THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 262, No. 24, Issue of August 25, pp. 11794-11800, 1987
Printed in U.S.A.

Primary Structure and Comparative Sequence Analysis of an Insect Apolipoprotein

APOLIPOPHORIN-III FROM *MANDUCA SEXTA**

(Received for publication, January 9, 1987)

Kenneth D. Cole‡, Germain J. P. Fernando-Warnakulasuriya‡, Mark S. Boguski§, Mark Freeman§, Jeffrey I. Gordon§¶‖, Wallace A. Clark‡, John H. Law‡, and Michael A. Wells‡

From the ‡Department of Biochemistry, Biosciences West, University of Arizona, Tucson, Arizona 85721 and the Departments of §Biological Chemistry and ¶Medicine, Washington University School of Medicine, St. Louis, Missouri 63110

The amino acid sequence of an insect apolipoprotein, apolipophorin-III from *Manduca sexta*, was determined by a combination of cDNA and protein sequencing. The mature hemolymph protein consists of 166 amino acids. The cDNA also encodes for an amino-terminal extension of 23 amino acids which is not represented in the mature hemolymph protein. The existence of a precursor protein was confirmed by *in vitro* translation of fat body mRNA. Computer-assisted comparative sequence analysis revealed the following points: 1) the protein is composed of tandemly repeating tetradecapeptide units with a high potential for forming amphiphilic helical structures. Compared to mammalian apolipoproteins the repeat units in the insect apolipoprotein show considerable length variability; 2) the sequence has a striking resemblance to several human apolipoproteins including apoE, AIV, AI, and CI. However, the homology seems to be entirely functional since, although the insect and mammalian apoproteins contain very similar types of amino acid residues, the actual degree of sequence identity is quite low. Whether the mammalian and insect apoproteins are derived from a common ancestral amphiphilic helix forming, lipid-binding protein, or arose by convergent evolution can not be determined at present. This represents the first complete amino acid sequence for an insect apolipoprotein.

Lipid transport in insect hemolymph differs in significant ways from similar processes in mammalian blood. The major lipoprotein, called lipophorin, functions as a recycling shuttle that carries fatty acids in the form of diacylglycerol, and in addition cholesterol, carotenes, and hydrocarbons (Chino *et al.*, 1981). The form of lipophorin varies in different life stages as the lipoprotein function changes (Ryan and Law, 1984). In the tobacco hornworm, *Manduca sexta*, the larval lipoprotein contains two apoproteins, apolipophorin-I (apoLp-I)[1] M_r = 240,000, and apolipophorin-II (apoLp-II) M_r = 80,000, and about 40% lipid, principally phospholipid and diacylglycerol (Pattnaik *et al.*, 1979; Shapiro *et al.*, 1984; Prasad *et al.*, 1986a). This lipoprotein is synthesized in the fat body and serves to shuttle digested lipids from the midgut to growing tissues and storage depots (Prasad *et al.*, 1986b).

In the adult moth, lipid transport is needed mainly to fuel sustained flight. In the resting adult the lipoprotein, which is called high density lipophorin-adult contains apoLp-I and apoLp-II, as well as two molecules of a third apoprotein, apoLp-III, and about 50% lipid (Shapiro and Law, 1983; Kawooya *et al.*, 1984; Wells *et al.*, 1987). During flight, diacylglycerol is mobilized from the fat body and added to high density lipophorin-adult to produce a diacylglycerol-rich, low density lipoprotein, low density lipophorin (Shapiro and Law, 1983; Ryan *et al.*, 1986; Wells *et al.*, 1987). In the course of diacylglycerol addition, several molecules of apoLp-III, which is abundant and free in hemolymph, are added to the growing lipoprotein particle until the fully loaded form has a total of 16 apoLp-III molecules (Wells *et al.*, 1987).

ApoLp-III is a small, highly asymmetric lipid-binding protein (Kawooya *et al.*, 1986), which contains no carbohydrate or other post-translational modifications (Kawooya *et al.*, 1984).

