Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 992–996, February 1985
Biochemistry

Comparative analysis of repeated sequences in rat apolipoproteins A-I, A-IV, and E

(full-length cDNA cloning/deletion mutations/high density lipoprotein structure/gene family structure and evolution)

MARK S. BOGUSKI*, NABIL ELSHOURBAGY†, JOHN M. TAYLOR†, AND JEFFREY I. GORDON*‡

Departments of *Biological Chemistry and ‡Medicine, Washington University School of Medicine, St. Louis, MO 63110; and †Gladstone Foundation Laboratories, Department of Physiology, Cardiovascular Research Institute, University of California, San Francisco, CA 94140

Communicated by Oliver H. Lowry, October 9, 1984

ABSTRACT **To understand the structural, functional, and evolutionary relationships among the principal protein components of rat high density lipoprotein particles, we undertook a systematic comparative analysis of the primary structures of apolipoproteins (apo)-A-I, -A-IV, and -E. Human apo-A-I and rat apo-A-IV have been shown previously to contain repeated sequences that presumably arose by intragenic duplication of 11- or 22-amino acid amphipathic segments. For apo-A-I, these segments are thought to be the structures responsible for lipid binding and activation of lecithin:cholesterol acyltransferase. From an analysis of the sequence of a full-length cDNA clone, rat apo-A-I is shown to contain eight tandem repetitions of a 22-amino acid segment. However, compared with human apo-A-I, the rat protein has undergone three deletions, two of which involve multiple amino acids in the repeated sequence domain. This disruption of the periodic structure of the protein raises the possibility of species-specific variation in the ability of rat apo-A-I to interact with high density lipoproteins and activate lecithin:cholesterol acyltransferase. Statistical analysis of the structure and organization of repeated sequences in apo-A-I, -A-IV, and -E demonstrates that all three proteins are paralogous members of a dispersed gene family. Despite overall similarity in sequence organization, different portions of these sequences have evolved at different rates. Diversification of a duplicated ancestral sequence has resulted in three lipid-binding proteins with distinct and shared functions.**

High density lipoproteins (HDL) have been implicated in the pathogenesis of atherosclerosis by virtue of the inverse relationship between their concentration in plasma and the incidence of coronary artery disease (1). The protein components of mammalian HDL are a heterogeneous group of macromolecules with diverse, yet overlapping, functions. The primary structures of many of the apolipoproteins are known (2). Still, in some cases, precise metabolic roles have yet to be defined. Also, species-specific differences in lipoprotein particle composition remain unexplained. The protein composition of rat HDL is considerably different from that of humans. The major protein components of human HDL are apolipoprotein (apo)-A-I and -A-II (1), whereas apo-A-I, -A-IV, and -E constitute the principal proteins of rat HDL (3).

Comparative sequence analysis is an important tool for studying the structural, functional, and evolutionary relationships among proteins. Often, the first or only clue to the biological significance of a newly discovered sequence is similarity to another sequence of known function. Detailed comparisons can provide evidence for structure–function relationships and address the question of common ancestry versus convergent evolution. Such analyses led to the discovery of intragenic duplication in human apo-A-I (4–6) and rat apo-A-IV (7) and suggested possible functions for these repeated sequences.

In the present study, we examine structural relationships among the three principal protein components of rat HDL particles: apo-A-I, -A-IV, and -E. The complete nucleotide and amino acid sequences of apo-E and -A-IV had been determined previously (7, 8). The isolation and sequencing of a full-length rat apo-A-I cDNA clone are described in the present work. Although the primary structure of rat apo-A-I is very similar to that of human apo-A-I (9), the rat sequence has undergone multiple deletions that may affect its physiologic function. Quantitative sequence comparisons show that apo-A-I, -A-IV, and -E have evolved from a common ancestral gene, with apo-A-I and -A-IV having diverged more recently. These results are discussed in terms of the distinct and shared functions of the apolipoproteins.

MATERIALS AND METHODS

The construction and screening of rat intestinal cDNA libraries and the sequencing of cDNA clones has been described previously (7). Plasmids containing apo-A-I cDNA were identified by hybridization selection and translation of the mRNA, followed by immunoprecipitation with specific antibodies. Computer programs (10) were used to conduct searches of biological sequence databases as described in the text. Optimal alignments of nucleic acid and protein sequences were carried out as described (10). Protein comparison matrices and associated statistics were calculated according to McLachlan (11, 12).

