Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 8085–8089, December 1985
Evolution

Evolution of apolipoprotein E: Mouse sequence and evidence for an 11-nucleotide ancestral unit

(repetitive structure/cDNA/amphipathic helix/tandem duplication/RNA sequencing)

TRIPATHI B. RAJAVASHISTH, JOHN S. KAPTEIN, KAREN L. REUE, AND ALDONS J. LUSIS*

Departments of Medicine and Microbiology, and Molecular Biology Institute, University of California, Los Angeles, CA 90024

Communicated by Susumu Ohno, July 26, 1985

ABSTRACT **Apolipoprotein E (apo E) is responsible for the binding of very low density lipoprotein and chylomicron remnants to cellular receptors thereby removing them from circulation. We have isolated and determined the sequence of a cDNA encoding 285 amino acids and the entire 3′ untranslated region of 112 nucleotides of mouse apo E. The remaining coding sequence was determined by sequencing mouse liver mRNA. Comparisons with rat and human apo E sequences showed a high degree of conservation although there were regions in each species that were characterized by unique insertions and deletions. Analysis of the sequence homologies within apo E revealed that the entire sequence is made up of repetitive units. The most primitive unit appeared to be an 11-nucleotide repeat within higher order repeats of 22 or 33 nucleotides. The 11-nucleotide unit -TCGGACGAGGC- is read in all three reading frames, and when tandemly repeated, it encodes the highly conserved amino acid sequence Xaa-(Glu/Asp)-(Glu/Asp)-Xaa-Arg-Xaa-Arg-Leu-Gly-Xaa-Xaa. We postulate that apo E and those other apolipoproteins related to it have arisen by duplications and subsequent modifications of this or a closely related 11-nucleotide ancestral sequence.**

Apolipoprotein E (apo E) plays a central role in mammalian lipoprotein metabolism by serving both as a structural component for a diverse group of lipoprotein types and as a mediator of lipoprotein catabolism by specific cell surface receptors (1–3). Mature human apo E is a 299-amino acid polypeptide of known sequence (4), and the nucleotide sequences of human and rat apo E cDNA clones have been reported (5–7). Apo E as well as other apolipoproteins contain 11- or 22-amino acid repeated regions as dominant features (8–12). These appear to encode largely amphipathic α-helices, which have been implicated in lipid binding (13, 14).

The mouse is being utilized as a model in the study of lipid transport and metabolism because of the advantages it offers for genetic analysis (15). We report here the nucleotide sequence of mouse apo E mRNA and a detailed analysis of internal homology within the sequence. We also present evidence that apo E and other members of the apolipoprotein gene family may have evolved from a tandemly repeated 11-base-pair unit, with each successive unit being read in a different reading frame. After three such units, the translational repeat length of 11 amino acids would also repeat, as has been observed in modern day apolipoproteins.

EXPERIMENTAL PROCEDURES

The cDNA library construction and screening were as described (16). DNA sequencing was done by both the dideoxy chain termination (17) and the chemical cleavage methods (18). Synthetic oligonucleotides used as sequencing primers were synthesized and purified following the protocol of Matteucci and Caruthers (19).

Apo E mRNA was sequenced as follows: A 14-base oligonucleotide corresponding to the 5′ end of the cDNA clone (nucleotides 89–102) was labeled with ^{32}P at its 5′ end by kinase treatment and annealed to mouse liver poly $(A)^+$ RNA at a molar ratio of approximately 1:1 [oligonucleotide/poly $(A)^+$ RNA] by heating to 95°C for 5 min and slow cooling to 42°C in annealing buffer (50 mM Tris·HCl, pH 8.3/60 mM NaCl/10 mM dithiothreitol/1 mM EDTA). The hybridized oligonucleotide was extended by using reverse transcriptase. The full-length cDNA extension product was purified by electrophoresis on a 7 M urea/8% polyacrylamide gel, and was sequenced by chemical degradation by taking advantage of the uniquely labeled 5′ end.

Sequences were analyzed and compared using the computer program described by Queen and Korn (20). Predictions of potential secondary structure were performed by Chou and Fasman analysis (21), and diagonal matrix analysis was performed by using the DIAGON program described by Staden (22).

