Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 3445–3449, May 1985
Medical Sciences

Nucleotide sequence and structure of the human apolipoprotein E gene

(*Alu* sequences/*apo-E4* genotype/5′ flanking region/S1 nuclease mapping/intron)

YOUNG-KI PAIK*, DAVID J. CHANG*, CATHERINE A. REARDON*, GLENN E. DAVIES*†, ROBERT W. MAHLEY*‡§, AND JOHN M. TAYLOR*‡¶‖

*Gladstone Foundation Laboratories for Cardiovascular Disease, ‡Cardiovascular Research Institute, ¶Department of Physiology and §Departments of Pathology and Medicine, University of California, San Francisco, CA 94140

Communicated by Stanley Cohen, January 7, 1985

ABSTRACT **The gene for human apolipoprotein E (apo-E) was selected from a library of cloned genomic DNA by screening with a specific cDNA hybridization probe, and its structure was characterized. The complete nucleotide sequence of the gene as well as 856 nucleotides of the 5′ flanking region and 629 nucleotides of the 3′ flanking region were determined. Analysis of the sequence showed that the mRNA-encoding region of the *apo-E* gene consists of four exons separated by three introns. In comparison to the structure of the mRNA, the introns are located in the 5′ noncoding region, in the codon for glycine at position −4 of the signal peptide region, and in the codon for arginine at position +61 of the mature protein. The overall lengths of the *apo-E* gene and its corresponding mRNA are 3597 and 1163 nucleotides, respectively; a mature plasma protein of 299 amino acids is produced by this gene. Examination of the 5′ terminus of the gene by S1 nuclease mapping shows apparent multiple transcription initiation sites. The proximal 5′ flanking region contains a "TATA box" element as well as two nearby inverted repeat elements. In addition, there are four *Alu* family sequences associated with the *apo-E* gene: an *Alu* sequence located near each end of the gene and two *Alu* sequences located in the second intron. This knowledge of the structure permits a molecular approach to characterizing the regulation of the *apo-E* gene.**

Apolipoprotein E (apo-E) is a component of various classes of plasma lipoproteins in all mammals that have been studied (for review, see refs. 1 and 2). It is a single chain polypeptide (M_r, 34,000) of 299 amino acids (3) that is synthesized initially with an 18-residue signal peptide that is removed cotranslationally (4, 5). The amino acid sequence as well as the mRNA nucleotide sequence are known for both the human (3, 6) and rat (7) species. The major site of synthesis is the liver, but relatively abundant levels of apo-E mRNA have been detected in many extrahepatic tissues, including the brain and the adrenals (8). In addition, apo-E is produced by mouse peritoneal macrophages, as well as human monocyte-derived macrophages (9).

A major function of apo-E is its mediation of the cellular uptake of specific lipoproteins through an interaction with apo-B,E(LDL) receptors on extrahepatic and hepatic cell surfaces and with distinct hepatic apo-E receptors (for review, see ref. 10). The receptor binding domain of human apo-E has been determined to be an arginine- and lysine-rich region in the vicinity of residues 140 and 160 (11, 12). Variant forms of apo-E with single amino acid substitutions in this region show decreased receptor binding activity (13–15) and are associated with type III hyperlipoproteinemia and accelerated cardiovascular disease (for review, see refs. 16 and 17). Apolipoprotein E with normal receptor binding activity is found in two common isoforms, the E3 and E4 phenotypes, with either cysteine or arginine, respectively, at residue position 112 (13).

Because of the central role that apo-E plays in the metabolism of cholesterol and other lipids, knowledge of the regulation of the *apo-E* gene is important in understanding the alterations in lipid metabolism that occur in normal and pathological processes. Therefore, to provide a molecular basis for examining its regulation, we have determined and analyzed the nucleotide sequence of the human *apo-E* gene and its proximal flanking regions.

