Human apolipoprotein B: analysis of internal repeats and homology with other apolipoproteins

Hans De Loof,* Maryvonne Rosseneu,* Chao-Yuh Yang,† Wen-Hsiung Li,†† Antonio M. Gotto, Jr.,† and Lawrence Chan†,**

Department of Clinical Biochemistry,* A. Z. St-Jan, B-8000 Brugge, Belgium; Department of Medicine† and Department of Cell Biology,** Baylor College of Medicine, Houston, TX 77030; and Center for Demographic and Population Genetics,†† University of Texas, Houston, TX 77030

Abstract Apolipoprotein B (apoB) is the major protein component of plasma low density lipoproteins (LDL) and, through its binding to the LDL receptor, it plays a prominent role in lipoprotein metabolism and in the development of atherosclerosis. Specially developed computer programs were applied to detect potential internal repeats in the human apoB sequence and homology of some of these repeats with other apolipoproteins. The simultaneous computer alignment of several (repeated) sequences, carried out in an iterative way to generate consensus sequences, showed the presence of repeated amphipathic helical regions and of repeated hydrophobic proline-rich domains. Extensive Monte-Carlo statistics were used to demonstrate the statistical significance of the internal repeats. Both classes of repeats may contribute to the specific lipid-binding characteristics of apoB. Additional homology, detected between apoB and apoE, the other apolipoprotein-ligand of the LDL receptor, further defined the structural requirements for this receptor-ligand interaction. The computer programs developed in this study should also be useful for detecting internal repeats in other proteins.—**De Loof, H., M. Rosseneu, C-Y. Yang, W-H. Li, A. M. Gotto, Jr., and L. Chan.** Human apolipoprotein B: analysis of internal repeats and homology with other apolipoproteins. *J. Lipid Res.* 1987. **28:** 1455–1465.

Supplementary key words lipid-binding • receptor-binding • atherosclerosis

Apolipoprotein (apo) B-100 is the protein ligand in low density lipoproteins (LDL) that binds to the LDL receptor (1). It is thus an important determinant that regulates LDL metabolism. Elevated plasma levels of LDL-apoB are strongly associated with increased risk of coronary artery disease (2, 3). Indeed, hyperapoB is a significant risk factor for atherosclerosis, even in the presence of normal serum cholesterol (4).

ApoB-100 is the largest protein component in human lipoproteins. The protein is characterized by its extreme insolubility in aqueous media after removal of the lipid, by its inability to transfer among lipoprotein particles, and by its high molecular weight (5).

Until recently, little was known of the apoB-100 primary structure. Attempts at its elucidation were hampered by its enormous size, insolubility, and tendency to aggregate. Recently, the primary structure of apoB-100 has been deduced from its cDNA sequence by four different laboratories (6–10). ApoB-100 turns out to be the largest monomeric protein ever studied. It comprises 4,536 amino acid residues, with a calculated molecular mass of 512,937 daltons.

When apoB-100 is digested with trypsin, the amino acid composition of the remaining parts of the protein shows little difference from that of the undigested protein (11, 12). Furthermore, the amino acid composition of the carboxyl-terminal fifth of apoB-100 (836 residues) differs only slightly from that of the whole molecule (13). These observations suggest that apoB-100 contains internal repeats. This is interesting because it might explain how such an exceptionally large protein has evolved and because internal repeats have been found in all the other apolipoproteins (14–17). Our initial analysis of the human apoB-100 sequence has indeed suggested the presence of numerous repeated sequences within this huge protein (8).

However, the existence of internal repeats in apoB-100 deserves a more careful study because the statistical significance of the potential repeats suggested in our previous study was not rigorously established and because two other laboratories failed to detect such repeats (7, 9) and a third laboratory could identify only uniquely repeated sequences of 6–12 residues, scattered throughout the apoB sequence (10).

Here, we have carefully re-analyzed the human apoB-100 sequence using specially developed computer programs that detect statistically significant repeats within

Abbreviation: LDL, low density lipoprotein.

extraordinarily long sequences such as apoB-100. The question of internal repeats in apoB-100 is important not only because it provides one mechanism for the evolution of such an exceptionally large protein, but more importantly, because the different repeats might delineate functionally important domains within this unique protein. In addition, we have analyzed for potential homology between apoB-100 and other apolipoproteins. Our results provide clues to some common structure-function relationship between these evolutionarily related proteins, especially with respect to the probable mechanisms of lipid-binding, and ligand-receptor interaction of apoB-100 and apoE with the LDL receptor.

MATERIALS AND METHODS

Comparison matrices

We analyzed the sequence reported by Yang et al. (8). Analysis of internal repeats in apoB and of its homology with the other apolipoproteins was based on the comparison-matrix method (18, 19). In this method, all possible segments of a given length from one sequence are compared with all segments of the same length from the other sequence using a scoring matrix. The same procedure can also be applied to a single sequence to identify and locate internal repeated sequences. By using a scoring matrix, this procedure allows the detection of segments that are not exact duplications, but in which amino acid substitutions have generally conserved the physicochemical properties of the residues. The scoring matrix used is that of Staden (20) and is derived from the mutation data matrix (PAM 250) devised by Dayhoff, Barker, and Hunt (18). All scores exceeding a certain threshold value are plotted in a two-dimensional graph with coordinates corresponding to the center of the compared residue spans. We used 25- and 40-residue-long segments and calculated the comparison scores between all segments. Each comparison score was divided by the segment length in order to obtain a mean score. In this way, calculations using different segment lengths can be compared. The threshold value was set at a mean score of 11, a level such that the well-documented internal repeats of apoA-I, A-IV, or E appear clearly on the comparison matrices (data not shown). Because a comparison matrix of very long sequences contains a large amount of data, the data were represented in a condensed way. The number of scores for the comparison of one span with the rest of the sequence, exceeding the threshold value, was plotted in the center of that span.

An alternative method of Kubota et al. (21) was applied to graphically locate the internal repeats. In this technique, cross correlation coefficients are calculated between 25-residue segments. The average of three correlation coefficients, obtained by using three different parameters for the 20 amino acids, was plotted in a comparison matrix when the average was higher than 0.475. These three parameters used were the conformational parameters determined by Levitt (22). They are based on the physicochemical characteristics of the amino acids, not on their mutability as in the calculations using the scoring matrix of Staden (20). No more than three parameters were used at a time due to limitations in computer time and memory. Computations with one parameter such as hydrophobicity (23) or bulkiness (24) were also carried out.

A third method was used to generate another type of comparison matrix using the programs FASTP and RDF of Lipman and Pearson (25). The complete apoB sequence was divided into 200-residue segments starting every 100 residues. All non-overlapping segments were compared and alignments were optimized. The data obtained by this comparison procedure were plotted in a two-dimensional matrix.