RESULTS AND DISCUSSION

Mouse Apo E Sequence. Approximately 1400 cDNA clones, made from mouse liver mRNA, and inserted into a pBR322 vector were screened by hybrid selection and translation (16), and clone p2C1-apo E was selected as being almost full length and was sequenced. The entire cDNA insert contains 968 nucleotides plus 21 adenine residues from the poly(A)-RNA tail (Fig. 1). There is only one large open reading frame consisting of 856 nucleotides, and it is followed by an untranslated region of 112 nucleotides before the poly(A)-RNA. The polyadenylylation signal AATAAA (23) is located 14 bases upstream of the polyadenylylation site. By analogy to the rat sequence (7), one concludes that the p2C1-apo E would code for the entire mature mouse apo E protein with the exception of 8 amino acids at the amino terminus. The coding sequence of apo E upstream from the cDNA clone (Fig. 1) was determined by sequencing of mouse liver apo E mRNA. This was done by a method in which a 5′-labeled-synthetic 14-base oligonucleotide primer complementary to apo E mRNA was extended by using reverse transcriptase to the end of the mRNA. The extended oligonucleotide was then purified by gel electrophoresis and sequenced by chemical degradation.

The complete amino acid sequence of mature mouse apo E, derived from the cDNA and mRNA sequencing, is shown in Fig. 2. The predicted polypeptide has a charge and size that compares favorably with values for mature mouse apo E. The

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.

Abbreviation: apo E, apolipoprotein E.

*To whom reprint requests and correspondence should be addressed at: Department of Medicine, University of California, Los Angeles, CA 90024.