EXPERIMENTAL PROCEDURES

DNA Library Screen. A human genome library of random, partially *Hae* III/*Alu* I-digested fragments of fetal liver DNA contained in the Charon 4A λ bacteriophage (18) was provided through the generosity of T. Maniatis (Harvard University). The phage was grown in *Escherichia coli* LE392 and screened essentially as described (19). About two million phage plaques were screened with a ^{32}P-labeled (20) restriction endonuclease fragment that was purified from a previously characterized (6) full-length cloned cDNA to human apo-E mRNA. A single recombinant bacteriophage was identified, and the DNA was prepared from plaque-purified material (19). All experiments were done in accordance with the National Institutes of Health Guidelines.

DNA Mapping, Subcloning, and Sequencing. Bacteriophage recombinant DNA was digested with various restriction endonucleases (Boehringer Mannheim and New England Biolabs) according to the suppliers' directions and was examined by electrophoresis in 0.8% agarose gels. The DNA was transferred to nitrocellulose filters by blotting (21), then hybridized to the ^{32}P-labeled apo-E cDNA probe, and examined by autoradiography to identify *apo-E* gene fragments. Based on these studies, *Eco*RI- and *Bam*HI-digested DNA fragments were subcloned into plasmid pUC9 (22), and *apo-E* gene-containing recombinants were selected as described above. The *apo-E* gene DNA inserts in the subclones were sequenced by the method of Maxam and Gilbert (23).

S1 Nuclease Mapping. A 67-base-pair *Bst*NI/*Hin*dIII restriction endonuclease fragment from an *apo-E* gene subclone was prepared (23) that contained a portion of the first exon, the transcription initiation site, and a portion of the 5′-terminal flanking region. The fragment was end-labeled at the 5′ ends by [γ-^{32}P]ATP and T4 polynucleotide kinase, and

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: apo, apolipoprotein; kb, kilobases.

†Present address: Boehringer Mannheim, P.O. Box 50816, Indianapolis, IN 46250.
‖To whom reprint requests should be addressed.

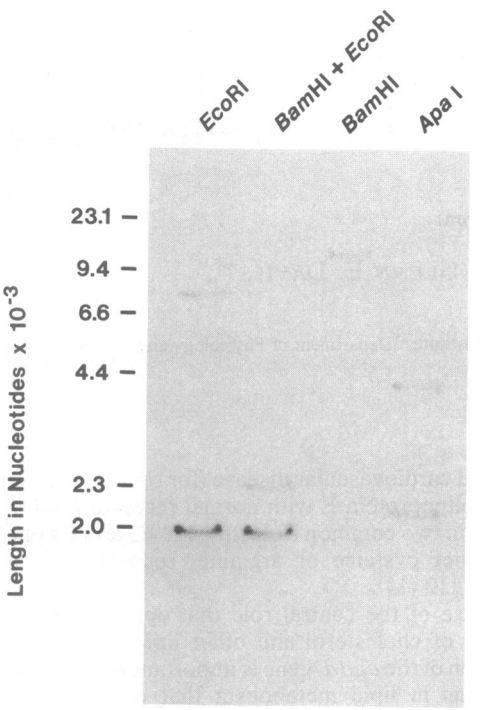

FIG. 1. Hybridization of apo-E cDNA to restriction endonuclease-digested human genome DNA. Ten micrograms of human genome DNA was digested with *Eco*RI, *Eco*RI/*Bam*HI, *Bam*HI, and *Apa* I. The digested DNA was analyzed by gel electrophoresis, blotting, and hybridization to a cloned ^{32}P-labeled apo-E cDNA. The molecular weight markers are *Hin*dIII-digested fragments of bacteriophage λ DNA.

a portion of the DNA was purified by electrophoresis in strand separating gels (23). Both double-stranded and single-stranded labeled fragments were hybridized to 5 μg of human liver poly(A)-containing RNA, and the double-stranded probe was hybridized to 20 μg of total human brain RNA at 42°C, essentially as described (24). The hybrids were digested with 100 units of S1 nuclease per ml (24) and then analyzed by electrophoresis on a 15% sequencing gel (23).

Genome DNA Analysis. In separate reactions, 10 μg of human genome DNA prepared (25) from total leukocytes or frozen liver was digested with 100 units of various restriction endonucleases, electrophoresed in 0.8% agarose gels, and transferred to nitrocellulose filters by blotting (21). The filters were hybridized to 10^6 cpm/ml of a 1117-base-pair *Aat* II/*Hin*fI restriction endonuclease fragment of the cloned apo-E cDNA (26). The 1117-base-pair fragment represented 96% of the length of the mRNA-encoding exons.