RESULTS AND DISCUSSION

Rat apo-A-I cDNA Cloning and Sequence Analysis. An 889-base-pair apo-A-I clone was isolated from a rat intestine cDNA library and its nucleotide sequence was determined (Fig. 1). Analysis of this sequence revealed one long open reading frame of 777 nucleotides, with 5′ and 3′ nontranslated regions of 43 and 69 nucleotides, respectively. The first 24 residues of the derived amino acid sequence agree with the primary structures of the signal peptide and prosegment that had been determined previously for the apo-A-I precursor (13). The amino-terminal sequence of the mature plasma protein also agrees with data published previously (13). The mature plasma form of rat apo-A-I contains 235 residues and has a calculated molecular weight of 27,369.

Poncin *et al.* have recently described a cDNA sequence for rat apo-A-I (14). Despite the use of the same strain of rat (Sprague–Dawley), our sequence differs from theirs in eight positions, seven of which occur in the coding region. In all cases, these discrepancies are substitutions of adenine for guanine or vice versa. Three of these differences result in

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; HDL, high density lipoprotein(s); LCAT, lecithin:cholesterol acyltransferase.

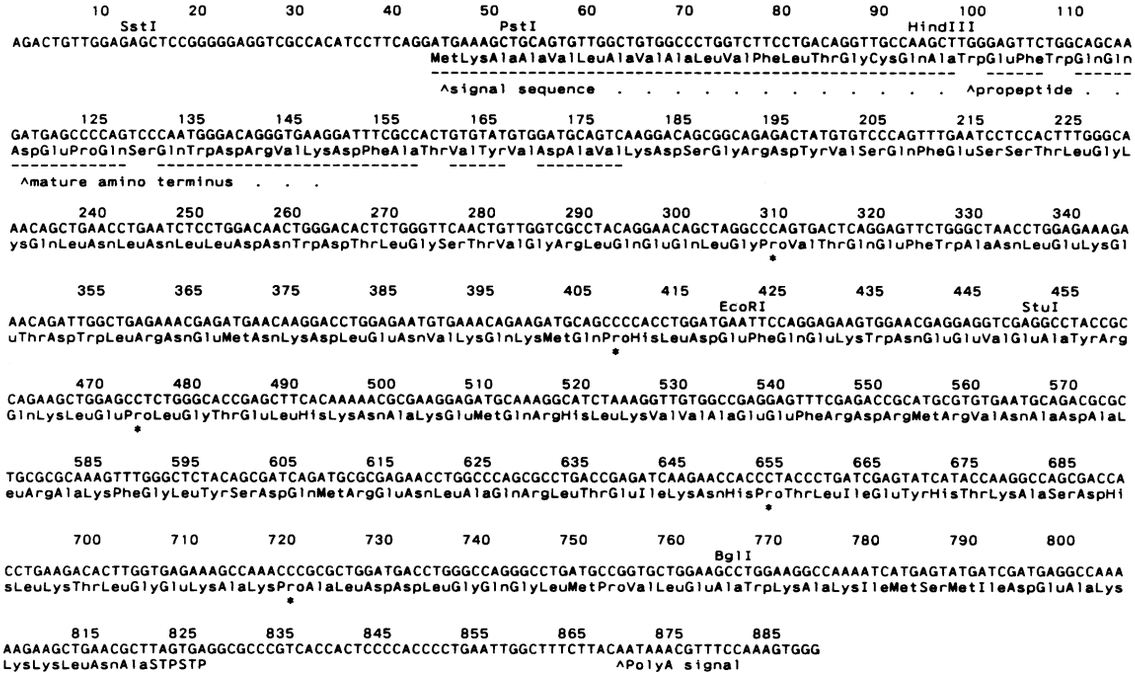

FIG. 1. Nucleotide sequence of rat apo-A-I cDNA. DNA sequencing was carried out on restriction fragments of the full-length cDNA that had been subcloned in phage M13, as described in ref. 7. Both strands were sequenced and junctions were confirmed by sequencing across the restriction sites indicated. Portions of the amino acid sequence determined previously (13) are underlined. Proline residues in the core repeated sequence block are marked by asterisks. Two in-phase termination codons are indicated by STP.

amino acid changes in the derived protein sequence (Fig. 3). All regions of our cDNA clone were sequenced on both strands, some regions were sequenced up to five times, and extensive regions of the sequence were confirmed in two additional isolates from a separate cDNA library.