We have determined the primary structure of apoLp-III of *M. sexta*, using both Edman degradation of the protein and peptides derived from it, and from the sequence of a cloned cDNA. The mature protein consists of 166 amino acids and has a molecular weight of 18,380. Computer-assisted comparative analysis of the sequence shows that the mature protein is composed almost entirely of repeated sequences with amphiphilic helical potential and that the protein has a striking similarity to several mammalian apolipoproteins.

MATERIALS AND METHODS

Purification of ApoLp-III—ApoLp-III was purified from adult *M. sexta* as described by Wells *et al.* (1985) except that the final concanavalin A column chromatography step was replaced by HPLC purification. Up to 40 mg of protein was dissolved in 10 ml of 0.25% trifluoroacetic acid and injected onto a Vydac C8 reverse-phase HPLC column (10 × 250 mm). The protein was eluted using a linear gradient formed between 0.25 and 0.20% trifluoroacetic acid in acetonitrile:water (70:30 v/v). The flow rate was 3 ml/min, and the run lasted 1 h. Peak fractions containing apoLp-III were combined and directly lyophilized. HPLC solvents were from Pierce Chemical Co. Double-distilled water was further purified by passage through a Milli-Q water purification system (Millipore Corp., Bedford, MA).

Cyanogen Bromide Cleavage—ApoLp-III was dissolved in 70% trifluoroacetic acid at a concentration of about 10 mg/ml and treated with a 100-fold molar excess of CNBr (Pierce Chemical Co.) with respect to methionine. The mixture was kept in the dark for 24 h, after which the sample was dried under a stream of nitrogen. The residue was dissolved in a small volume of water and lyophilized. The peptides were purified by HPLC as described above.

Protease Digestion of ApoLp-III—ApoLp-III, 0.5 mg, was dissolved

* This work was supported by National Institutes of Health Grants HD10954, GM29238, and DK30292. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‖ An Established Investigator of the American Heart Association.

[1] The abbreviations used are: apoLp-I, apolipophorin-I; apoLp-II, apolipophorin-II; apoLp-III, apolipophorin-III; HPLC, high performance liquid chromatography; SDS, sodium dodecyl sulfate.

in 200 μl of 100 mM ammonium bicarbonate and digested as follows: *Staphylococcus aureus* V8 protease (Miles Scientific, Naperville, IL): 25 μl of a 1 mg/ml solution in water was added, the mixture incubated for 16 h at 37 °C, the solution lyophilized, and the residue subjected to HPLC. *Thermolysin* (Behring Diagnostics): 10 μl of a 1 mg/ml solution was added, the mixture incubated for 2 h at 25 °C, and treated as described for the V8 protease. *Clostripain* (Sigma): a 2 mg/ml solution of the enzyme was prepared in 1 mM $CaCl_2$ and 2.5 mM dithiothreitol and left overnight at 5 °C to allow activation of the enzyme, after which it was diluted to 0.5 mg/ml with the above activating solution. Two mg of apoLp-III were dissolved in 500 μl of 50 mM sodium phosphate buffer, pH 7.6, 80 μl of the enzyme solution added, and the mixture incubated at 25 °C for 10 min. The reaction was stopped by adding 1 drop of glacial acetic acid. After dilution to 10 ml with 0.25% trifluoroacetic acid, the mixture was separated by HPLC. *Pyroglutaminase aminopeptidase* (Sigma): the peptide (1.5 mg) was dissolved in 800 μl of 100 mM ammonium bicarbonate, 1 mM EDTA, and 10 mM 2-mercaptoethanol. After addition of 50 μg of enzyme, the reaction was allowed to proceed at 37 °C for 4 h. The sample was lyophilized, and the product was purified by HPLC.