RESULTS AND DISCUSSION

Analysis of Genome DNA. The *apo-E* gene in human genome DNA was examined by restriction endonuclease analysis (Fig. 1). The enzymes used and the corresponding approximate lengths of the genome DNA fragments that hybridized to the apo-E cDNA were as follows: *Eco*RI, 1.9 and 8.0 kilobases (kb); *Eco*RI/*Bam*HI, 1.9 and 2.3 kb; *Bam*HI, 12.0 kb; *Apa* I, 4.2 and 2.1 kb. Because only one or two fragments hybridized to the probe in each case, it is likely that the human haploid genome contains only one copy of the *apo-E* gene.

Characterization of the Cloned *apo-E* Gene. The recombinant λ bacteriophage DNA containing the *apo-E* gene, which had been selected by cDNA screening, was examined by restriction endonuclease mapping and by hybridization of ^{32}P-labeled apo-E cDNA to DNA filter blots (data not shown). The probe hybridized to fragments of the same size as those found in the genome DNA digests for *Apa* I, *Eco*RI, and *Eco*RI/*Bam*HI, suggesting that no rearrangements in the structure of the *apo-E* gene had occurred upon cloning. The recombinant phage contained ≈20 kb of inserted human DNA, which included the intact *apo-E* gene, which was subcloned and mapped in detail. The results indicated that the mRNA coding region of the *apo-E* gene consisted of four coding segments (exons) that were interrupted by three noncoding segments (introns).

Nucleotide Sequence Analysis. The strategy used to determine the complete nucleotide sequence of the *apo-E* gene is shown in Fig. 2. In addition to the exon and intron segments, 856 nucleotides of the 5′ flanking region and 629 nucleotides of the 3′ flanking region were determined.

The complete nucleotide sequence of the human *apo-E* gene and its proximal flanking sequences are shown in Fig. 3. A comparison of this sequence to the previously determined (6) nucleotide sequence of the apo-E cDNA identified the locations of the exon–intron junctions. All introns begin with the nucleotides G-T and end with the nucleotides A-G, which is consistent with the consensus sequence for exon–intron splice junctions for eukaryotic genes (28). In this regard, the precise locations of the third and fifth exon–intron junctions of the *apo-E* gene were established by taking the consensus sequence into account. The lengths of the exons are 44, 66, 193, and 860 nucleotides, and the intron lengths are 760, 1092, and 582 nucleotides in their 5′ to 3′ order. In comparison to the corresponding mRNA sequence (6), the first intron occurs in the 5′ noncoding region following guanine at position −78 of the mRNA (G 900 in Fig. 3); the second intron occurs in the codon for glycine at position −4 of the signal peptide region following guanine at position −12 of the mRNA (G 1660 in Fig. 3); and the third intron occurs in the codon for arginine at position +61 of the mature plasma