Comparison of Orthologous Sequences: Rat and Human apo-A-I. Strictly speaking, homologous sequences are related by descent from a common ancestor. Homologous sequences can be divided into *orthologous* and *paralogous* categories (15). Orthologous sequences reflect the phylogenetic branching order of the species in which they are found and have identical functions. Paralogous sequences are the product of a gene duplication that was fixed prior to speciation. Divergence of such sequences usually results in the evolution of new functions. In this section we compare orthologous apolipoproteins A-I and consider the functional implications of sequence differences. In the next section we demonstrate that apo-A-I, -A-IV, and -E are paralogous proteins and discuss the possible origins of their unique functions.

Rat apo-A-I is eight amino acids shorter than its human homolog (Fig. 2). A proline residue at position 28 in the human sequence is deleted in the rat protein. The carboxyl-terminal amino acid in human A-I is a glutamine, specified by the codon CAG. At the corresponding location in the rat sequence (nucleotide 821), a substitution of thymine for cytosine results in a nonsense mutation causing translation to terminate one residue prematurely relative to human A-I mRNA (Figs. 1 and 2).

The remaining six amino acids absent from the rat sequence can be accounted for by two deletions in a region of the protein that, for human apo-A-I, has been shown to consist of repeat units of 11 or 22 amino acids (4–6). As shown in Fig. 3, rat apo-A-I contains eight 22-amino acid segments (or sixteen 11-amino acid segments), which constitute a repeated sequence domain. The tandemly arrayed docosapeptides in apo-A-I are not exact duplications, but amino acid substitutions have generally conserved the chemical types of residues in corresponding positions of the repeats. It has been shown that, for human apo-A-I, these docosapeptides are able to assume amphipathic (also termed amphiphilic) helical conformations that are thought to be the structural basis of the protein's activities in binding lipids and activating lecithin:cholesterol acyltransferase (LCAT) (16). Deletions and substitutions of single residues in human apo-A-I variants have been reported to decrease the ability of the protein to activate LCAT *in vitro* (17). Thus the deletions in rat apo-A-I may be of considerable functional significance.

The precise locations of these deletions, and the number of residues involved, were determined by optimal alignment (10) of the amino acid sequences of rat and human apo-A-I (Fig. 2) as well as by optimal alignment of the corresponding nucleotide sequences (not shown). These deletions result in gaps of four and two residues relative to the human apo-A-I sequence. The first deletion occurs in the central region of

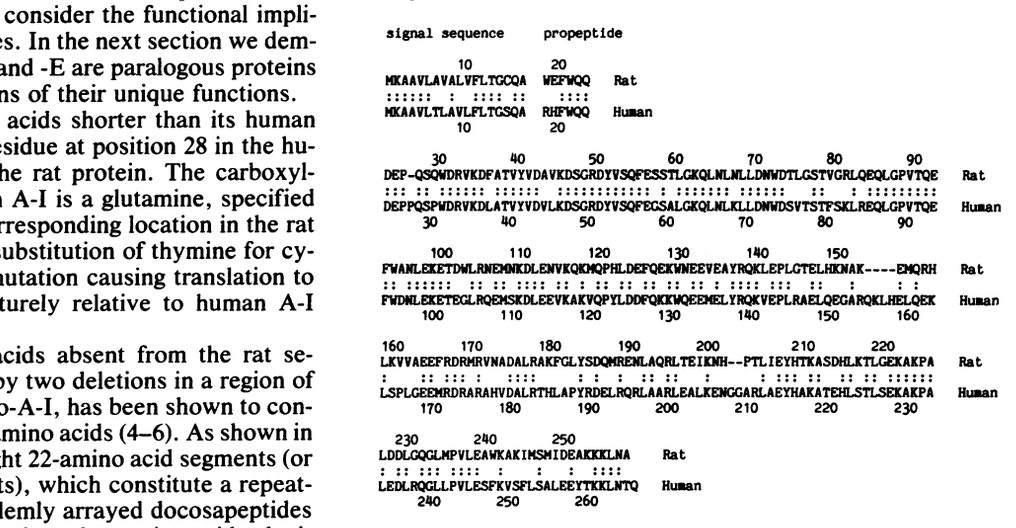

FIG. 2. Optimal alignment of rat and human apo-A-I. Sequences were aligned by computer with a *k*-tuple of 1, a window size of 20, and a gap penalty of 2 as described in ref. 10. The standard one-letter code for amino acids is used, and dots indicate identities.