Multiple sequence alignment

Alignment of the repeated sequences in apoB was carried out using specially developed computer programs. For this purpose, all consecutive apoB segments with a certain length (a "window" was moved through the sequence) were compared to a query sequence of the same length. These sequences were aligned using the Needleman-Wunch algorithm (26) and a comparison score was calculated using the scoring matrix of Staden (20). Multiple gaps were allowed but penalized (n = number of consecutive gaps, penalty = $20 + [n - 1] \times 5$). All non-overlapping segments exceeding a certain score were saved and used to generate a new query sequence. This procedure was repeated and carried out in an iterative way to generate optimized consensus sequence. Usually after 5–15 cycles, the consensus sequence remained unchanged and was then considered as optimized.

Initial query sequences were chosen as those domains of apoB that yielded the highest homology with other parts of the sequence or with the other apolipoproteins (see Fig. 1, B and D). This procedure was first validated by aligning the multiple repeats within the apoA-I sequence. The various parameters (threshold score, minimal alignment, gap penalty) were selected such that the computer-generated alignments were in agreement with those obtained by manual alignment (17). We tested this procedure with the apoE sequence, a repetitive protein whose repeats are not as well defined and are not punctuated by proline residues. The overall result obtained using our program is very similar to the one reported by Boguski et al. (19) or Luo et al. (17) (data not shown).

For the longer consensus sequences (>25 residues), the process was started with shorter segments. When a consensus was obtained, it was used as the core of longer

query sequences (+ 3 residues on both sides). This procedure was repeated until the number of aligned sequences or the average alignment score decreased to a minimum.

Statistical analysis

In order to validate these consensus sequences and to locate the homologous regions throughout the sequence, we used the method developed by Kubota et al. (27). Average cross-correlation coefficients, based on ten parameters, were calculated for all consecutive segments of apoB, compared with the consensus sequences. Correlation coefficients greater than 0.3 were considered to be significant as originally proposed by Kubota et al. (27). This method has been successfully used to detect the internal repeats in rat apoA-IV (19).

The use of a computer-algorithm for the alignment of several sequences enables quantification of the statistical significance of the repeated sequences by the Monte-Carlo technique (18). The distribution of scores in the Dayhoff matrix and homologies with the consensus sequence for randomized sequences was obtained by simulation. The random sequences used have the same length and the same amino acid composition as the sequence under study. The probability of obtaining a number of scores equal to or larger than a given value X is expressed as the number of standard deviations (SD value) from the mean value of the randomized sequences to the actual number of segments equal to or larger than X (18). Randomizations were carried out 100 times.

Control calculations were carried out by using the distance matrix developed by Bacon and Anderson (28), which is based on the Euclidean distances derived from a number of physicochemical characteristics of the 20 amino acids. These computations ruled out the possibility that the high values for cysteine, tryptophan, or proline in the Staden distance matrix (18) contribute to the statistical significance of the repeats. Validation of this method was carried out using two non-repetitive proteins: β-globin (human) and serum retinol-binding protein, a lipid-binding protein. Positive controls were apoA-I and apoE. Randomizations in the RDF program (25) were carried out 1,000 times.

Analysis of secondary structure potential

The secondary structure of the repeated sequences was analyzed by the methods of Chou and Fasman (29) and Garnier, Osguthorpe, and Robson (30). Helical hydrophobic moments were calculated as previously described (31).

RESULTS

The comparison matrix of apoB with itself is shown in **Fig. 1A**. Repeats are indicated as relatively short diagonals, parallel to the main diagonal. These computations were carried out using 25-residue segments (upper right half of the figure) and 40-residue segments (lower left half). The plot consists only of segments with a mean comparison score equal to or higher than 11. As shorter segments have a higher probability of chance similarity than longer segments (see below), the background is higher with the 25-residue segments. A close inspection shows that in both parts of the figure, the same regions of apoB contain clusters of diagonals, indicative of internal repeats. The comparison scores for each of the 25-residue segments were generally only moderately higher than 11; all were below 12.9. For comparison, the maximum score for the segments in apoA-I, a protein known to have numerous internal repeats, was 12.4.

In order to estimate the statistical significance of these internal repeats, we compared the distribution of scores obtained for the real sequence with those of the randomized sequences **(Fig. 2, A and B)**. For comparison, the cumulative probability plots for apoA-I, apoE, β-globin, and serum retinol binding protein are presented in Fig. 2, C–F. A clear shoulder can be seen in the cumulative probability distribution of the scores for apoB and for both apoA-I and apoE. The SD value for apoB reaches a value of 24 for segments with a score higher than or equal to 11. This SD value increases at higher scores, up to 139 for the 34 comparisons with a score of 12.5 or more. The use of an independent scoring matrix of Bacon and Anderson (28) confirmed the statistical significance of the observation (data not shown). The distributions for apoA-I and apoE both show marked differences between the actual and the randomized sequences indicative of the presence of internal repeats. The two overlapping distributions for the negative control proteins are also clear (Fig. 2, E–F). Differences in the latter case are always smaller than 3 SD units.

Since all the other apolipoproteins are believed to have a common evolutionary origin and share common repeats (17, 19, 32–34), comparison matrices were constructed between apoB and the apolipoproteins A-I, A-II, A-IV, E, C-I, C-II, and C-III (Fig. 1C). These comparison matrices show that homology exists between some regions of apoB and the other apolipoproteins. The total number of segments in the small apolipoproteins exceeding the threshold value in the comparison with one segment of apoB were also plotted (Fig. 1D). This lineplot shows that the homology is located within certain distinct domains of the apoB protein. Two large domains that contain most of the homologous sequences are located between residues 2035–2506 and between residues 4002–4527. Some smaller domains, as detailed in the legend of Fig. 1, were also detected.

An estimation of the chance occurrence of these homologies was obtained by randomization of the small apolipoprotein sequences followed by a comparison with the