RNA Preparation and Translation—Fat body tissue from newly emerged adults was removed, rinsed with ice-cold phosphate-buffered saline (0.1 M sodium phosphate, pH 7.0, 0.15 M NaCl), quick frozen, and stored at −70 °C until use. RNA was prepared by the guanidine thiocyanate method of Chirgwin *et al.* (1979). Polyadenylated RNA was isolated using oligo(dT) (Collaborative Research, Inc., Lexington, MA, Type 3) and the procedure of Aviv and Leder (1972). Polyadenylated RNA was translated in the wheatgerm system (Anderson *et al.*, 1983) using [^{35}S]methionine, specific activity = 1000 Ci/mmol (New England Nuclear). The translation mixture in 25 μl, containing 5 μCi of labeled methionine, was incubated for 1 h at 30 °C, and then diluted with 75 μl of phosphate-buffered saline and 50 μl of normal rabbit serum. After 1 h at 37 °C, 25 μl of Pansorbin (Behring Diagnostics) was added and the incubation continued for 1 h at 4 °C. The mixture was centrifuged, and 50 μl of anti-apoLp-III antiserum was added to the supernatant. After incubation overnight at 4 °C, 25 μl of Pansorbin was added, and the mixture was incubated for 1 h at 4 °C. After centrifugation, the pellet was washed three times with ice-cold 150 mM NaCl, 50 mM Tris, pH 7.4, 5 mM EDTA, and 0.05% Nonidet P-40. The pellet was washed once with ice-cold phosphate-buffered saline and resuspended in SDS sample buffer.

SDS-Polyacrylamide Gel Electrophoresis—SDS-polyacrylamide gel electrophoresis was carried out in 12.5% polyacrylamide gels (Laemmli, 1970). Fluorography was carried out as described by Prasad *et al.* (1986b).

cDNA Library Construction and Screening—Polyadenylated RNA from the fat body of newly emerged adult males was used to prepare cDNA using a commercial kit (cDNA Synthesis System, Amersham Corp.). *Eco*RI linkers (Bethesda Research Laboratories), were added, the cDNA was size-fractionated and ligated to phosphorylated λ gt11 arms (Promega Biotech, Madison, WI) as described by Huynh *et al.* (1985). The DNA was then packaged (Promega Biotech, Madison, WI), used to infect *Escherichia coli* Y-1090, plated, and transferred to nitrocellulose according to Huynh *et al.* (1985). The filters were blocked using nonfat milk (Johnson *et al.*, 1984) and incubated with anti-apoLp-III antiserum, followed by ^{125}I-protein A (ICN Radiochemicals, Irvine, CA). Filters were placed at −70 °C with Kodak X-Omat AR film with a Cronex Lighting Plus intensifying screen. Ten thousand recombinant plaques yielded 6 positives that were purified to homogeneity. DNA was prepared by the method of Benson and Taylor (1984) and the *Eco*RI inserts subcloned into pUC-8 and transformed into *E. coli* JM 83.

Blotting and Hybridization—A 20-mer synthetic oligonucleotide was prepared complimentary to the codons for amino acids 9–15 established from the amino acid sequence of apoLp-III. This nucleotide had the sequence T-G-(T/C)-T-T-(T/C)-T-C-C-A-T-(T/C)-T-C-(T/C)-T-C-(A/G)-A-A and was labeled using [^{32}P]ATP and T4 polynucleotide kinase (Bethesda Research Laboratories). A Southern transfer (Southern, 1975) of the six positive clones was hybridized to the labeled probe in 0.9 M NaCl, 6 mM EDTA, 90 mM Tris, pH 8.0 (6 × NET), 0.2% Ficoll, 0.2% polyvinylpyrollidone, and 0.2% bovine serum albumin (10 × Denhardt's solution), 0.1% SDS, and 100 μg/ml of heat-denatured salmon sperm DNA at 45 °C overnight (Berent *et al.*, 1985). The filters were washed four times at room temperature in 6 × NET, 0.1% SDS, and once at 45 °C with each wash lasting 15 min.

RNA was fractionated in formaldehyde gels and transferred to nitrocellulose (Maniatis *et al.*, 1982). Hybridization with the nick-translated 708-base pair *Eco*RI insert was done in 0.6 M NaCl, 0.08 M Tris, pH 7.8, 4 mM EDTA (4 × SET), 10 × Denhardt's, 0.1% SDS, and 0.1% sodium pyrophosphate at 65 °C overnight. The final wash was with 0.5 × SET, 0.1% SDS, and 0.1% sodium pyrophosphate at 65 °C for 1 h.

DNA Sequencing—The 708-base pair *Eco*RI inserts and restriction fragments were inserted into either M13, mp 18, or mp 19 DNA and transformed into *E. coli* JM 101 (Yanisch-Perron *et al.*, 1985). The single-stranded DNA was then sequenced using the dideoxy chain termination method of Sanger *et al.* (1977).