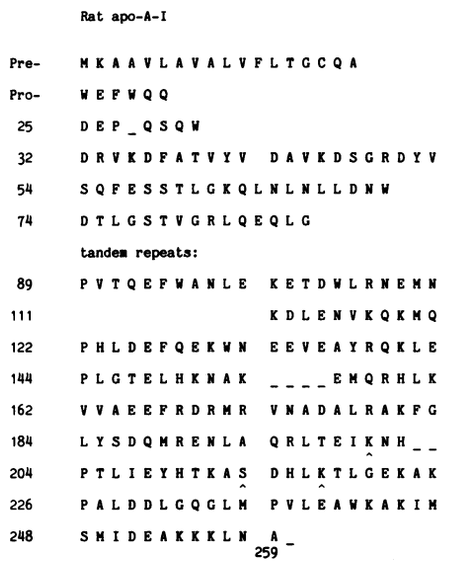

FIG. 3. Primary structure of rat apo-A-I arranged to emphasize the organization of repeated sequences. Gaps in the sequence, relative to human apo-A-I, are indicated by an underscore. Discrepancies with the sequence of Poncin *et al.* (14) are indicated by carets.

the third docosapeptide repeat (Fig. 3, after residue 154). Assuming that the docosapeptides in apo-A-I have α-helical conformations (16), the consequence of a four-residue deletion would be to eliminate about one turn (3.6 residues) of the helix. The definition of an amphipathic helix (18) specifies the distribution of hydrophilic and hydrophobic side chains on opposite faces of the helix. Deletion of one helical turn would not drastically affect the positions of side chains relative to a particular face of the helix and would thus be expected to preserve the general amphipathic character.

The second deletion of two glycine residues occurs at what might be considered a "hinge" region between the fifth and sixth docosapeptide repeats (Fig. 3). According to rules for empirical prediction of secondary structure (19), glycine is a strong α-breaking, β-turn-forming residue. Thus the deletion of these two glycines from rat apo-A-I might be expected to considerably decrease the likelihood of a change in backbone direction between adjacent amphipathic helices. However, this deletion may be compensated to some extent by the presence of a helix-breaking proline in the next position (residue 204).

The two deletions in the core repeated sequence block of rat apo-A-I disrupt its periodic structure (see below) and may therefore affect functions such as LCAT activation. However, functional abnormalities in apo-A-I may be compensated by apo-A-IV. There is considerable sequence homology between these two proteins (7), including a shared repeated sequence block which, in apo-A-IV, has not undergone any deletions. It is noteworthy that apo-A-IV is a major component of rat HDL, with an abundance that could permit a functional complementarity to apo-A-I.

In an optimal alignment, human and rat preproapolipoproteins A-I are identical in 166 positions for an overall sequence homology of 64% (71% on the nucleotide level). Sequence identities, however, are not uniformly distributed (Fig. 3). Preproapo-A-I can be divided into three separate domains: signal sequence and propeptide (residues 1–24), amino-terminal domain (residues 25–88), and core repeated sequence block (residues 89–259). These domains roughly correspond to exons in the human apo-A-I gene (20, 21). Sequence conservation in the signal sequence and propeptide has already been noted (21, 22). The 64-residue amino-terminal domain of mature rat apo-A-I is the most highly conserved region of the molecule, with 77% identity to the human protein. In contrast, the core repeated sequence block is only 64% homologous. (The majority of nonidentical residues are conservative substitutions in both regions.) Clearly, natural selection has tolerated fewer mutations in the amino-terminal domain.

Comparison of Paralogous Sequences: Rat apo-A-I, -A-IV, and -E. A common evolutionary origin for the apolipoproteins has long been suspected on the basis of an 11-residue sequence periodicity (6). Human apo-A-I contains six to eight repetitions of a 22-amino acid segment that was itself composed of two related 11-mers (4–6). Similar 11-residue segments were noted in apo-A-II, -C-II, and -C-III, but these repeats were considerably more subtle, requiring statistical tests to rule out chance occurrence (6). However, on the basis of a different statistical approach (5), McLachlan suggested that similarities between apo-A-I and -C-III could be the result of convergent evolution rather than divergence from a common ancestral gene.