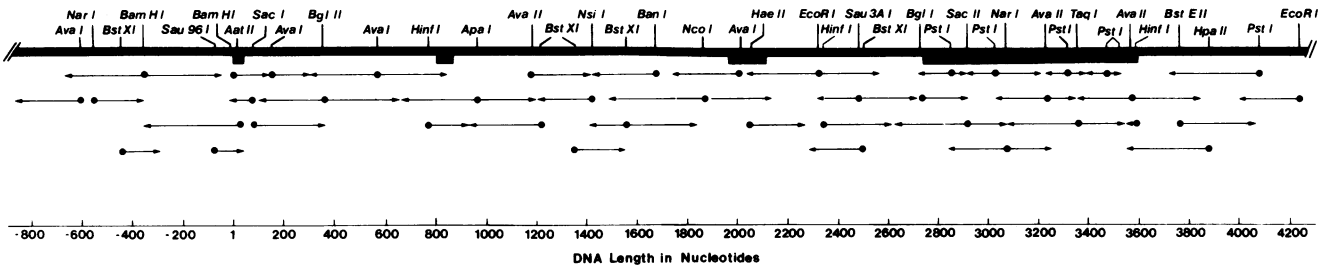

FIG. 2. Restriction endonuclease map and nucleotide sequencing strategy for the *apo-E* gene. The *apo-E* gene and its flanking regions are represented on the top line, with relative positions of the exons shown by boxes below the line. Restriction endonuclease sites indicated are only those used in the sequence determination and do not necessarily reflect the total number of sites present. Solid circles represent sites of ^{32}P labeling of the restriction endonuclease fragments. Arrows represent the direction and length of fragment sequencing. The site of *Bst*EII cleavage in the 3′ flanking region of the gene did not correspond to the typical sequence (27) recognized by this enzyme, and it was not digested by some of the enzyme preparations.