Protein and Peptide Sequence Analysis—Intact apoLp-III or peptides derived from it (5 nmol) were sequenced by automated Edman degradation (Edman and Begg, 1967), using a Beckman 890M instrument (Beckman Instruments). Polybrene was added to peptide solutions (Tarr *et al.*, 1978) and the Beckman Quadrol program 05-22-85 was used and runs were 8–62 cycles. Phenylthiohydantoin derivatives were analyzed by HPLC using a Beckman 110 system, a C-18 reversed-phase column, and a linear gradient consisting of *a*) 10% acetonitrile with 0.02 M sodium acetate and *b*) 100% acetonitrile. Repetitive yields were ≥95%.

Computer-assisted Analysis of Sequence Data—RELATE and all other programs used in this study are described in detail in Boguski *et al.* (1986a). A MicroVax II (Digital Equipment Corp.) running in the VMS operating system (Version 4.4) was used for all computations.

RESULTS AND DISCUSSION

Nucleotide Sequence of ApoLp-III cDNA—Positive clones were identified on the basis that they produced a fusion protein which cross-reacted with anti-apoLp-III antibody and contained an *Eco*RI insert which hybridized with a specific oligonucleotide probe. All six clones contained an insert of approximately 700 base pairs. The strategy used for sequencing the cDNA insert is shown in Fig. 1 and the nucleotide and deduced amino acid sequence in Fig. 2 (the protein is numbered beginning with the chain initiating Met). The 708-base pair sequence, which includes the 8-base pair *Eco*RI linker additions at both ends, has one large open reading frame beginning with an ATG codon at position 43 and extending to position 609. This sequence codes for a protein with 189 residues, including the chain initiating Met. Although the poly(A) tail was lost during cloning, the consensus poly(A) addition signal, AATAAA, is found beginning at position 684. The 5 nucleotides upstream from the ATG have a sequence TCACT, which is similar to the proposed consensus eukaryotic initiation site (CCACC) (Kozak, 1984) except for substitution of T for C.

Confirmation of Nucleotide Sequence by Protein Sequencing—In Fig. 2 the residues which were confirmed by protein sequencing are *capitalized* and *underlined* and were determined as follows: residues 1–40, the intact protein; residues 13–52, the largest CNBr peptide after treatment with pyro-

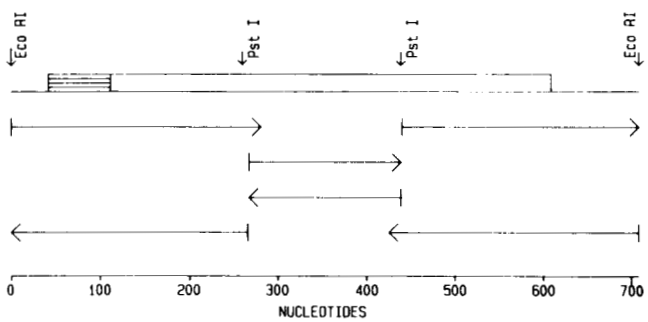

FIG. 1. **Sequencing strategy for apolipophorin-III.** The *box* indicates the location of the coding region in the 708 base pair cDNA. The *shaded area* indicates the leader sequence. The site of restriction enzyme cleavage is indicated by the *vertical arrows* and the extent and direction of sequencing of the various subclones is indicated by the *horizontal arrows*.