The rat apo-A-I sequence now permits a detailed comparative analysis of relationships among the principal rat HDL-associated apolipoproteins, A-I, A-IV, and E. As a first step in analyzing these relationships, database searches were conducted, using rat apo-A-IV as the query sequence. We have used apo-A-IV as the touchstone for analysis of repeated sequences because it has the greatest number of most highly conserved repeats of any known apolipoprotein (7). Among the 1175 mammalian nucleotide sequences in the Genetic Sequence Data Bank[§], the three best matches were to two human apo-A-I entries (J00098, J00100) and the human apo-E sequence (K00396). The comparison scores for these three matches were 12, 8.8, and 7.3 standard deviations above the mean score for all mammalian sequences. A search of the Protein Sequence Database[¶] yielded the following top three matches: human proapo-A-I (LPHUA1), human apo-E (LPHUE), and dog apo-A-I (LPDGA), with comparison scores that were 4.2, 3.7, and 3.5 standard deviations above the mean score for 2511 sequences. It is notable that apo-A-II, -C-I, -C-II, and -C-III were *not* among the 40 sequences with highest similarity scores to apo-A-IV, despite their presence in the Protein Sequence Database. This may indicate either distant evolutionary relationships or structural convergence.

The complete nucleotide sequences coding for rat apo-A-IV, -E, and -A-I are known (refs. 7 and 8 and the present work). However, the analyses that follow employ the derived protein sequences because it has been demonstrated that DNA comparisons are less informative and subject to more background "noise" (23). Indeed, Karathanasis *et al.* concluded that there are only six repeats in apo-A-I, on the basis of a computer analysis of DNA sequence data alone (20). However, inspection of the rat protein sequence (Fig. 3), as well as the sequence of human apo-A-I, shows that there are at least eight repeat units.

The comparison matrix method (11, 12) was used to quantitatively analyze the relationships among rat apolipoproteins A-I, A-IV, and E. This method computes matching probability scores for spans of amino acids based upon observed frequencies of amino acid replacements in homologous proteins. One sequence is compared against another and a matrix of scores representing all possible alignments of

[§]Bolt Beranek and Newman Inc., Cambridge, MA. Sequence entries are referenced by their accession numbers. Release 19.0 of the data bank was searched with a *k*-tuple of 4, window of 20, and gap penalty of 4 (10).

[¶]National Biomedical Research Foundation, Washington, DC. Sequence entries are referenced by their retrieval key codes. The January 1984 release was searched with a *k*-tuple of 2, window of 20, and gap penalty of 10.

the two sequences is constructed. All scores exceeding a threshold value are plotted to obtain a graphical representation of the locations of related sequences within the two proteins. The probability of chance occurrence of a particular score is estimated from the theoretical distribution of matching scores for infinite random sequences with the same amino acid compositions. In proteins that are closely related but lack a repeating structure, the highest scores tend to fall on a single main diagonal. Repeated sequences reveal themselves as shorter diagonals offset from the main diagonal.

Fig. 4*A* shows the comparison matrix for preapo-A-IV and preproapo-A-I. Immediately apparent are the strong repeats extending from residue 89 to about residue 350 in apo-A-IV and from residue 89 to about residue 200 in apo-A-I. Most of the diagonals correspond to repeat spacings that are multiples of 11 or 22 residues. Starting from about residue 200 in apo-A-I, there is a blank vertical zone indicating a major discontinuity in the periodic structure of the molecule. It is important to note that such zones do not indicate deletions per se but rather an absence of spans with comparison scores as high as elsewhere in the sequence. This phenomenon may be the result of a shift in the register of a periodic pattern (caused by a deletion or insertion) or the presence of a span of divergent amino acids. A register shift caused by the two-residue gap in the fifth docosapeptide repeat is part of the explanation in this case, but sequence divergence is also an important factor. The nonconservative substitution of isoleucine (residue 207) in an "acidic" region (7) of the repeat units is notable. The four-residue deletion after residue 154 in apo-A-I has less drastic consequences, causing only a slight shift off the main diagonal (arrow b in Fig. 4*A*).

Another striking feature of the comparison matrix in Fig. 4*A* is the long main diagonal shared by the two sequences, indicating a colinear relationship between apo-A-I and apo-A-IV. The close correspondence between apo-A-IV residues 34–73 and apo-A-I residues 41–80 is most surprising, since this region spans the first two repeats of the apo-A-IV molecule. Close inspection of this domain revealed two related 11-mers (residues 32–42 and 43–53 in Fig. 3). These tandem repeats lie in that region of the apo-A-I molecule that is the most highly conserved domain in the rat and human proteins.

A comparison of preapo-A-IV vs. preapo-E (Fig. 4*B*) shows that the overall pattern of sequence organization in apo-E is remarkably similar to that in apo-A-I. There is a striking main diagonal indicating colinearity with apo-A-IV. The blank vertical zone indicates disruption of a periodic pattern of repeats in the apo-E molecule. Also, as in apo-A-I, the periodic pattern resumes after the interruption but extends farther in preapo-E because this protein is 311 residues long compared with 259 residues for preproapo-A-I.