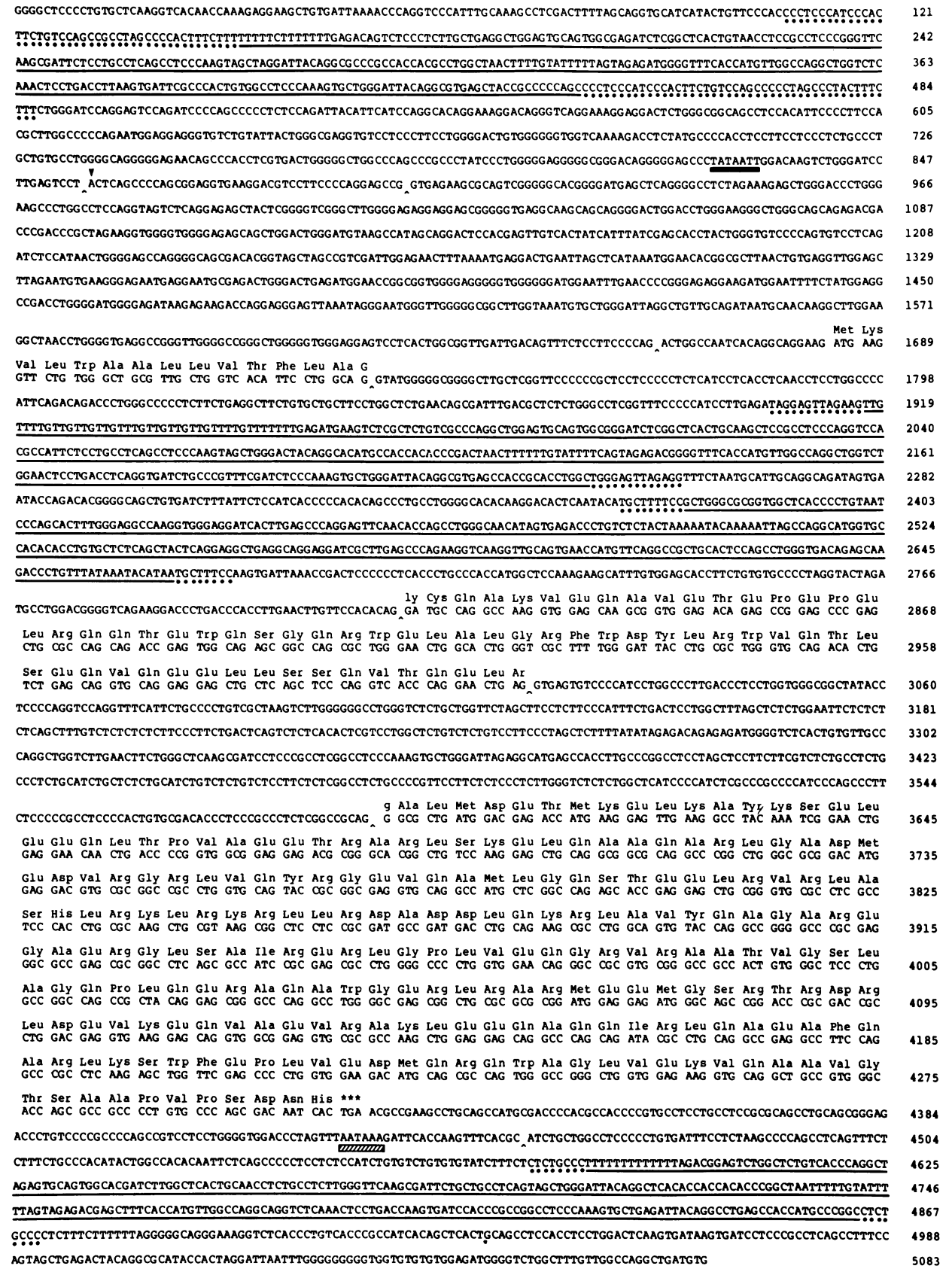

FIG. 3. Nucleotide sequence of the *apo-E* gene and proximal flanking regions. Complete nucleotide sequence of the human *apo-E* gene and proximal flanking regions are shown. Beginning and end of the gene, corresponding to the final mRNA product, as well as the exon–intron junctions are indicated by marks (^) below the sequence. Blanks at these positions and between the amino acid codons in the exons are for clarity only and do not indicate missing nucleotides. Major transcription initiation site is indicated at the beginning of the first exon (▼). The group of three asterisks indicates the translation termination codon in the fourth exon. The TATA box element in the 5′ flanking region (■), and the A-A-T-A-A-A poly(A) signal (▨) in the fourth exon are underlined. Locations of *Alu* family members (—) and their flanking repeated sequences (·····) are underlined.

protein region following guanine at position 182 of the mRNA (G 3011 in Fig. 3). Thus, the overall length of the *apo-E* gene is 3597 nucleotides, which encodes a mRNA of 1163 nucleotides.

The nucleotide sequences of the exons of this *apo-E* gene differed from the previously reported normal E3 apo-E cDNA in four positions, all of which were located in the coding region for the mature plasma protein. One of these nucleotide differences, thymine instead of cytosine at base 334 of the mRNA (base 3745 in Fig. 3), predicts an arginine at amino acid residue 112 instead of a cysteine. Thus, the gene sequence reported here encodes the E4 variant of apo-E. The other nucleotide differences were guanine substituted for adenine in the third base position of the codon for glutamine. These positions in the protein are at glutamine +55, +58, and +248, corresponding to mRNA base positions 165, 174, and 744 (bases 2994, 3003, and 4155 in Fig. 3). Thus, these nucleotide changes do not result in amino acid changes.

S1 Nuclease Mapping. The 5′ terminus of the corresponding apo-E mRNA was determined by S1 nuclease mapping. Because the sequence of 61 nucleotides of the 5′ noncoding region of apo-E mRNA had been determined previously (6) from the cDNA sequence (beginning at nucleotide 7 in Fig. 4*B*), a restriction endonuclease fragment from a gene subclone was prepared that included the distal portion of this region as well as a portion of the 5′ upstream region. The fragment was ^{32}P-labeled at the 5′ ends, hybridized to liver mRNA, and digested with S1 nuclease. As shown in Fig. 4*A*, apo-E mRNA protected two clusters of subfragments from S1 nuclease digestion, suggesting that two or more transcription initiation sites might be present in the *apo-E* gene. The same digestion pattern was observed whether single-stranded or double-stranded probes were used and with different amounts of S1 nuclease. The appearance of minor subfragments may be caused by the bulky 5′ cap structure on the mRNA, which could interfere with duplex formation at the corresponding end of the DNA·mRNA hybrid and allow additional S1 nuclease digestion as reported (29). Since most eukaryotic mRNAs start with adenine (30), the likely 5′ terminus of the major portion of apo-E mRNA lies 67 nucleotides upstream (at nucleotide 1 in Fig. 4*B*) from the initiation codon. It is also probable that apo-E mRNA has at least one (nucleotide −3 in Fig. 4*B*) or more additional 5′-terminal start sites. In addition, no differences in digestion patterns were observed between the reactions with liver RNA (Fig. 4*A*, lanes A and B) and brain RNA (lane D), which suggests that the initiation sites for the *apo-E* gene are the same in both tissues.

The 5′ Flanking Region of the *apo-E* Gene. An examination of the nucleotide sequence of the 5′ flanking region of the *apo-E* gene adjacent to the transcription initiation site revealed several potentially important sequence elements. The sequence T-A-T-A-A-T-T begins at nucleotide −33 (Fig. 4*B*). This sequence is homologous to the "TATA box" sequence that has been identified as a component of the promoter region for most eukaryotic genes (30).