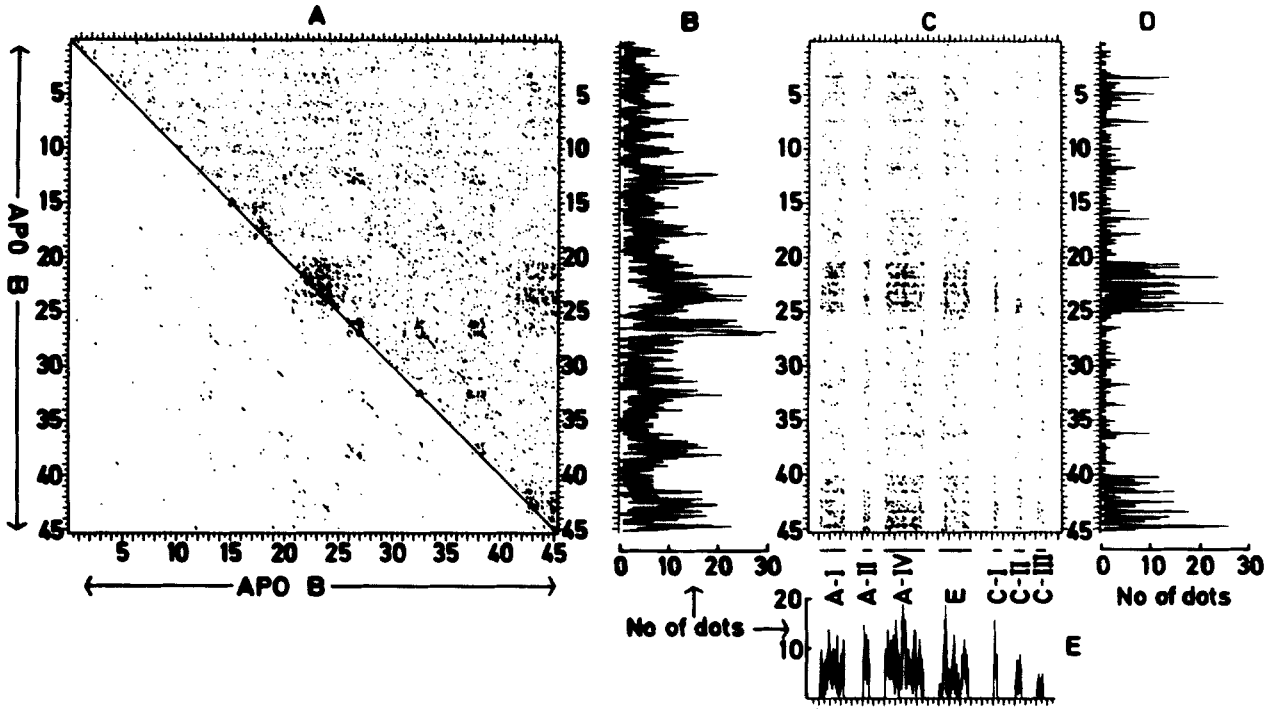

Fig. 1. (A) Sequence comparison matrix of apoB with itself. Two different segment lengths were used. In the upper right part of the figure, 25-residue segments, and in the lower left part, 40-residue segments are compared. Comparison scores, calculated using the matrix proposed by Staden (20), exceeding a mean score of 11 are plotted as dots with coordinates corresponding to the centers of the segments for the 25-residue segments and at residue 20 for the 40-residue long segments. (B) Lineplot showing the number of segments of the complete protein homologous to a particular 25-residue long segment, with a mean score exceeding the threshold value of 11. These numbers are plotted in function of the center of these segments. (C) Comparison matrix of apoB with human apoA-I, A-II, A-IV, E, C-I, C-II, and C-III calculated as in Fig. 1A. Sequences are taken from Mahley et al. (37) and Boguski et al. (19). The segment length is 25. (D) Lineplot showing the number of segments within the different small apolipoproteins having homology with one segment of apoB using 25-residue spans. Two large domains that contain most of the homology are situated between residues 2035–2506 and between residues 4002–4527. Homology in these domains is interrupted by several nearly blank zones, e.g., between residues 2200–2230, 4095–4126, and 4163–4240. Additional smaller peaks are situated around residues 332, 484, 743, 1174, 1643, 1784, and 3623. (E) Lineplot showing the domains within the different small apolipoproteins that show homology with apoB. Homology seems to be of the same order of magnitude for most of the apolipoproteins. Detailed analysis of the plot for apoE, for example, shows that homology is nearly interrupted between residues 160–215. This corresponds to the central divergent zone of exon 4 in apoE as described by Boguski et al. (19).

complete apoB sequence. Although the SD values were lower than those for the internal repeats of apoB, they were statistically significant. The number of scores exceeding the threshold value of 11 yielded SD values of 6.9, 5.3, 10.5, and 8.5 for apoA-I, A-II, A-IV, and E, respectively.

Fig. 1E shows that the homology of apoB with each of the other apolipoproteins is of approximately the same order of magnitude and that the homology is mainly located in the domains with potential amphipathic helices. A more detailed analysis, for example with apoE, shows that homology with apoB is interrupted by a nearly blank zone extending from residues 160–215. This corresponds to the central divergent zone of exon 4 as described by Boguski et al. (19), the domain showing the lowest homology with apoA-IV or with the other amphipathic repeats within apoE.

The identification of the apoB domains with homology to the other apolipoproteins is useful for the interpretation of the comparison of apoB plot with itself. This is obvious from the analysis of the apoB domains, homologous to 15 segments or more (Fig. 1B), but not homologous with the other apolipoproteins (Fig. 1D). These domains correspond to some of the multiple proline-associated regions identified by Knott et al. (7) and are located in the segments containing residues 1132–1397, 2528–2760, 3120–3300, and 3620–3875. The comparison plots using the cross correlation coefficients show a similar pattern as shown in **Fig. 3A**. Similar plots, using only one parameter such as hydrophobicity (23), also allowed the localization of these two classes of repeats (data not shown). This finding indicates that neither the computation method nor the initial data set influenced the final results.

The presence of internal repeats within apoB is also confirmed by analysis using the FASTP program (25). The results of these comparisons, plotted on Fig. 3B which shows the presence of many long repeats, are also in agreement with the conclusions from the other methods.