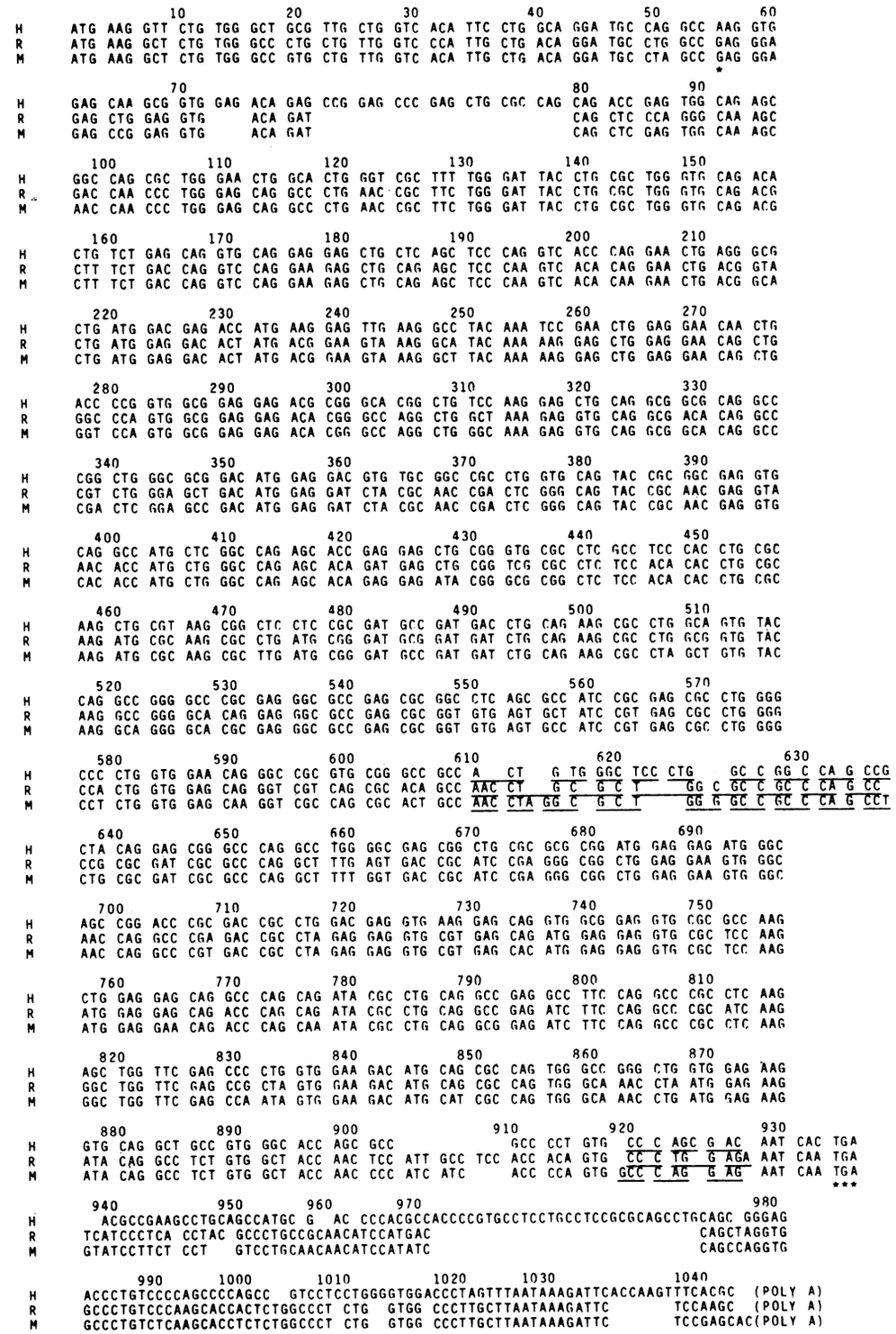

FIG. 1. Alignment of human, rat, and mouse apo E cDNA sequences. The alignment was made so as to demonstrate maximum homology. In the coding region, the nucleotides are grouped in sets of three corresponding to the codons. Underlined codons are those which are read differently in the three species. In the regions of frameshifting other alignments are possible which reduce the number of deletions/insertions but which reduce homology. Gaps correspond to nucleotides present in one species but absent in another. The beginning of the mature protein is indicated by an asterisk.

mature protein is preceded by an 18-amino acid signal peptide with the sequence Met-Lys-Ala-Leu-Trp-Ala-Val-Leu-Leu-Val-Thr-Leu-Leu-Thr-Gly-Cys-Leu-Ala. This sequence is highly homologous to the signal peptides of rat and human apo E (6, 7).

Species Comparisons. The nucleotide sequence of mouse apo E was aligned with the human (6) and rat (7) cDNA sequences (Fig. 1). The overall homology between mouse and human is 78%, between mouse and rat is 93%, and between rat and human is 78%. To maximally align the three sequences it was necessary to introduce gaps for nucleotides present in one species that were absent in the others (Fig. 1). These occur predominantly in three regions (nucleotides 73–78, 611–633, and 904–927). Some of the changes in these

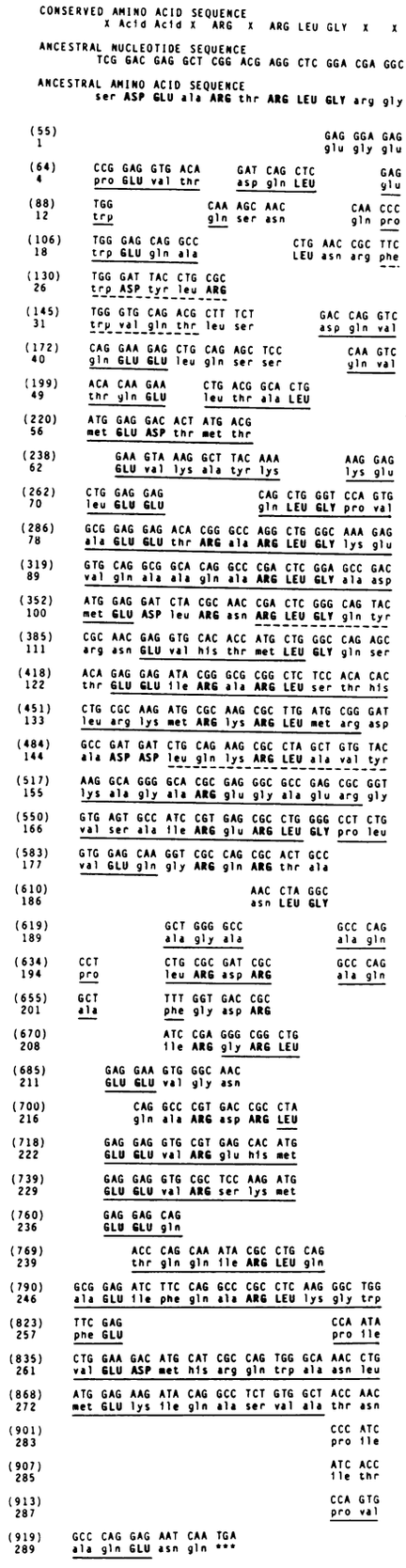

FIG. 2. Nucleotide and amino acid sequence of mature mouse apo E: Alignment of internal repeats with proposed ancestral sequence. Shown at the top are the positions of highly conserved amino acids of apo E along with the proposed ancestral nucleotide and amino acid sequences. The apo E nucleotide and amino acid sequences have been aligned with the central 33-nucleotide repeats. Amino acids which correspond to the conserved sequence are highlighted. The predicted secondary structure of mouse apo E is indicated by solid underlining (α-helix) or broken underlining (β-sheet). Amino acids are numbered starting with the first residue of the mature protein. Nucleic acids are numbered (in parentheses) as in Fig. 1.