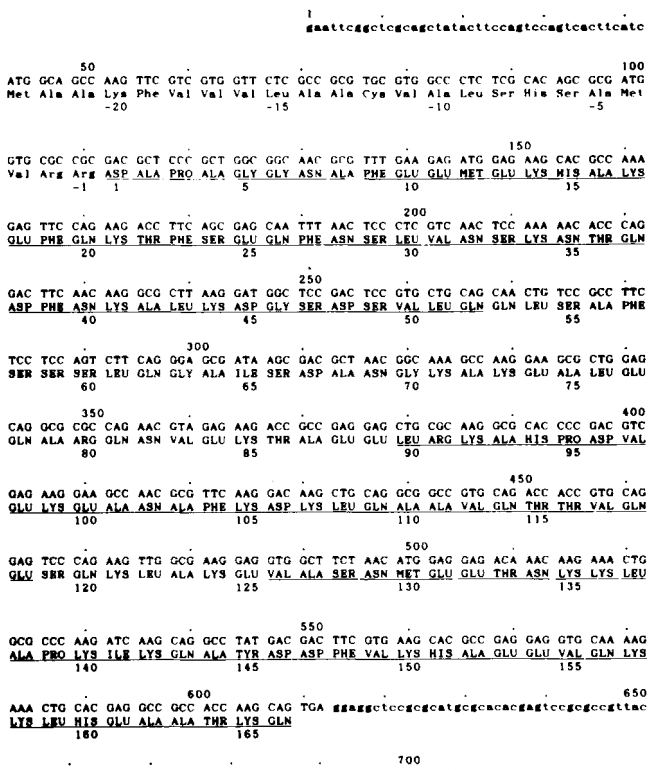

FIG. 2. **cDNA and amino acid sequence of apolipophorin-III.** The numbering of the amino acid sequence is as follows: amino acids in the mature hemolymph protein are capitalized and assigned positive numbers, whereas those in the signal peptide are in *lower case* and assigned negative numbers. Those portions of the sequence *underlined* were confirmed by amino acid sequencing. Bases in the coding region are capitalized.

glutaminase; residues 90–98, a V8 peptide; residues 92–107, a clostripain peptide; residues 101–118, a V8 peptide; residues 126–133, a thermolysin peptide; residues 131–164, a CNBr peptide; residues 156–166, a thermolysin peptide.

Evidence for a Precursor—When the immunoprecipitated product produced in the *in vitro* translation system was analyzed by SDS-polyacrylamide gel electrophoresis, it was found to have a molecular weight about 2,000 greater than the hemolymph protein. This difference corresponds to the difference between the nucleotide-derived sequence (M_r = 20,663) and the isolated protein (M_r = 18,380).

We have attempted to predict the site of cotranslational proteolytic processing in this secreted protein using the empirical method of Von Heijne (1983, 1985). The results presented in Table I include (in descending order of probability) cleavage after Ser (−8), Ser (−6), or Ala (−10). Given the fact that the Asp at position 24 of the primary translation product forms the amino-terminal residue of the mature protein, the implication of this analysis is that apoLp-III undergoes both co- and post-translational proteolytic processing. The presence of two arginines just proximal to Asp-1 is compatible with the predominant structural feature of eukaryotic prosegments (Steiner *et al.*, 1980). If cotranslational cleavage occurs at Ser (−6), an amino-terminal pentatpeptide prosegment (A-M-V-R-R) would be defined which bears striking sequence similarity to the pentapeptide prosegment of human apoAII (A-L-V-R-R). Further studies are needed to verify these suggestions.

Comparative Sequence Analysis—Definition of the primary structure of an insect apolipoprotein now provides an opportunity to examine the development of lipid-binding activity

TABLE I

Processing probability analysis of the amino-terminal 28 residues of the primary translation product of apoLp-III

The weight matrix analysis technique described by Von Heijne (1985) was applied to generate the values shown. The accuracy of this predictive method is 75–80%.

Cleavage position	Probability
Between −11, −10	−7.7
Between −10, −9	2.0
Between −9, −8	−0.6
Between −8, −7	2.9
Between −7, −6	−3.0
Between −6, −5	2.6
Between −5, −4	1.5
Between −4, −3	−6.7
Between −3, −2	0.4
Between −2, −1	−9.3
Between −1, 1	−2.5
Between 1, 2	−6.6
Between 2, 3	−9.3

over more than 500 million years of metazoan evolution.

As a first step in analyzing the sequence of apoLp-III, we used the computer program RELATE to determine if apoLp-III bore any similarity to two well-known (but unrelated) families of vertebrate lipid-binding sequences: (i) the human apolipoproteins (plus chicken apoVLDL-II) and (ii) a group of fatty acid/retinol-binding proteins from several mammalian species (the relationships between members of this latter family are reviewed in Sacchettini *et al.*, 1986). Briefly, the RELATE program was designed to detect statistically significant similarities among a group of sequences and to assign a value (SD score) for the degree of similarity between two sequences. Scores of >3.0 SD units are generally considered to indicate a significant relationship (Dayhoff *et al.*, 1983; see Table II).