In contrast to apoA-I and apoA-IV where the repeats

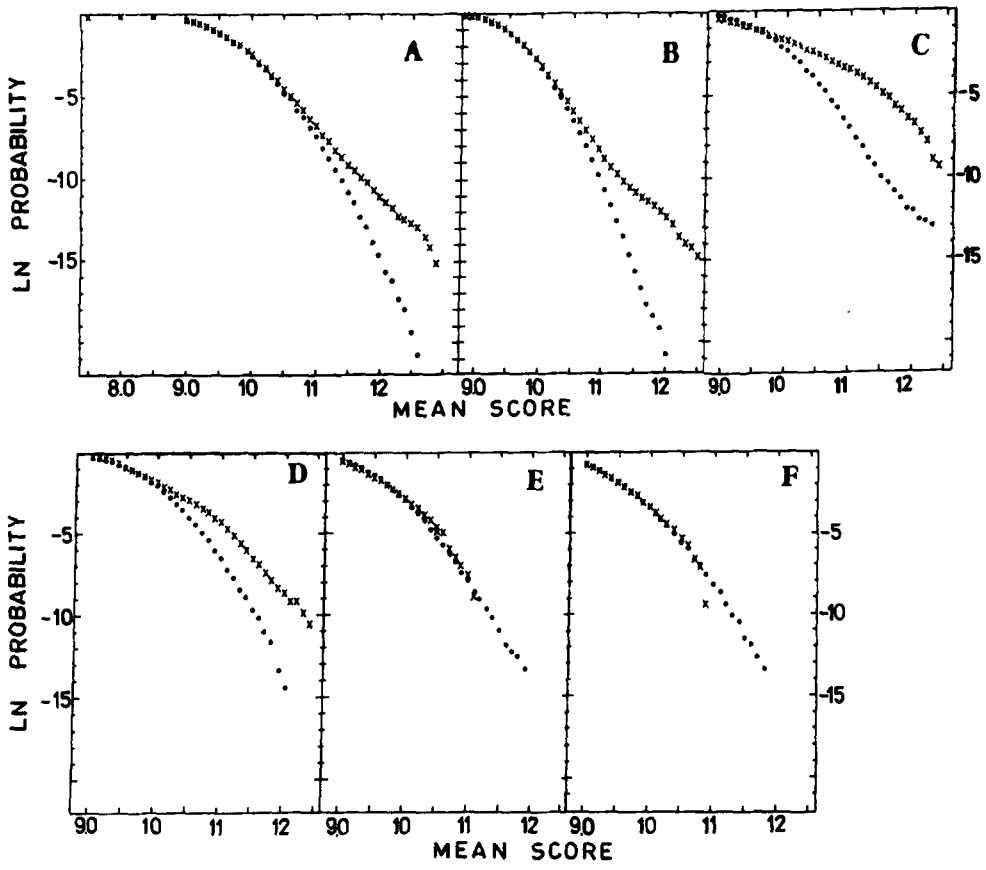

Fig. 2. Cumulative probability (LN probability) distribution of scores in the comparison matrices of real sequences (x) and randomized sequences (O) of apoB with itself using 25-residue segments (A) and 40-residue segments (B). Cumulative probability distribution of the comparison of apoA-I (C), apoE (D), β-globin (E), and serum retinol binding protein (F) with itself using 25-residue segments. For apoB, apoA-I, and apoE, a shoulder can be seen indicative of the non-randomness of the occurrence of repeats. Sequences were taken from reference 36 and the NBRF data bank.

are delineated by proline residues, no sharp boundaries are found between the repeats and, as repeat spacing is variable, alignments were performed by computer procedures. Iterative multiple alignments using the Needleman-Wunsch algorithm (26) were carried out with different query sequences and resulted in the optimal alignment of the two classes of repeats with the consensus sequences as shown in **Table 1** and Table 2. In agreement with the comparison matrix (Fig. 1), the aligned segments belong to distinct domains of the apoB sequence. The proline-rich repeats, homologous to the first 52-residue-long consensus sequence, occur in sequences starting at residues 1283, 2574, 2666, 3245, 3711, and 3805. All have a mean comparison score greater than 11. Moreover, the correlograms confirm the location and statistical significance of these consensus sequences with the complete sequence **(Fig. 4).** Four major peaks, starting at residues 1283, 2666, 3245, and 3805 have correlation coefficients exceeding 0.39, well above the confidence limit of 0.3 used by Kubota et al. (27). They are located within the four proline-rich domains described above. In addition to these four major peaks, smaller ones, starting at residues 1278, 2567, 2606, 2646, 2661, 2671, and 3366 have correlation coefficients around 0.3. This suggests that smaller repeats may build up the larger ones as is evidenced by the correlogram of 25-residue-long consensus sequences. Peaks with correlation coefficients greater than 0.45 are identified as sequences starting at residues 1289, 1296, 2585, 2704, 3219, 3251, 3717, and 3823. The aligned sequences (scores >11.5) start at residues 1296, 2585, 2629, 2704, 3219, 3251, 3717, and 3823.

Extensive Monte-Carlo simulations (18) were performed to show the statistical significance of the repeats and to eliminate the possibility that the consensus sequences contained a bias, due to the pooling together of certain residues without a specific sequence. Randomized consensus sequences were compared with the actual apoB sequence preserving the amino acid composition of the various domains. None of the 100 randomizations of the 52-residue-long sequence yielded a single alignment with

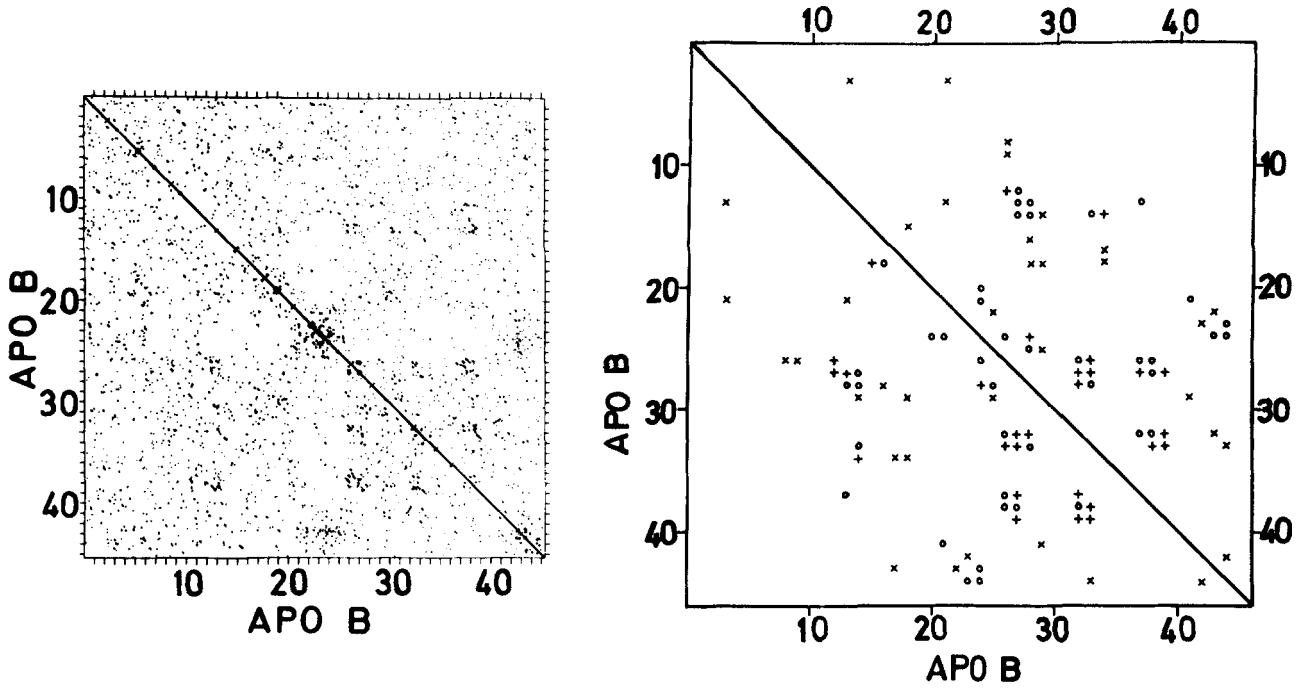

Fig. 3. Comparison matrices using alternative methods. (A) Left: Comparison matrix of apoB with itself obtained by the method developed by Kubota et al. (21) using the conformational parameters of Levitt (22). Correlation coefficients exceeding 0.475 are plotted. The clusters of diagonals are evident in the same regions of apoB as shown in Fig. 2A. (B) Right: Comparison of 200-residue fragments of apoB with all other non-overlapping 200-residue fragments using the FASTP program (ktup = 1); (+) indicates a score equal to or higher than 60; (x) indicates alignments equal to or longer than 90 residues; and (○) combines the latter two criteria. Most of the comparisons indicated by a circle have SD values, determined by the RDF program (25) using 1000 randomizations of three or more.

a mean score equal to or greater than 11. The inverse approach of randomizing the apoB sequence also failed to yield such alignments. The seven 25-residue-long consensus sequences with a mean score of 11.83 or higher reached an SD value of 40. The FASTP and RDF program of Lipman and Pearson (25) applied to these consensus sequences compared to different 1500-residue segments of apoB yielded several single optimal alignments with SD values between 9.5 and 12.3 for the first consensus sequence, and SD values between 5.6 and 7.9 for the 25-residue consensus sequence.

