | 94         |
| Intron 4        | 93                  | 93                | 91         |
| Exon 5          | 124                 | 124               | 96         |
| Intron 5        | 150                 | 148               | 87         |
| Exon 6          | 87                  | 87                | 94         |
| Intron 6        | 137                 | 146               | 88         |
| Exon 7          | 1253                | 1278              | 88         |
| 3’ Flanking     | 1771                | (304)             | 75         |

In the translating region of exon 7, there is 93% similarity. The lengths of all corresponding exons, except exon 7, are identical and similarities range from 86 to 96%. The lengths of all corresponding introns are less than 1.6-fold different, with intron 4 identical; similarities range from 79 to 91%. The close similarities in this table, as compared with those in Table II, demonstrate that most of the exon and intron differences between P1-450 and P3-450 occurred between the time that P1-450 and P3-450 split (about 65 million years ago) and the rat-mouse divergence (about 17 million years ago).

In exon 2, there is a 492-bp stretch with 96% similarity between P1-450 and P3-450 (12); a similar stretch of similarity is seen between rat P-450c (46) and P-450d (49). If an intra-exon gene conversion event had occurred prior to the rat-mouse split 17 million years ago, all divergences (P1-450 versus P3-450; P-450d versus P-450c; P-450d versus P-450c; and P-450d versus P-450d) should be the same. Because they are not, this finding argues in favor of conservation of this exon 2 region related to gene function.

In the first 172 bases upstream of the cap site, there are 157 matches (91%) between mouse P1-450 and rat P-450c. Among the three highly homologous boxes upstream from the cap site (Figs. 2 and 3), the 3’-most box is 75% similar (9/12), the middle box is 100% similar (11/11), and the 5’-most box cannot be compared because the reported rat P-450c sequence (48) does not extend far enough upstream. The (PuPy), stretches noted in the 5’ flanking region of rat P-450c (48) are not present in mouse P-450, suggesting that this DNA having Z-DNA-forming potential probably has no important regulatory function. Conservation of portions of the 5’ flanking region between mouse P1-450 and rat P-450c thus suggests the presence of important regulatory elements.

REFERENCES

1. Negishi, M., and Nebert, D. W. (1979) J. Biol. Chem. 254, 11015–11023
2. Ohyama, T., Nebert, D. W., and Negishi, M. (1984) J. Biol. Chem. 259, 2675–2682
3. Reik, L. M., Levin, W., Ryan, D. E., and Thomas, P. E. (1982) J. Biol. Chem. 257, 3950–3957
4. Norman, R. L., Johnson, E. F., and Muller-Eberhard, U. (1978) J. Biol. Chem. 253, 8640–8647
5. Haugen, D. A., van der Hoeven, T. A., and Coon, M. J. (1975) J. Biol. Chem. 250, 3561–3570
6. Nebert, D. W., Negishi, M., Lang, M. A., Hjelmeland, L. M., and Eisen, H. J. (1982) Adv. Genet. 21, 1–52
7. Eisen, H. J., Hannah, R. R., Legraverend, C., Okey, A. B., and Nebert, D. W. (1984) in Biochemical Actions of Hormones (Litwack, G., ed) pp. 227–258, Academic Press, New York
8. Owens, I. S. (1977) J. Biol. Chem. 252, 2927–2933
9. Kumaki, K., Jensen, N. M., Shire, J. G. M., and Nebert, D. W. (1979) Gene (Amst.) 29, 281–292
10. Gonzalez, F. J., Mackenzie, P. L., Kimura, S., and Nebert, D. W. (1984) Gene (Amst.) 29, 281–292
11. Kimura, S., Gonzalez, F. J., and Nebert, D. W. (1984) Nucleic Acids Res. 12, 2917–2928
12. Kimura, S., Gonzalez, F. J., and Nebert, D. W. (1984) J. Biol.
Mouse P_{1}\text{-}450 and P_{2}\text{-}450 Genes

13. Tukey, R. H., Lalley, P. A., and Nebert, D. W. (1984) *Proc. Natl. Acad. Sci. U. S. A.* **81**, 3163–3166

14. Gonzalez, F. J., Tukey, R. H., and Nebert, D. W. (1984) *Mol. Pharmacol.* **26**, 117–121

15. Tukey, R. H., Hannah, R. R., Negishi, M., Nebert, D. W., and Eisen, H. J. (1982) *Cell* **31**, 275–284

16. Guenther, T. M., and Nebert, D. W. (1978) *Eur. J. Biochem.* **91**, 449–456

17. Ikeda, T., Altieri, M., Chen, Y.-T., Nakamura, M., Tukey, R. H., Nebert, D. W., and Negishi, M. (1983) *Eur. J. Biochem.* **134**, 13–18