The results of the RELATE analysis for apoLp-III compared with the eight human proteins (apoA-I, A-II, A-IV, B, C-I, C-II, C-III, E) plus avian apoVLDL-II are shown in Table II. The SD scores ranged from 1.139 for apoLP-III and apoVLDL-II to 8.754 for apoLp-III and apoE. The sequence of apoLp-III thus bears a remarkable resemblance to several human apolipoproteins including apoE, A-IV, A-I and C-I. It is interesting to note that the SD scores obtained from comparing apoLp-III with these sequences were considerably higher than several of the SD scores for comparisons of the human apolipoproteins with each other. In contrast, apoLp-III appeared to have no significant homology to the intra- or extracellular fatty acid and retinol-binding proteins listed in Table II.

The human apolipoproteins (particular A-I, A-IV, and E) are composed of tandemly repeated sequences that are multiples of 11 amino acids (Karathanasis *et al.*, 1983; Das *et al.*, 1985; Paik *et al.*, 1985; Boguski *et al.*, 1986a; Karathanasis *et al.*, 1986), so we next examined the structure of apoLp-III for evidence of internal homology or sequence periodicity. This was done using comparison matrix analysis (reviewed in Boguski *et al.*, 1986a). ApoLp-III was first compared against itself (an intrasequence comparison) and then against the sequence of a human apolipoprotein (an intersequence comparison matrix).

Fig. 3 displays intrasequence comparison matrices for apoLp-III at two different plotting thresholds. Inspection of Fig. 3*A* reveals the presence of multiple regions of internal homology as evidenced by the numerous, short diagonal lines offset from, but parallel to, the main diagonal. The main diagonal represents a colinear alignment (one-to-one correspondence) of the apoLp-III sequence with itself. As the

TABLE II

Intersequence comparisons using the RELATE algorithm

The RELATE algorithm was used to compare the 166 residue mature apoLp-III sequence to the other protein sequences listed. Segment comparison scores (expressed in SD units) were generated using spans of 17 residues and the mutation scoring system (Dayhoff *et al.*, 1983). SD units are computed by first noting the differences in the mean score obtained for the real sequence comparisons and the mean score from multiple (in this case 100) comparisons of randomly shuffled sequences having the same amino acid composition as the real sequence. This difference is then divided by the standard deviation of the scores from the random shuffle. The relationship between SD units and the probability (P) of achieving the SD score by chance is as follows: SD = 1.0, P = 0.159; SD = 2.0, P = 0.227×10^{-1}; SD = 3.0, P = 0.135×10^{-2}; SD = 4.0, P = 0.317×10^{-4}; SD = 5.0, P = 0.287×10^{-6}; SD = 6.0, P = 0.987×10^{-9}; SD = 7.0, P = 0.128×10^{-11}; SD = 8.0, P = 0.622×10^{-15}; SD = 9.0, P = 0.133×10^{-18}; SD = 10.0, P = 0.762×10^{-23}.

	apoLp-III	A-IV	A-I	A-II	C-I	C-II	C-III
Human preapoA-IV "LPHUA4"[a]	8.204						
Human preproapoA-I "LPHUA1"	6.774	23.080					
Human preproapoA-II "LPHUA2"	3.594	3.378	5.175				
Human preapoC-I "LPHUC1"	5.474	4.548	3.861	5.912			
Human preapoC-II "LPHUC2"	1.191	2.166	3.042	3.155	4.699		
Human preapoC-III "LPHUC3"	2.611	1.585	3.103	3.861	6.290	6.805	
Human preapoE "LPHUE"	8.754	19.552	13.628	3.085	4.938	1.274	2.497
Human preapoB "LPHUB"	2.820						
Chicken preapoVLDL-II "VLCH1"	1.139	2.104	1.725	1.403	1.632	−0.005[b]	2.015
Human preproalbumin "ABHUS"	2.393						
Rat liver FABP "FZRTL"	−0.048						
Rat intestinal FABP "FZRTI"	0.025						
Rat heart FABP "FZRTH"	1.122						
Human serum retinol binding protein "VAHU"	−0.244						
Rat CRBP "RJRTO"	1.824						
Bovine CRABP "RJBOA"	2.332						

[a] National Biomedical Research Foundation (NBRF) Protein Identification Resources (PIR) Database retrieval key codes.