In addition, two major inverted repeated sequences are located within the 150 nucleotides adjacent to the mRNA start site. The proximal element is located between nucleotides −76 and −46, and the distal element is located between nucleotides −144 and −108. These sequences are illustrated in Fig. 4*B*, and they include all potential base pairs. The large number of G·C base pairs in both sequences suggests that these palindrome-like structures might be stable naturally occurring elements.

***Alu* Family Sequences of the *apo-E* Gene.** An examination of the introns and proximal flanking regions of the *apo-E* gene shows that there are four members of the *Alu* family of repeated sequences (31) associated with the gene. Two of these sequences are located in the second intron, and there is an *Alu* sequence located close to each end of the gene in the nontranscribed flanking regions (Fig. 5). Their lengths range from 280 to 324 nucleotides. In their structural orientation, one of the *Alu* sequences located in the second intron is

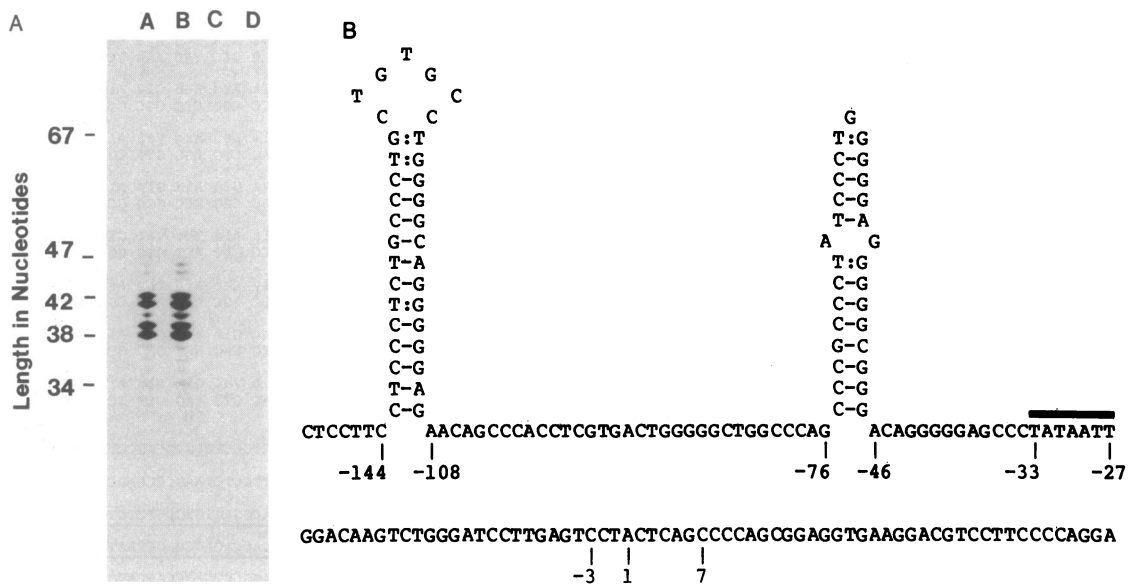

FIG. 4. Analysis of transcription initiation site and 5′ flanking region of the *apo-E* gene. (*A*) S1 nuclease protection analysis of transcription initiation site of the *apo-E* gene. The 67-base-pair ^{32}P-end-labeled gene fragment was used in three separate reactions as either a double-stranded (lane A), single-stranded noncoding strand (lane B), or single-stranded coding strand (lane C) hybridization probe for liver poly(A)-containing RNA. The double-stranded probe was also hybridized to total brain RNA (lane D). Bands shown are the DNA fragments that were protected from S1 nuclease digestion. A trace amount of residual undigested probe is visible at the 67-base-pair length marker. Nucleotide lengths were determined from examination of the partial degradation products of a standard nucleotide sequence reaction run in an adjacent lane. (*B*) Nucleotide sequence of 150 nucleotides of the proximal 5′ region adjacent to the transcription initiation site and of 40 nucleotides of the first exon of the *apo-E* gene. Numbers indicate nucleotide positions relative to the initiation site (position 1). The TATA box site is indicated by a bar. Inverted repeated sequences are shown with all potential base pairs, with G·T base pairs indicated by a colon. First and second inverted repeats have a calculated ΔG of −33 and −26 kcal/mol, respectively.