18. Rimn, D. L., Hogness, J., Kucera, J., and Blattner, F. R. (1980) *Nature (Lond.)* **282**, 480–486

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

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

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

22. Sanger, F., and Coulson, A. R. (1978) *FEBS Lett.* **87**, 107–110

23. Staden, R. (1980) *Nucleic Acids Res.* **8**, 3673–3694

24. Hagenbuchle, O., Bovery, R., and Young, R. A. (1980) *Cell* **21**, 179–187

25. Brethnach, R., and Chambon, P. (1981) *Annu. Rev. Biochem.* **50**, 349–383

26. Okayamn, H., and Berg, P. (1983) *Mol. Cell. Biol.* **3**, 280–289

27. Corden, J., Wasylyk, B., Buchwalder, A., Sassone-Corsi, P., Kedinger, C., and Chambon, P. (1980) *Science (Wash. D. C.)* **209**, 1405–1414

28. Efstratiatis, A., Pasakony, J. W., Maniatis, T., Lawn, R. M., O’Connell, C., Spritz, R. A., DeRiel, J. K., Forget, B. G., Weissman, S. M., Slightom, J. L., Blechl, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C., and Proudfoot, N. J. (1980) *Cell* **21**, 653–668

29. Benoist, C., O’Hare, K., Brethnach, R., and Chambon, P. (1980) *Nucleic Acids Res.* **8**, 127–142

30. Scheideret, C., Geisse, S., Westphal, H. M., and Beato, M. (1983) *Nature (Lond.)* **304**, 749–752

31. Jost, J.-P., Seldran, M., and Geiser, M. (1984) *Proc. Natl. Acad. Sci. U. S. A.* **81**, 429–433

32. Borgmeyer, U., Nowock, J., and Sippel, A. E. (1984) *Nucleic Acids Res.* **12**, 4295–4311

33. Fowikes, D. M., Mullis, N. T., Comeau, C. M., and Crabtree, G. R. (1984) *Proc. Natl. Acad. Sci. U. S. A.* **81**, 2313–2316

34. Ostrowski, M. C., Huang, A. L., Kessel, M., Wolford, R. G., and Hager, G. L. (1984) *Eur. Mol. Biol. Organ.* **3**, 1891–1899

35. Rosenthal, N., Kress, M., Gruss, P., and Khoury, G. (1983) *Science (Wash. D. C.)* **222**, 749–755

36. Tuan, D., and London, I. M. (1984) *Proc. Natl. Acad. Sci. U. S. A.* **81**, 2718–2722

37. Brady, J., Radonovich, M., Thoren, M., Das, G., and Salzman, N. P. (1984) *Mol. Cell. Biol.* **4**, 133–141

38. Wang, A. H.-J., Quigley, G. J., Kolpak, F. J., Crawford, J. L., van Boom, J. M., van der Marel, G., and Rich, A. (1979) *Nature (Lond.)* **282**, 680–686

39. Desiderio, S., and Baltimore, D. (1984) *Nature (Lond.)* **308**, 860–862

40. Shimizu, A., and Honjo, T. (1984) *Cell* **36**, 801–803

41. Tautz, D., and Renz, M. (1984) *Nucleic Acids Res.* **12**, 4127–4138

42. Ozols, J., Heinemann, F. S., and Johnson, E. F. (1981) *J. Biol. Chem.* **256**, 11405–11408

43. Mizukami, Y., Sogawa, K., Suwa, Y., Muramatsu, M., and Fujii-Kuriyama, Y. (1983) *Proc. Natl. Acad. Sci. U. S. A.* **80**, 3958–3962

44. Atchison, M., and Adesnik, M. (1983) *J. Biol. Chem.* **258**, 11285–11295

45. Phillips, I. R., Shepherd, E. A., Ashworth, A., and Rabin, B. R. (1983) *Gene* **24**, 41–52

46. Nebert, D. W., Kimura, S., and Gonzalez, F. J. (1985) in *Microsomes and Drug Oxidations* (Boobis, A., Caldwell, J., and Davies, D., eds) Taylor and Francis Ltd., London, in press

47. Wilbur, W. J., and Lipman, D. J. (1983) *Proc. Natl. Acad. Sci. U. S. A.* **80**, 726–730

48. Sogawa, K., Gotob, O., Kawaiji, K., and Fujii-Kuriyama, Y. (1984) *Proc. Natl. Acad. Sci. U. S. A.* **81**, 5066–5070

49. Kawajiri, K., Gotob, O., Sogawa, K., Tagashira, Y., Muramatsu, M., and Fujii-Kuriyama, Y. (1984) *Proc. Natl. Acad. Sci. U. S. A.* **81**, 1649–1653