regions have occurred since the divergence of mouse and rat, implying both a recent and a rapid evolutionary change. The distance between the polyadenylylation signal and the actual site of polyadenylylation is different in rat, human, and mouse sequences, adding circumstantial evidence that the site of polyadenylylation is determined by secondary structural features in the precursor mRNA in addition to proximity to the signal sequence (24).

In terms of the amino acid sequence of apo E, mouse is 91% homologous to rat and 70% homologous to human. All three species contain an 11-amino acid repeat in the central region of the protein with the conserved amino acid sequence: Xaa-(Glu/Asp)-(Glu/Asp)-Xaa-Arg-Xaa-Arg-Leu-Gly-Xaa-Xaa (see Fig. 2 for mouse).

The sequences of the mature apo E proteins from human, rat, and mouse were analyzed to predict secondary structure using the rules of Chou and Fasman (21). Fig. 2 shows the predicted secondary structure of mouse apo E. Overall, the predicted structures of the three proteins are nearly identical with α-helical regions comprising two-thirds of the protein in 14 areas and β-sheet comprising $\approx$10% of the protein in three areas. The helices are predominantly amphipathic in nature (data not shown). The receptor binding site of human apo E is thought to be localized around human amino acid residues 140–150 (25–27). This corresponds to mouse residues 132–142, since human apo E has an additional 8 amino acids near the amino terminus (Fig. 1). The potential structure of this region is predicted to be an amphipathic α-helix for all three species, and the only differences are three conservative substitutions of hydrophobic amino acids.

Allelic variants have been identified for human apo E that differ in at least six amino acid positions (25, 28, 29). At all of these positions both the rat (7) and mouse have the same amino acid residue as the human E4 allele. Thus, it is likely that the E4 is the primal human allele, and not the E3 as has been suggested (29).

Internally Repetitive Regions. Several reports indicate that there is considerable internal homology within individual apolipoproteins as well as between the different apolipoproteins (8–12). An analysis, therefore, was undertaken to characterize the internal homologies found in the mouse apo E gene. Because of the extensive correlation with both human and rat sequences (Fig. 1) the mouse results also pertain to these species.

Searches of both the amino acid and nucleotide sequences for internal homologies yielded extensive data sets that indicated almost any region of the gene showed homology to one or more regions elsewhere in the gene. This can be depicted most clearly by the dot matrix diagram (22) shown in Fig. 3. In this analysis, all segments of a given length in the sequence of apo E are compared to all other segments of similar length, and the position of the middle of the two segments is plotted along the two axes. A positive score is given and a point is plotted only if the two segments are sufficiently homologous. Homologous sequences show up as diagonal lines that are offset from the central line of identity by a distance equal to the separation between the homologous regions. The most prominent feature of the dot matrix analysis for mouse apo E is the region between nucleotides 200 and 600, which indicates a repeated structure 33-base-pairs long shown by the large arrows. This region would contain 12 such unit repeats. Each of these shows good homology to neighboring repeats (the closed arrows) but those which are widely separated no longer show up as being