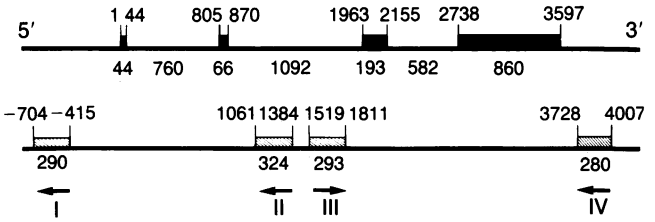

FIG. 5. *Alu* family sequences associated with the human *apo-E* gene. Schematic outlines of the *apo-E* gene with relative positions of exons (solid boxes, upper line) and *Alu* sequences (shaded boxes, lower line) are indicated. Nucleotide sequence positions of the first and last nucleotide of each element relative to transcription initiation site (position 1) are shown above the lines. Lengths of sequence elements are shown below the lines. Arrows show orientation of the *Alu* sequences relative to the coding strand of the *apo-E* gene. Roman numerals indicate the individual *Alu* sequence family members.

oriented with the same polarity as the mRNA coding sequences of the exons, whereas the other three *Alu* family sequences have the opposite orientation.

Upon alignment for maximum sequence homology, the individual *Alu* family members show an 81%–90% identity in their nucleotide positions. The *Alu* family sequence located in the 5′ flanking region of the *apo-E* gene is bounded on each end by an unusually long directly repeated sequence of 45 nucleotides. Short repeated sequences of 13, 8, and 8 nucleotides, respectively, flank the other three *Alu* sequence family members.

Structural Comparison of the *apo-E* Gene to the *apo-A-I* and *apo-C-III* Genes. The overall structural organization of the human *apo-E* gene is similar to that of the human *apo-A-I* and *apo-C-III* genes (32–34), which also are composed of four exons and three introns. The relative locations of the introns are quite similar in all three genes, with the second intron of the *apo-E* and *apo-A-I* genes located at exactly the same place in the signal peptide coding region of the corresponding mRNAs. Furthermore, the second exon is nearly the same length in all three genes, and it encodes most of the signal peptide for the respective proteins. In contrast, the lengths of the introns vary substantially among these genes at each position.

The general structure of these three apolipoprotein genes suggests that their evolutionary development may have been influenced by common exonic requirements. In this regard, the genes give rise to secreted proteins having homologous amphipathic lipid-binding regions encoded by their fourth exons (reviewed in ref. 2), with length differences in these exons relating closely to the lengths of the corresponding proteins. The third exons encode the amino-terminal regions of the mature secreted proteins in each case. These regions have no obvious interrelationships and may contribute to the functional differences among the apolipoproteins. The first exon is relatively short in each gene and is contained within the 5′ nontranslated portion of the corresponding mRNA.

However, despite the similar organization of these apolipoprotein genes, they have substantial differences in their nucleotide and derived amino acid sequences, in the functions of their encoded proteins, and in the regulation of their expression. Thus, a broad understanding of the evolutionary relationships among the apolipoprotein genes may require the additional knowledge of the sequence and structure of the other members of this gene family.

Helpful suggestions by Chris Lau and Mary Ann Gholson are gratefully appreciated. The graphics assistance of James X. Warger and Norma Jean Gargasz as well as the editorial assistance of Barbara Allen and Sally Gullatt Seehafer are acknowledged. Gratitude is expressed to Dr. Hugo Martinez for making available the computer facilities of the Biomathematics Computation Laboratory of the University of California at San Francisco.