The second class of internal repeats, together with the 22-residue consensus, is shown in **Table 2**. Analysis of the consensus sequence by predictive algorithms and its presentation in an Edmundson-wheel diagram (35) **(Fig. 5A)** both indicate that it has an amphipathic structure. Hydrophobic amino acids are located on one side of the helix while the polar residues are located on the other side. However, as pointed out by Boguski et al. (19), consensus sequences are only approximations of real sequences. In order to examine whether each of the internal repeats identified in Table 2 is amphipathic in nature, we have calculated the mean hydrophobicity and the helical hydrophobic moment of each sequence individually. Data (not shown) indicate that all the sequences have a high helical hydrophobic moment consistent with an amphipathic structure. Furthermore, presentation of the sequences in Edmundson-wheel diagrams clearly indicates that indeed *each* of them forms an amphipathic helix (data not shown).

Alignments starting at residues 2079, 2135, 2173, 2384, 2407, 4150, 4237, 4397, and 4463 have a mean score exceeding 11.5. On the correlogram, the highest peaks are also concentrated in the two domains described above; peaks with a score higher than 0.4 start at residues 2079, 2173, 2384, 2407, 2500, 2507, 4038, 4150, and 4237. Statistical significance, calculated in the same way as for the "proline-rich" repeats, yielded an SD value of 21.3 for the nine repeats with a score higher than or equal to 11.59.

A correlogram of this consensus sequence with the human apoA-IV sequence revealed the tandemly organized repeat structure (19) of apoA-IV (Fig. 4D), confirming the amphipathic helical characteristics of this consensus sequence. A final test using the RDF program (25) shows that only certain domains yield optimized alignments with SD values higher than 3 (Fig. 4E).

Knott et al. (36) reported some homology of residues 140–150 in apoE with residues 3357–3367 in apoB, although they also reported that there was no significant homology between apoB and any apolipoprotein, including apoE (7, 36). Based on our homology calculations, the receptor-binding domains for apoB and apoE seem to in-

TABLE 1. "Proline-rich" consensus sequences

First Residue	Sequence	Score
A. 52 Residues		
1283	LKMLETVRTPALHFKSVGFHLPSREFQVPTFTIPKLYQLQVP-LLGVLDLSTN	11.88
2574	EVSLQALQKATFQTPDFIVPLTDLRIPSVQINFKDLKNIKIPSRFSTPEFTI-	11.31
2666	LRDLKVEDIPLARITLPDFRLPEIAIPEFIIPTLNLNDFQVP-DLHIPEFQLP	12.28
3245	YVFPKAVSMPSFSILGSDVRVPSYTLILPSLELPVLHVPRNL-KLSLPDFKEL	11.98
3711	NDLNSVLVMPTFHVPFTDLQVPSCKLDFREIQIYKKLRTSSF-ALNLPTLPEV	11.40
3805	SDGIAALDLNAVANKIADFELPTIIVPEQTIEIPSIKFSVPA-GIAIPSFQAL	12.26
Consensus	LDSLKALDMPTFHIPSSDFRLPSITIPEPTIEIPKLKNSQVP-ALSIPDFQEL	
B. 25 Residues		
1296	FKSVGFHLPSREFQVPTFTIPKLYQ	11.84
2585	FQTPDFIVPLTDLRIPSVQINFKDL	12.76
2629	FHIPSFTIDFVEMKVKIIRTIDQML	11.68
2704	FQVPDLHIPEFQLPHISHTIEVPTF	12.72
3219	FQIPGYTVPVVNVEVSPFTIEMSAF	12.80
3251	VSMPSFSILGSDVRVPSYTLILPSL	12.24
3717	LVMPTFHVPFTDLQVPSCKLDFREI	12.54
3823	FELPTIIVPEQTIEIPSIKFSVPAG	12.64
	FQMPSFHVPETDLEVPSITIEVPAL	

Fifty two- and 25-residue-long consensus sequences derived by the iterative alignment procedure. Identical residues are printed in bold face type and related amino acids are underlined. The mean scores are calculated using the Staden (20) scoring matrix. Gaps are penalized as described in Methods.

clude more residues than originally proposed (36, 37). The comparison of the two domains yields additional interesting information about the structural requirements for receptor-protein interaction. The homologous residues of the two sequences were plotted in an Edmundson-wheel diagram (35) (Fig. 5B) which shows that most residues match and that the amphipathic nature of the two segments is very well preserved (38).

The general similarities in structure notwithstanding, interesting differences were also observed. Residues Lys-143 and Asp-154 of apoE are located on the same side of the helix and have opposite charges. Theoretically, they may neutralize each other and reduce their importance in the polar interactions with the LDL receptor. In apoB, the two corresponding residues, Leu-3360 and Ala-3371, are not charged, thus maintaining a neutral environment for receptor interaction. Futhermore, residue Arg-136 in apoE and residue Lys-3353 in apoB are both positively charged. When the receptor-binding domains of human (37), rat (39) and mouse (40) apoE are aligned with that of apoB, a positively charged amino acid is again present at this position (Fig. 5B). This would extend the residues potentially important for receptor-binding in apoE to 136–160 which is compatible with the region identified by hydrophobicity analysis (31). The significance of this region of homology was tested by comparing the 20-residue segment of apoB (3352–3371) with the whole apoE sequence. Using 1,000 randomizations, we obtained a z-score of 7.84. The similarity of these protein domains does not go beyond these 20 residues because the statistical significance decreases markedly when more residues were aligned.

DISCUSSION

Because of the importance of apoB in the development of atherosclerosis and in genetically determined hyperlipidemias, and because of the unique nature of its lipid-binding properties, the elucidation of the apoB structure has been a subject of intense research and a source of frustration for numerous investigators. The sequence of this protein is of special interest as apoB seems to be a highly polymorphic protein (41) and mutations in some functionally important domains may interfere with specific LDL-apoB functions such as receptor binding. Studies on this clinically important protein will provide insight into the structure-function relationship of its various subdomains in analogy with studies on apoE and with the LDL receptor (1).