In the amino-terminal domains, homology between apo-E and apo-A-IV is greater than for apo-A-IV and apo-A-I. For apo-A-IV vs. apo-A-I (Fig. 4*A*), the main diagonal fades away in three locations, whereas for apo-A-IV vs. apo-E (Fig. 4*B*) the corresponding diagonal is almost completely contiguous and contains a greater number of identical residues. Furthermore, two tandem repeats, parallel to the region indicated by the arrow in Fig. 4*B*, are readily apparent.

In the repeated sequence block, apo-E is less highly conserved than apo-A-I at the levels of both individual residues and the organization of repeated units. The wider blank vertical zone indicates more extensive sequence rearrangements. This divergence is particularly evident in the cumulative distributions of matching scores (Table 1). For example, the apo-A-I/A-IV matrix had 18 spans with scores of 157, whereas only 3 spans reached this score in the apo-E/A-IV comparison. Furthermore, the ratio of observed frequency to expected frequency is about 10-fold higher for apo-A-I/A-IV at this score level.

Our results demonstrate that apolipoproteins A-I, A-IV, and E contain sequences that derive from a common ancestral gene. This ancestral gene must have been the product of *intragenic* duplication of 33- and 66-nucleotide segments (4–7, 20). *Intergenic* duplication probably next gave rise to a primordial apo-E gene and an apo-A-I/A-IV precursor. Sub-

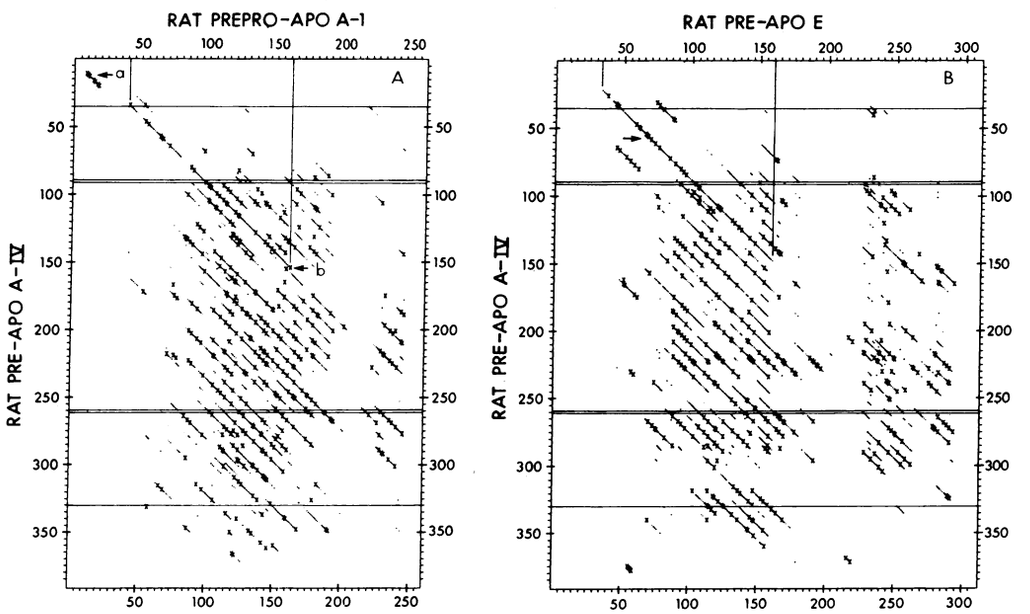

FIG. 4. Sequence comparison matrices with a span of 29 residues and weights of unity (5). A comparison or matching score is the unweighted sum of similarity scores across the span. X and · indicate the centers of spans with comparison scores of 125 or greater. In addition, X indicates an amino acid identity between the compared sequences. Single horizontal lines at residues 33 and 330 mark the repeated sequence domain in apo-A-IV. Double horizontal lines at residues 89 and 259 delineate the core repeated sequence block (7). Vertical lines indicate the main diagonals. (*A*) Rat preapo-A-IV vs. rat preproapo-A-I; 1801 spans with scores of 125 or greater (Table 1*A*) are displayed. In random sequences of the same amino acid compositions, only about 1 in 1000 (or 92 out of 101,269) spans should reach this score by chance. The signal peptide is indicated by the arrow labeled a. (*B*) Rat preapo-A-IV vs. rat preapo-E; 1955 spans with scores of 125 or greater (Table 1*B*) are displayed. The frequency of reaching this score by chance is about 1 in 1000 (133 out of 121,601).