1. Mahley, R. W., Innerarity, T. L., Rall, S. C., Jr. & Weisgraber, K. H. (1984) *J. Lipid Res.* **25,** 1277–1294.
2. Mahley, R. W. (1979) in *Atherosclerosis Reviews,* eds. Paoletti, R. & Gotto, A. M., Jr. (Raven, New York), Vol. 5, pp. 1–34.
3. Rall, S. C., Jr., Weisgraber, K. H. & Mahley, R. W. (1982) *J. Biol. Chem.* **257,** 4171–4178.
4. Lin-Lee, Y. C., Tanaka, Y., Lin, C. T. & Chan, L. (1981) *Biochemistry* **20,** 6474–6480.
5. Reardon, C., Driscoll, D., Hay, R., Reddy, G., Kohler, H. & Getz, G. S. (1981) *Circulation* **64,** IV-16.
6. McLean, J. W., Elshourbagy, N. A., Chang, D. J., Mahley, R. W. & Taylor, J. M. (1984) *J. Biol. Chem.* **259,** 6498–6504.
7. McLean, J. M., Fukazawa, C. & Taylor, J. M. (1983) *J. Biol. Chem.* **258,** 8993–9000.
8. Elshourbagy, N. A., Liao, W. S., Mahley, R. W. & Taylor, J. M. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 203–207.
9. Basu, S. K., Ho, Y. K., Brown, M. S., Bilheimer, D. W., Anderson, G. W. & Goldstein, J. L. (1982) *J. Biol. Chem.* **257,** 9788–9795.
10. Mahley, R. W. & Innerarity, T. L. (1983) *Biochim. Biophys. Acta* **737,** 197–222.
11. Weisgraber, K. H., Innerarity, T. L., Harder, K. J., Mahley, R. W., Milne, R. W., Marcel, Y. L. & Sparrow, J. T. (1983) *J. Biol. Chem.* **258,** 12348–12354.
12. Innerarity, T. L., Friedlander, E. J., Rall, S. C., Jr., Weisgraber, K. H. & Mahley, R. W. (1983) *J. Biol. Chem.* **258,** 12341–12347.
13. Weisgraber, K. H., Innerarity, T. L. & Mahley, R. W. (1982) *J. Biol. Chem.* **257,** 2518–2521.
14. Rall, S. C., Jr., Weisgraber, K. H., Innerarity, T. L. & Mahley, R. W. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 4696–4700.
15. Rall, S. C., Jr., Weisgraber, K. H., Innerarity, T. L., Bersot, T. P., Mahley, R. W. & Blum, C. B. (1983) *J. Clin. Invest.* **72,** 1288–1297.
16. Mahley, R. W. & Angelin, B. (1984) *Adv. Intern. Med.* **29,** 385–411.
17. Mahley, R. W. (1983) *Klin. Wochenschr.* **61,** 225–232.
18. Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G. K. & Efstratiadis, A. (1978) *Cell* **15,** 687–701.
19. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
20. Tu, C. P. & Wu, R. (1980) *Methods Enzymol.* **65,** 620–638.
21. Southern, E. M. (1979) *Methods Enzymol.* **68,** 152–176.
22. Vieira, J. & Messing, J. (1982) *Gene* **19,** 259–268.
23. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65,** 499–560.
24. Berk, A. & Sharp, P. A. (1977) *Cell* **12,** 721–732.
25. Goosens, M. G. & Kan, Y. W. (1981) *Methods Enzymol.* **76,** 805–817.
26. Feinberg, A. P. & Vogelstein, B. (1983) *Anal. Biochem.* **132,** 6–13.
27. Lautenberger, J. A., Edgell, M. H. & Hutchinson, C. A., III (1980) *Gene* **12,** 171–174.
28. Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. & Chambon, P. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 4853–4857.
29. Weaver, R. F. & Weissmann, C. (1979) *Nucleic Acids Res.* **7,** 1175–1193.
30. Breathnach, R. & Chambon, P. (1981) *Annu. Rev. Biochem.* **50,** 349–383.
31. Jelinek, W. R. & Schmid, C. W. (1982) *Annu. Rev. Biochem.* **51,** 813–844.
32. Karathanasis, S. K., Zannis, V. I. & Breslow, J. L. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 6147–6151.
33. Shoulders, C. C., Kornblihtt, A. R., Munro, B. S. & Baralle, F. E. (1983) *Nucleic Acids Res.* **9,** 2827–2837.
34. Protter, A. A., Levy-Wilson, B., Miller, J., Bencen, G., White, T. & Seilhamer, J. J. (1984) *DNA* **3,** 449–456.