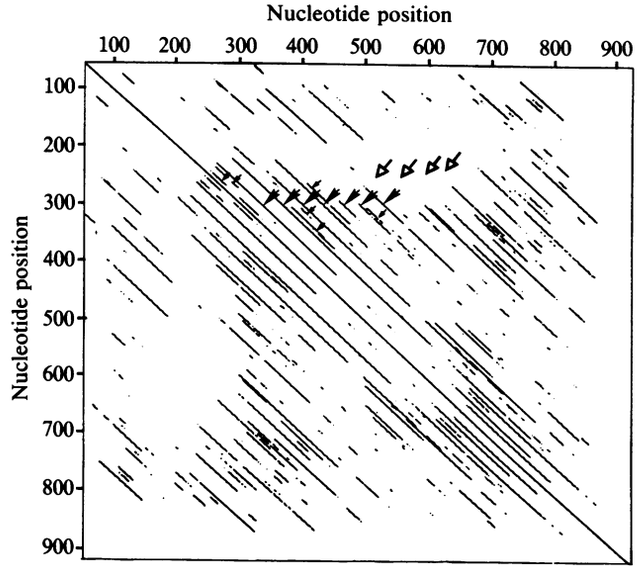

FIG. 3. Diagonal dot matrix analysis of mouse apo E nucleotide sequence. The analysis was performed using the DIAGON portion of the Staden program (22). The nucleotide sequence was compared against itself in all possible alignments. Along each alignment, all stretches of 131 nucleotides were examined and scored as positive (i.e., a dot was placed at the location of the nucleotide in the center of the 131-nucleotide long sequence examined) if 51 out of 131 were identical. These parameters were chosen empirically because they clearly demonstrate the major 33-nucleotide repeats. Closed arrows, good homology to neighboring sequences; open arrows, homologous but widely separated sequences; small arrows, 11-nucleotide repeats.

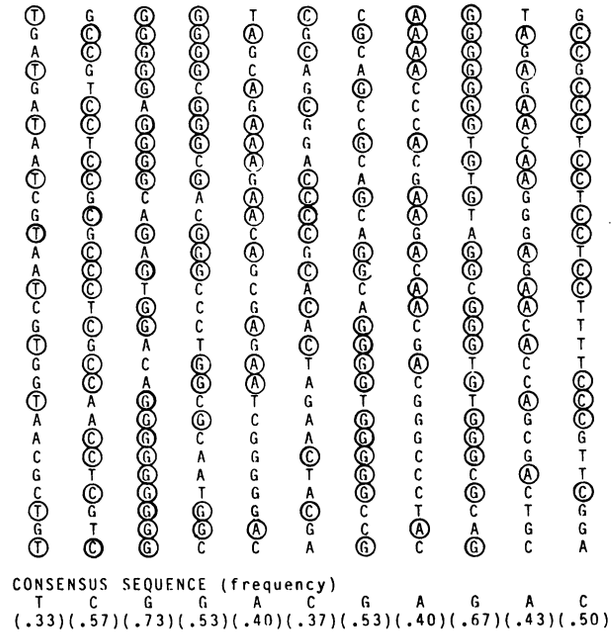

FIG. 4. Apo E is composed of 11-nucleotide repeats. The sequence of the central region of mouse apo E showing tandem replication of an 11-nucleotide repeat is shown. The sequence extends continuously from nucleotides 275–604, and, thus, each consecutive 11-nucleotide repeat is translated in a different reading frame. The consensus sequence and the frequency of occurrence of the nucleotide is given below. The nucleotides corresponding to this consensus are circled. Other regions of mouse apo E are also homologous to this 11-nucleotide repeat (see Fig. 2), but because in some cases only portions of the sequence have been duplicated, their alignment is less certain. This consensus sequence differs from the proposed ancestral unit (see text) at only the tenth position, where the choice between guanine and adenine is uncertain. We tentatively have inserted a guanine at this position on the basis of inspection of the entire apo E coding region, but either nucleotide yields the same sequence of conserved amino acids.

homologous (the open arrows) under the criteria used. There are also diagonal patterns indicative of a more primitive repeat of 11 base pairs (the small arrows).

Data from amino acid and nucleotide homology searches were compiled to yield a composite of the repeated structure within mouse apo E. Since the dot matrix analysis revealed the general 33-base-pair repeat structure, those homologies found by the Queen and Korn homology searches (20) that were approximately 33 bases apart were aligned. Within these repeat units, individual nucleotides were aligned by inspection to yield maximal homology. A similar alignment of homologies outside the region where the 33-base-pair unit is predominant was also done, and these were then compared to the block of 33-base-pair repeats. Homology searches also revealed evidence for repeats of 11 nucleotides within the 33-nucleotide repeats (Fig. 4). These correspond to the 11-nucleotide lines of homology depicted by the small arrows in the diagonal matrix (Fig. 3). Examination of the corresponding amino acid sequences indicated that the 11-nucleotide repeats are read in all three reading frames. We have chosen to depict the repetitive apo E structure as shown in Fig. 2, aligning the nucleotides in triplet codons to illustrate the resulting amino acid homologies as well as the resemblances to the proposed apo E ancestral sequence (see below). It should be noted that this representation maximizes neither amino acid nor nucleic acid homologies, and, obviously, other alignments such as a 66-nucleotide repeat would also show homologies.