Table 1. Comparison statistics for rat apo-A-IV vs. apo-A-I and apo-E

Score, *S*	No. spans of score *S* or higher	Observed frequency	Expected frequency	Observed/ expected
	A. Preproapo-A-I vs. preapo-A-IV			
86	50,317	0.499	0.639	0.78
100	18,710	0.185	0.166	1.11
110	7,943	0.784×10^{-1}	0.305×10^{-1}	2.50
125	1,801	0.178×10^{-1}	0.761×10^{-3}	23.4
132	705	0.696×10^{-2}	0.866×10^{-4}	80.0
143	119	0.117×10^{-2}	0.163×10^{-5}	718
157	18	0.178×10^{-3}	0.394×10^{-8}	45,200
168	2	0.197×10^{-4}	0.160×10^{-10}	1,230,000
	B. Preapo-E vs. preapo-A-IV			
88	60,788	0.499	0.630	0.79
100	27,039	0.222	0.206	1.10
110	11,028	0.910×10^{-1}	0.411×10^{-1}	2.21
125	1,955	0.161×10^{-1}	0.115×10^{-2}	14.0
136	438	0.360×10^{-2}	0.363×10^{-4}	99.2
146	58	0.477×10^{-3}	0.866×10^{-6}	551
157	3	0.246×10^{-4}	0.749×10^{-8}	3280

Comparison parameters were the same as in Fig. 4 and identical to those used to demonstrate the repeated sequences in human apo-A-I (5). The number of elements in a comparison matrix is the product of the number of residues in each of the two sequences being compared. Thus the preapo-A-IV vs. prepro-A-I matrix contained $391 \times 259 = 101{,}269$ spans. Similarly, the preapo-A-IV vs. preapo-E matrix contained 121,601 spans. The mean scores for shuffled sequences of the same length and amino acid composition were 86 for -A-IV vs. -A-I and 88 for -A-IV vs -E, with standard deviations of 15 and 14, respectively.

sequent evolution led to separate apo-A-I and A-IV genes. We have assumed that all three sequences diverged with comparable mutation rates and also that unequal crossovers (nonreciprocal recombinations) between interrupted genes are rarer genetic events than point mutations. That the divergence of apo-A-I and apo-A-IV was a more recent event is supported by two independent lines of evidence: rank order of sequence similarity in the nucleic acid and protein database searches and, more importantly, double matching probability score distributions (Table 1). This phylogeny dictates that, at least at some time in their evolutionary history, these genes or their progenitors were closely linked. In humans, apo-A-I and apo-E are on different chromosomes (24), indicating that the apolipoproteins constitute a dispersed gene family. The possible structural and evolutionary relationships of other apolipoproteins to the apo-A-I/A-IV/E family remain to be convincingly demonstrated.

Finally, a few observations can be made about the possible origins of those domains that confer unique properties on the individual apolipoproteins. Although the hexapeptide prosegment is unique to apo-A-I, the amino terminus of apo-A-IV contains a hexapeptide (Trp-Asp-Tyr-Phe-Thr-Gln, residues 32–37) similar to the apo-A-I prosegment (Trp-Glu-Phe-Trp-Gln-Gln). A functional propeptide may have arisen after divergence of the apo-A-I and apo-A-IV genes or may have been present in a primordial lipid-binding protein but subsequently lost from apo-A-IV and -E. It is quite clear from the comparison matrices (Fig. 4) that the amino-terminal domains of apo-A-I and -E were once composed of repeated sequences. However, the amino-terminal domain of apo-A-I has diverged considerably from the repeated sequence block. Most importantly, once these mutations occurred, selection acted to conserve the changes in this domain more highly than in other regions of the molecule (Fig. 2), which indicates an important, and perhaps yet undiscovered, function for this domain in apo-A-I and perhaps in apo-E as well. The rates of mutation and fixation have also been nonuniform within the apo-E sequence: the amino terminus is more highly conserved and the repeated sequence domain is less highly conserved, relative to apo-A-I. For human apo-E, the domain that interacts with the apo-B,E (low density lipoprotein) receptor has been mapped to a central region of the sequence that spans residues 140–160 (25). Note that the corresponding region in the orthologous rat protein (residues 150–170 in preapo-E) falls entirely within the repeated sequence domain. However, this region is precisely where the apo-E sequence begins to diverge most, relative to apo-A-IV (Fig. 4*B*). Thus it seems that many of the unique properties of the principal rat HDL-associated proteins have resulted from functional diversification of repeated sequences in the gene for a primitive lipid-binding protein. Given the expansionary manner in which these repeated sequences have propagated themselves, the existence of other genes (apo-B?) or pseudogenes based upon this repeated sequence motif is an intriguing possibility.