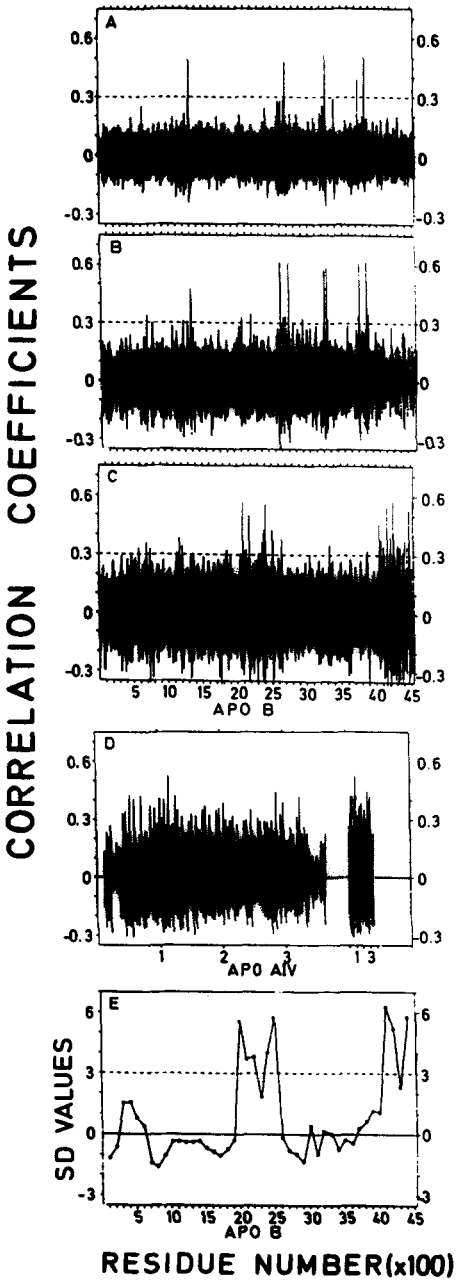

Fig. 4. Correlograms for sequence comparison according to Kubota et al. (27). (A) Correlogram of the total apoB sequence with the 52-residue "proline-rich" consensus sequence. (B) Correlogram as in Fig. 4B with the 25-residue long consensus sequence. (C) Correlogram with the 22-residue consensus sequence. (D) Correlogram of the 22-residue consensus sequence with the human apoA-IV sequence. (E) Comparison of the 22-residue consensus sequence with 200-residue fragments of apoB using the RDF program. The SD values for the optimized scores obtained after 1000 randomizations (ktup = 1) are plotted in function of the midpoint of the 200-residue fragments. The location of amphipathic-helic repeats within the two zones, as described in the text, is clearly visible as zones with scores exceeding the statistical significance limit of 3 SD values.

Considering the importance of this sequence, we have analyzed it for the existence of internal repeats and of homologies with other apolipoproteins, hoping to gain insight into its structure-function relationship.

Studies in other laboratories failed to identify internal repeated sequences or homology with other apolipoproteins (7, 9, 19). One study suggested that similarity to other apolipoproteins is restricted to a single 11-residue segment between apoE and apoB, a domain with the putative LDL receptor-binding domain of both proteins (36). The negative conclusions are probably due to the extreme length of this sequence, making computations very time-consuming, to the different methods employed (methods generally developed for shorter sequences with highly conserved internal repeats), and to the fact that no repeats or homologous domains have very high similarity scores. In our previous study (8), we suggested that many potential internal repeats exist in apoB. The analysis, however, was not rigorous because we had not performed any randomization tests to see how often segments with high similarities can arise by chance. The present analysis is based on computer programs developed specially to detect internal repeats within long sequences at a lower threshold. Special attention was paid to the statistical significance of the apparent homologous regions identified. An iterative procedure was applied to the search of an optimal consensus sequence for the apoB repeats detected on the comparison matrix.

Our computation procedure has identified the amphipathic helical segments (38) in apoB that are homologous to those in other apolipoproteins. It has demonstrated significant homology between putative receptor-binding domains in apoB and apoE, and has further revealed the existence of "proline-rich" repeats characteristic for apoB.

The homologous repeats that are common to apoB and the other apolipoproteins show characteristic amphipathic helices (Fig. 5A). These amphipathic helices have been extensively investigated by different groups using a variety of methods (38, 42-45), and are thought to be important for phospholipid binding. Interestingly, on intact LDL particles, the domains containing these repeats are generally inaccessible to trypsin (8), further supporting the hypothesis that these domains are involved in lipid binding.

In contrast, the "proline-rich" repeats unique to apoB are characterized by the preponderance of hydrophobic residues. Their secondary structure is predicted (29, 30) to be composed of predominantly β-sheets and β-turns (due to proline residues). We speculate that they interact with lipids in a different way. Computer modeling (Brasseur, R., H. De Loof, M. Rosseneu, and J-M. Ruysschaert, unpublished data) of these proline-rich sequences in the presence of dipalmitoylphosphatidylcholine suggests that the first part of such a segment consists of a β-sheet that might penetrate into the acyl chains. After a turn around a proline residue, the segment can form a second β-sheet parallel to the first one, but with a reverse orientation. The relative symmetry of these structures can account for

TABLE 2. Twenty-two residue consensus

First Residue		Score	Mean Hydrophobicity	Mean Helical Hydrophobic Moment
2079	QFVRKYRAALGKLPQQANDYLN	12.54	−.20	0.98
2135	DAKINFNEKLSQLQTYMIQFDQ	11.63	−.08	0.89
2173	NIIDEIIEKLKSLDEHYHIRVN	12.09	−.06	1.02
2384	TFIEDVNKFLDMLIKKLKSFDY	11.59	.06	1.01
2407	QFVDETNDKIREVTQRLNGEIQ	12.31	−.32	1.02
4150	RVTQEFHMKVKHLIDSLIDFLN	12.68	.03	1.00
4237	DVISMYRELLKDLSKEAQEVFK	11.95	−.12	0.99
4397	EYIVSASNFTSQLSSQVEQFLH	11.85	.13	0.69
4463	DYHQQFRYKLQDFSDQLSDYYE	12.90	−.31	0.82
Consensus	DFIDEFNEKLKDLSDQLNDFLN		−.13	0.98

Twenty-two-residue-long consensus sequence derived by the iterative alignment procedure. This consensus sequence is shown in an Edmundson wheel (35) representation in Fig. 5A. Identical residues are printed in boldface type and related amino acids are underlined. The mean hydrophobicity and mean helical hydrophobic moment are calculated for all segments as previously described (31). The mean hydrophobicity of the different segments is close to zero. The mean helical hydrophobic moment is, however, close to unity for most segments. This is indicative of the amphipathic nature of these segments when oriented in a helical conformation. The presentation of the individual segments in an Edmundson-wheel diagram confirms that each segment can form an amphipathic helix.

the partial overlapping of these repeats. Such a structure would be able to penetrate more deeply into the LDL than the amphipathic helices.

Cooperativity in the lipid-binding of these two different classes of subdomains, i.e., amphipathic versus proline-rich regions, might account for the observation that, in contrast to the smaller apolipoproteins, apoB does not exchange between different lipoprotein particles (5). Nevertheless, it is noteworthy that there are no segments within the apoB sequence with a hydrophobicity comparable to the membrane-spanning segments of integral membrane proteins (6).

The genomic structure of apoB has recently been reported by Blackhart et al. (46). The intron-exon junctions do not clearly define the relative positions of the various internal repeats, except that the last intron-exon boundary, occurring after the residue 4002, delineates the large COOH-terminal domain homologous with the other apolipoproteins.