An 11-Nucleotide Ancestral Sequence. The concept that genes have evolved by duplications of primitive nucleotide sequences has been documented (30). In particular, repetition of nucleotide sequences that are not integral multiples of three in length and that contain no termination codons in any of the three reading frames would lead to (*i*) long open reading frames, (*ii*) a repetitious nature of the amino acid sequence and hence also of structural features such as α-helices, and (*iii*) resistance to mutational silencing since all three reading frames would be open and would encode similar protein sequences.

Since the above analysis of apo E internal homology suggested the presence of an 11-nucleotide repeat length, we attempted to derive the sequence of an 11-nucleotide ancestral unit which could give rise to the present day apo E by tandem duplication and mutation. The repeating unit -TCGGACGAGGC- was found to match the highly conserved apo E amino acid sequence of Xaa-(Glu/Asp)-(Glu/Asp)-Xaa-Arg-Xaa-Arg-Leu-Gly-Xaa-Xaa and to satisfy the overall nucleotide sequence homology requirements (Fig. 2). It is essentially identical to the consensus 11-nucleotide repeat sequence present in the continuous stretch of 33-nucleotide repeats of the central region of apo E (Fig. 4). Also, its base composition (C, 27%; T, 9%; G, 45%; A, 18%) is not unlike that of present-day mouse apo E (C, 28%; T, 14%; G, 35%; A, 23%). The 11-nucleotide repeat can be read from any of the three reading frames and from any starting position to yield the same repeating amino acid sequence.

The peptide structure encoded by the tandemly repeated 11-nucleotide unit (each successive unit being read in a different reading frame) is an 11-amino acid repeating sequence consisting of a potential α-helical stretch (-Asp-Glu-Ala-Arg-Thr-Arg-Leu-) interspersed with a region (-Gly-Arg-Gly-Ser-) that would break the α-helical structure. This α-helix would be amphipathic in nature and would exhibit

structural features thought to be important in binding to hydrophobic surfaces containing phospholipids (i.e., hydrophobic and hydrophilic faces separated by positively charged residues). Once a particular reading frame was established by an initiator codon, the repeat would become fixed as a 33-nucleotide sequence. Also, two such helices could be stacked together by mutationally changing the four amino acids of the spacer into residues compatible with α-helical structure, thereby resulting in a α-helix 18 amino acids in length and conversion of the repeat structure into a 66-nucleotide sequence. Such a conversion of a 4-amino acid spacer in an α-helical region would permit nearly continuous alignment of the corresponding faces of the previously separate amphipathic α-helices. Indeed, such 66-nucleotide repeats are a major feature of some apolipoproteins investigated thus far (31).

The central region of apo E, from amino acids 73–185, appears to have arisen by repeated duplication of a sequence derived from three tandem repeats of the ancestral 11-nucleotide sequence (Fig. 2). This 33-nucleotide sequence probably mutated prior to repeated duplication, because at several positions the predominant nucleotide observed differs from that present in the predicted ancestral sequence. Mutations occurring early in the duplication process would also account for the fact that adjacent repeats of the apo E sequence are more similar than widely separated repeats. Outside of the central region of apo E, shorter repeats of the sequence have been duplicated (Fig. 2), while the signal sequence has no obvious homology to the region encoding the mature protein.

Analysis of amino acid sequence data indicates that the apolipoproteins constitute a family of structurally and functionally related genes which have arisen from a common ancestral sequence. Thus, it is likely that the ancestral 11-nucleotide unit proposed here or a closely related sequence gave rise by duplications to a primitive gene which evolved into the various apolipoproteins comprising this family. Indeed, the repetitive structure of other apolipoproteins, including apo A-I and apo A-IV, appear to be entirely consistent with this possibility. Analysis of the nucleotide sequence data for apo A-I (11) and apo A-IV (12, 31) indicates that the 33- and 66-nucleotide repeats identified contain more primitive 11-nucleotide repeats which are similar to the ancestral sequence for apo E proposed here (data not shown).

In conclusion, the conserved amino acid repeats of mammalian apo E have retained a clear vestige of their 11-nucleotide ancestral unit, providing a striking example of a gene built by duplication of oligomeric repeats. These data and our hypothesis lend strong support to the work of Ohno and others who have postulated a similar mechanism for the origin of class 1 major histocompatibility antigens, immunoglobulins, and other proteins (30).