We thank Jens Birktoft, Mark Freeman, Larry Holden, and Richard Wrenn for assistance during various phases of this project. This work was supported in part by National Institutes of Health Grants AM 30292 and AM 31615. M.S.B. is supported by a Medical Scientist Training Program Grant GM 07200.

1. Herbert, P. N., Assman, G., Gotto, A. M., Jr., & Fredrickson, D. S. (1983) in *The Metabolic Basis of Inherited Disease*, eds. Stanbury, J. B., Wyngaarden, J. B., Fredrickson, D. S., Goldstein, J. L. & Brown, M. S. (McGraw-Hill, New York), 5th Ed., pp. 589–621.
2. Scanu, A. M., Edelstein, C. & Shen, B. (1982) in *Lipid Protein Interactions*, eds. Jost, P. C. & Griffith, O. H. (Wiley, New York), Vol. 1, pp. 259–316.
3. Swaney, J. B., Braithwaite, F. & Eder, H. A. (1977) *Biochemistry* **16**, 271–279.
4. Fitch, W. M. (1977) *Genetics* **86**, 623–644.
5. McLachlan, A. D. (1977) *Nature (London)* **267**, 465–466.
6. Barker, W. C. & Dayhoff, M. O. (1977) *Comp. Biochem. Physiol. B* **57**, 309–315.
7. Boguski, M. S., Elshourbagy, N., Taylor, J. M. & Gordon, J. I. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 5021–5025.
8. McLean, J. W., Fukazawa, C. & Taylor, J. M. (1983) *J. Biol. Chem.* **258**, 8993–9000.
9. Brewer, H. B., Jr., Fairwell, J., LaRue, A., Ronan, R., Houser, A. & Bronzert, T. J. (1978) *Biochem. Biophys. Res. Commun.* **80**, 623–630.
10. Wilbur, W. J. & Lipman, D. J. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 726–730.
11. McLachlan, A. D. (1971) *J. Mol. Biol.* **61**, 409–424.
12. McLachlan, A. D. (1972) *J. Mol. Biol.* **64**, 417–437.
13. Gordon, J. I., Smith, D. P., Andy, R., Alpers, D. H., Schonfeld, G. & Strauss, A. W. (1982) *J. Biol. Chem.* **257**, 971–978.
14. Poncin, J. E., Martial, J. A. & Gielen, J. E. (1984) *Eur. J. Biochem.* **140**, 493–498.
15. Fitch, W. M. & Margoliash, E. (1970) *Evol. Biol.* **4**, 67–109.
16. Kaiser, E. T. & Kezdy, F. J. (1984) *Science* **223**, 249–254.
17. Rall, S. C., Weisgraber, K. H., Mahley, R. W., Ogawa, Y., Fielding, C. J., Uterman, G., Haas, J., Steinmetz, A., Menzel, H. J. & Assman, G. (1984) *J. Biol. Chem.* **259**, 10063–10070.
18. Segrest, J. P., Jackson, R. L., Morrisett, J. D. & Gotto, A. M. (1974) *FEBS Lett.* **38**, 247–253.
19. Chou, P. Y. & Fasman, G. D. (1978) *Annu. Rev. Biochem.* **47**, 251–276.
20. Karathanasis, S. K., Zannis, V. I. & Breslow, J. L. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 6147–6151.
21. Shoulders, C. C., Kornblihtt, A. R., Munro, B. S. & Baralle, F. E. (1983) *Nucleic Acids Res.* **11**, 2827–2837.
22. Gordon, J. I., Sims, H. F., Lentz, S. R., Edelstein, C., Scanu, A. M. & Strauss, A. W. (1983) *J. Biol. Chem.* **258**, 4037–4044.
23. McLachlan, A. D. (1983) *J. Mol. Biol.* **169**, 15–30.
24. Utermann, G., Steinmetz, A., Paetzold, R., Wilk, J., Feussner, G., Kaffarnik, H., Mueller-Eckhardt, C., Seidel, D., Vogelberg, K.-H. & Zimmer, R. (1982) *Hum. Genet.* **61**, 329–337.
25. Mahley, R. W. & Innerarity, T. L. (1983) *Biochim. Biophys. Acta* **737**, 197–222.