In conclusion, using a combination of computation

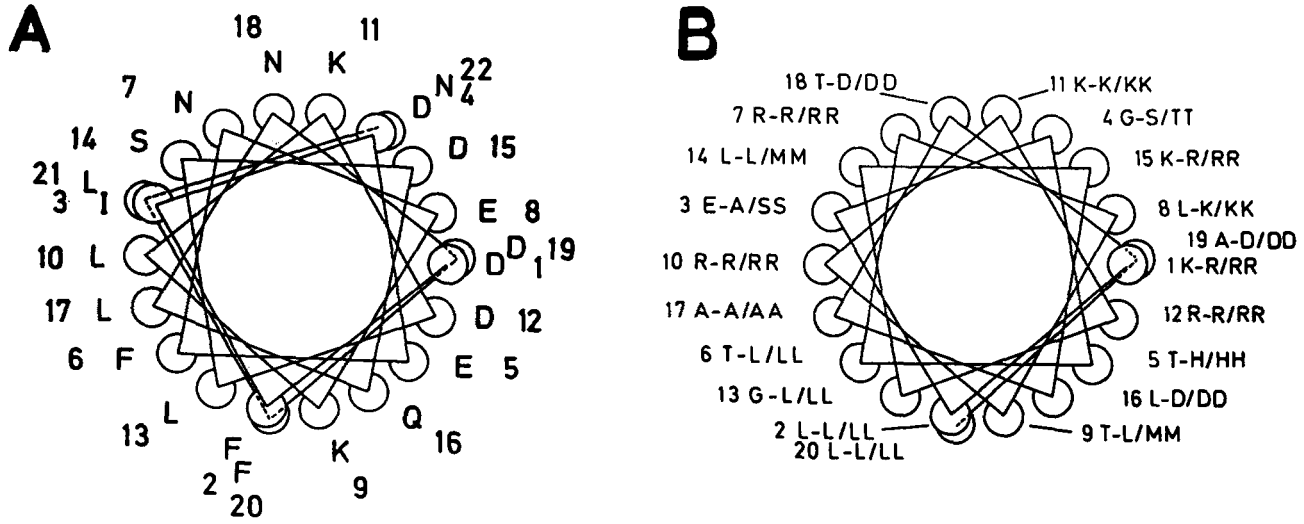

Fig. 5. (A) Edmundson-wheel diagram (35) of the 22-residue consensus sequence. One side of the helix clearly has highly hydrophobic amino acids while non-hydrophobic amino acids are present on the other side of the helix. (B) Edmundson-wheel representation of the region of the apoB sequences (3352–3371) homologous to the receptor binding region of human apoE (136–155) (31, 37) and corresponding residues of the rat apoE (39), and mouse apoE sequence (residues 128–147) (40). Individual amino acids, labeled from left to right, represent residues in human apoB, human apoE/rat apoE, mouse apoE, respectively, e.g., 18 T-D/D D.

procedures, we have identified the presence of different types of internal repeats in apoB-100 and sequence homology between this protein and the soluble apolipoproteins. The internal repeats share interesting physical properties which might bear on the overall physicochemical behavior of apoB-100. The methods used in this study can be directly applied to the identification of intra- and intersequence homologies among other protein sequences of different lengths.

Dr. Hans De Loof is a recipient of an I. W. O. N. L. fellowship. This work was supported by NIH grants H27341 and HL16512 to L. Chan and GM-30998 to W-H. Li, and by a grant from the March of Dimes Birth Defects Foundation.

Manuscript received 31 March 1987 and in revised form 15 June 1987.

REFERENCES

1. Brown, M. S., and J. L. Goldstein. 1986. A receptor-mediated pathway for cholesterol homeostasis. *Science.* **232:** 34–47.
2. The National Institutes of Health Consensus Development Conference. 1985. Lowering blood cholesterol to prevent heart disease. *JAMA.* **253:** 2080–2086.
3. Lipid Research Clinics Program. 1984. The Lipid Research Clinics Coronary Primary Prevention Trial results. *JAMA.* **251:** 351–374.
4. Sniderman, A., S. Shapiro, D. Marpole, B. Skinner, B. Teng, and P. O. Kwiterovich, Jr. 1980. Association of coronary atherosclerosis with hyperapobetalipoproteinemia (increased protein but normal cholesterol levels in human plasma low density lipoproteins). *Proc. Natl. Acad. Sci. USA.* **77:** 601–608.
5. Kane, J. P. 1983. Apoprotein B: structural and metabolic heterogeneity. *Annu. Rev. Physiol.* **45:** 637–650.
6. Chen, S-H., C-Y. Yang, P-F. Chen, D. Setzer, M. Tanimura, W-H. Li, A. M. Gotto, and L. Chan. 1986. The complete cDNA and amino acid sequence of human apolipoprotein B-100. *J. Biol. Chem.* **261:** 12918–12921.
7. Knott, T. J., R. J. Pease, L. M. Powell, S. C. Wallis, S. C. Rall, T. L. Innerarity, B. Blackhart, W. H. Taylor, Y. Marcel, R. Milne, D. Johnson, M. Fuller, A. J. Lusis, B. J. McCarthy, R. W. Mahley, B. Levy-Wilson, and J. Scott. 1986. Complete protein sequence and identification of structural domains of human apolipoprotein B. *Nature.* **323:** 734–738.
8. Yang, C. Y., S-H. Chen, S. H. Gianturco, W. A. Bradley, J. T. Sparrow, M. Tanimura, W-H. Li, D. A. Sparrow, H. DeLoof, M. Rosseneu, F-S. Lee, Z-W. Gu, A. M. Gotto, Jr., and L. Chan. 1986. Sequence, structure, receptor binding domains and internal repeats of human apolipoprotein B-100. *Nature.* **323:** 738–742.
9. Law, S. W., S. M. Grant, K. Higuchi, A. Hospattankar, K. Lackner, N. Lee, and H. B. Brewer, Jr. 1986. Human liver apolipoprotein B-100 cDNA: complete nucleic acid and derived amino acid sequence. *Proc. Natl. Acad. Sci. USA.* **83:** 8142–8146.
10. Cladaras, C., M. Hadzopoulos-Cladaras, R. T. Notte, D. Atkinson, and V. I. Zannis. 1986. The complete sequence and structural analysis of human apolipoprotein B-100: relationship between apoB-100 and apoB-48. *EMBO J.* **5:** 3495–3507.
11. Margolis, S., and R. G. Langdon. 1966. Human serum beta-$_1$ lipoprotein. III. Enzymatic modifications. *J. Biol. Chem.* **241:** 485–493.
12. Chapman, M. J., S. Goldstein, and G. L. Mills. 1978. Limited tryptic digestion of human serum low-density lipoprotein: isolation and characterization of the protein-deficient particles and of its apoprotein. *Eur. J. Biochem.* **87:** 475–488.
13. Wei, C. F., S. H. Chen, C. Y. Yang, Y. L. Marcel, R. W. Milne, W-H. Li, J. T. Sparrow, A. M. Gotto, Jr., and L. Chan. 1985. Molecular cloning and expression of partial cDNAs and deduced amino acid sequence of a carboxyl-terminal fragment of human apolipoprotein B-100. *Proc. Natl. Acad. Sci. USA.* **82:** 7265–7269.
14. Karathanasis, S. K., V. I. Zannis, and J. L. Breslow. 1983. Isolation and characterization of the human apolipoprotein A-I gene. *Proc. Natl. Acad. Sci. USA.* **80:** 6147–6151.
15. Boguski, M. S., N. Elshourbagy, J. M. Taylor, and J. I. Gordon. 1985. Comparative analysis of the repeated sequences in rat apolipoproteins A-I, A-IV, and E. *Proc. Natl. Acad. Sci. USA.* **82:** 992–996.
16. Shelley, C. S., C. R. Sharpe, F. E. Baralle, and C. C. Shoulders. 1985. Comparison of the apolipoprotein genes. ApoA-II presents a unique functional intron-exon junction. *J. Mol. Biol.* **186:** 43–51.
17. Luo, C-C., W-H. Li, M. N. Moore, and L. Chan. 1986. Structure and evolution of the apolipoprotein multigene family. *J. Mol. Biol.* **187:** 325–340.
18. Dayhoff, M. O., W. C. Barker, and L. T. Hunt. 1983. Establishing homologies in protein sequences. *Methods Enzymol.* **91:** 524–545.
19. Boguski, M. S., M. Freeman, N. A. Elshourbagy, J. M. Taylor, and J. I. Gordon. 1986. On computer-assisted analysis of biological sequences: proline punctuation, consensus sequences, and apolipoprotein repeats. *J. Lipid Res.* **27:** 1011–1034.
20. Staden, R. 1982. An interactive graphics program for comparing and aligning nucleic acid and amino acid sequences. *Nucl. Acid Res.* **10:** 2951–2961.
21. Kubota, Y., K. Nishikawa, S. Takahashi, and T. Ooi. 1982. Correspondence of homologies in amino acid sequence and tertiary structure of protein molecules. *Biochim. Biophys. Acta.* **701:** 242–252.
22. Levitt, M. 1978. Conformational preferences of amino acids in globular proteins. *Biochemistry.* **17:** 4277–4284.
23. Eisenberg, D. 1984. Three-dimensional structure of membrane and surface proteins. *Annu. Rev. Biochem.* **53:** 595–623.
24. Zimmerman, J. M., N. Eliezer, and R. Simha. 1968. The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.* **21:** 170–177.
25. Lipman, D. J., and W. R. Pearson. 1985. Rapid and sensitive protein similarity searches. *Science.* **227:** 1425–1441.
26. Needleman, S. B., and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48:** 443–453.
27. Kubota, Y., S. Takahashi, K. Nishikawa, and T. Ooi. 1981. Homology in protein sequences expressed by correlation coefficients. *J. Theor. Biol.* **91:** 347–361.
28. Bacon, D. J., and W. F. Anderson. 1986. Multiple sequence alignment. *J. Mol. Biol.* **191:** 153–161.