We are grateful for the technical assistance of Stuart Rich and Diana Quon. We also thank Dr. Susumo Ohno for invaluable discussions. This work was supported in part by National Institutes of Health Grants HL 27481 and AM 27008. This work was carried out in facilities provided by the Jonsson Comprehensive Cancer Center and the Molecular Biology Institute, University of California, Los Angeles. K.L.R. was a United States Public Health Service Trainee supported by National Research Service Award GM 07104.

1. Shore, V. G. & Shore, B. (1973) *Biochemistry* **12,** 502–507.
2. Innerarity, T. L. & Mahley, R. W. (1978) *Biochemistry* **17,** 1440–1447.
3. Mahley, R. W. & Innerarity, T. L. (1983) *Biochim. Biophys. Acta* **737,** 197–222.
4. Rall, S. C., Weisgraber, K. H. & Mahley, R. W. (1982) *J. Biol. Chem.* **257,** 4171–4178.
5. Taylor, J. M. (1984) *J. Biol. Chem.* **259,** 6498–6504.
6. Breslow, J. L., McPherson, J., Nussbaum, A. L., Williams, H. W., Lofquist-Kahl, F., Karathanasis, S. K. & Zannis, V. I. (1982) *J. Biol. Chem.* **257,** 14639–14641.
7. McLean, J. W., Fukazawa, C. & Taylor, J. M. (1983) *J. Biol. Chem.* **258,** 8993–9000.
8. Barker, W. C. & Dayhoff, M. O. (1977) *Comp. Biochem. Physiol.* **B57,** 309–315.
9. Fitch, W. M. (1977) *Genetics* **86,** 623–644.
10. McLachlan, A. D. (1977) *Nature (London)* **267,** 465–466.
11. Karathanasis, S. K., Zannis, V. I. & Breslow, J. L. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 6147–6151.
12. Boguski, M. S., Elshourbagy, N., Taylor, J. M. & Gordon, J. I. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 5021–5025.
13. Segrest, J. P., Jackson, R. L., Morrisett, J. D. & Gotto, A. M. (1974) *FEBS Lett.* **38,** 247–253.
14. Sparrow, J. T., Morrisett, J. D., Pownall, J. H., Jackson, R. L. & Gotto, A. M. (1975) in *Peptides: Chemistry, Structure and Biology*, eds. Walter, R. & Meienhofer, J. (Ann Arbor Science, Ann Arbor, MI), pp. 597–602.
15. Lusis, A. J. & LeBoeuf, R. C. (1985) *Methods Enzymol.*, in press.
16. Reue, K. L., Quon, D. H., O'Donnell, K. A., Dizikes, G. J., Fareed, G. C. & Lusis, A. J. (1984) *J. Biol. Chem.* **259,** 2100–2107.
17. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
18. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 560–564.
19. Matteucci, M. D. & Caruthers, M. H. (1980) *Tetrahedron Lett.* **21,** 719–722.
20. Queen, C. & Korn, L. J. (1980) *Methods Enzymol.* **65,** 595–609.
21. Chou, P. Y. & Fasman, G. D. (1978) *Annu. Rev. Biochem.* **47,** 251–276.
22. Staden, R. (1982) *Nucleic Acids Res.* **10,** 2951–2961.
23. Proudfoot, N. J. & Brownlee, G. G. (1976) *Nature (London)* **263,** 211–214.
24. McDevitt, M. A., Imperiale, M. J., Ali, H. & Nevins, J. R. (1984) *Cell* **37,** 993–999.
25. Weisgraber, K. H., Rall, S. C. & Mahley, R. W. (1981) *J. Biol. Chem.* **256,** 9077–9083.
26. Innerarity, T. L., Friedlander, E. J., Rall, S. C., Weisgraber, K. H. & Mahley, R. W. (1983) *J. Biol. Chem.* **258,** 12341–12347.
27. 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.
28. Weisgraber, K. H., Innerarity, T. L. & Mahley, R. W. (1982) *J. Biol. Chem.* **257,** 2518–2521.
29. Rall, S. C., Weisgraber, K. H., Innerarity, T. L. & Mahley, R. W. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 4696–4700.
30. Ohno, S. (1984) *J. Mol. Evol.* **20,** 313–321.
31. Boguski, M. S., Elshourbagy, N. E., Taylor, J. M. & Gordon, J. I. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 992–996.