29. Chou, P. Y., and G. D. Fasman. 1978. Empirical predictions of protein conformation. *Annu. Rev. Biochem.* **47:** 251–276.
30. Garnier, J., D. J. Osguthorpe and B. Robson. 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120:** 9–120.
31. De Loof, H. M. Rosseneu, R. Brasseur, and J. M. Ruysschaert. 1986. Use of hydrophobicity profiles to predict receptor binding domains on apolipoprotein E and the low density lipoprotein apolipoprotein (B-E) receptor. *Proc. Natl. Acad. Sci. USA.* **83:** 2295–2299.
32. Fitch, W. M. 1977. Phylogenies constructed by the crossover process as illustrated by human hemoglobins and a thirteen-cycle, eleven amino acid repeat in human apolipoprotein A-I. *Genetics.* **86:** 623–644.
33. McLachlan, A. D. 1977. Repeated helical pattern in apolipoprotein A-I. *Nature.* **267:** 465–466.
34. Barker, W. C., and M. O. Dayhoff. 1977. Evolution of lipoproteins deduced from protein sequence data. *Comp. Biochem. Physiol. B.* **57:** 309–315.
35. Schiffer, M., and A. B. Edmundson. 1967. Use of helical wheels to represent the structures of proteins and to identify segments with helical potential. *Biophys. J.* **7:** 121–125.
36. Knott, T. J., S. C. Rall, T. L. Innerarity, S. F. Jacobson, M. S. Urdea, B. Levy-Wilson, L. M. Powell, R. J. Pease, R. Eddy, H. Nakai, M. Beyers, L. M. Priestley, E. Robertson, L. B. Rall, C. Betsholz, T. B. Shows, R. W. Mahley, and J. Scott. 1985. Human apolipoprotein B: structure of carboxyl-terminal domains, sites of gene expression, and chromosomal localization. *Science.* **230:** 37–43.
37. Mahley, R. W., T. L. Innerarity, S. C. Rall, and K. H. Weisgraber. 1984. Plasma lipoproteins: apolipoprotein structure and function. *J. Lipid Res.* **25:** 1277–1294.
38. Segrest, J. P., R. L. Jackson, J. D. Morrisett, and A. M. Gotto, Jr. 1974. A molecular theory of lipid-protein interactions in the plasma lipoproteins. *FEBS Lett.* **38:** 247–258.
39. McLean, J. W., C. Fukazawa, and J. M. Taylor. 1983. Rat apolipoprotein E mRNA. Cloning and sequencing of double stranded cDNA. *J. Biol. Chem.* **258:** 8993–9000.
40. Rajavashisth, T. B., J. S. Kaptein, K. L. Reue, and A. J. Lusis. 1985. Evolution of apolipoprotein: mouse sequence and evidence for an 11-nucleotide ancestral unit. *Proc. Natl. Acad. Sci. USA.* **82:** 8085–8089.
41. Chan, L., P. Van Tuinen, D. H. Ledbetter, S. P. Daiger, A. M. Gotto, and S. H. Chen. 1985. The human apolipoprotein B-100 gene: a highly polymorphic gene that maps to the short arm of chromosome 2. *Biochem. Biophys. Res. Commun.* **133:** 248–255.
42. Segrest, J. P., and R. J. Feldmann. 1977. Amphipathic helixes and plasma lipoproteins: a computer study. *Biopolymers.* **16:** 2053–2065.
43. Anantharamaih, G. M. 1986. Synthetic peptide analogs of apolipoproteins. *Methods Enzymol.* **128:** 627–647.
44. Anantharamaih, G.M., J. L. Jones, C. G. Brouilette, C. F. Schmidt, B. H. Chung, T. A. Hughes, A. S. Bhown, and J. P. Segrest. 1985. Studies of synthetic peptide analogs of the amphipathic helix: structure of complexes with dimyristoylphosphatidylcholine. *J. Biol. Chem.* **260:** 10248–10255.
45. Sparrow, J. T., and A. M. Gotto, Jr. 1981. Apolipoprotein/lipid interactions: studies with synthetic polypeptides. *CRC Crit. Rev. Biochem.* **12:** 87–107.
46. Blackhart, B. D., E. M. Ludwig, V. R. Pierotti, L. Caiati, M. A. Onasch, S. C. Wallis, L. Powell, R. Pease, T. J. Knott, M-L. Chu, R. W. Mahley, J. Scott, B. J. McCarthy, and B. Levy-Wilson. 1986. Structure of the human apolipoprotein B gene. *J. Biol. Chem.* **261:** 15364–15367.