[b] A negative SD score indicates that the segment comparison scores generated from the randomly shuffled sequences were greater than those obtained with the real sequences.

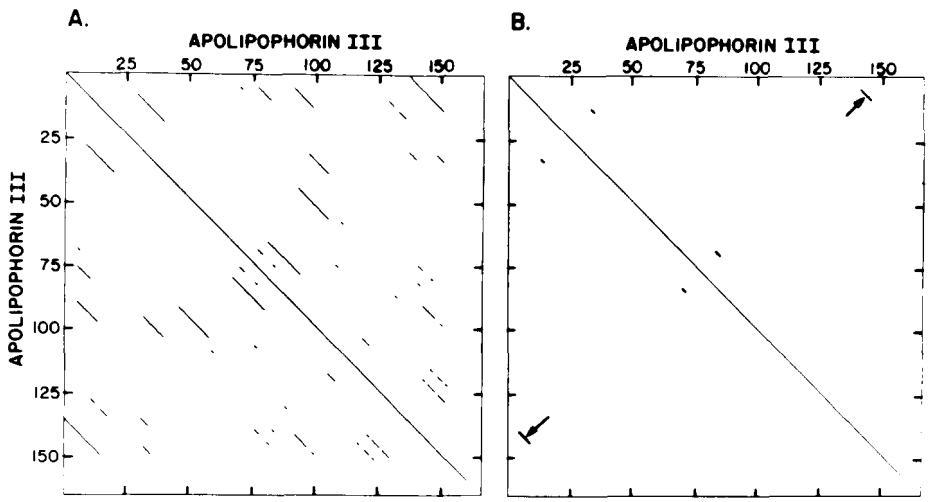

FIG. 3. **Intrasequence comparison matrix of apolipophorin-III.** The 166-residue mature hemolymph apoLp-III polypeptide was compared against itself using a span of 17 residues and the CMPSEQ84 algorithm of McLachlan (1983). Selection of a span length is somewhat arbitrary. However, the length should generally be less than 10% of the length of the intact protein and greater than the size of the fundamental repeat (Boguski *et al.*, 1986a). Matching scores generated for each set of segments that are compared represent the sum of similarity scores for each aligned pair of elements in the span. The mutation data matrix (250 PAMs, Dayhoff *et al.*, 1983) was used to score segment comparisons. If a matching score for two spans exceeds a predetermined value, a point corresponding to the center of the span is plotted in a two-dimensional array. In *panel A*, the threshold score had a probability of occurring by chance alone of less than 1 in 1,000. In *panel B*, the plotting threshold was raised to 1 in 100,000.

plotting threshold is raised (Fig. 3*B*), regions with relatively greater degrees of homology are defined. In this manner, we identified the two subsequences (residues 9–22 and 152–165) of apoLp-III that are most alike (indicated by the *arrows* in Fig. 3*B*).

It is often difficult to determine the precise length of a repeating unit from a comparison matrix alone, especially if there is considerable length variability among the repeats. Thus we once again employed the RELATE program in order to define better the periodicity in apoLp-III. In generating an SD score, RELATE compares all overlapping subsequenes of a user-specified length and ranks their relative similarities. The program reports the positions of the first residues of all pairs of subsequences and the distances between them (displacements). When sequences are composed of multiple tandem repeats, the displacements of the highest scoring seg-

ments tend to be multiples of the repeat length.

Table III shows displacements of the top 76 highest scoring segments for an intrasequence comparison of apoLp-III. Several conclusions can be drawn from an analysis of these data. First, the length of the repeat unit in apoLp-III appears to be 14 residues. Second, this repeat unit length is not very well preserved in the periodic structure of the protein because the longer displacements are not exact multiples of 14 (compare this table with Table III in Boguski *et al.* (1986b) which shows that the displacements in apoAIV are multiples of 11 residues). Finally, the second most frequent displacement value corresponds approximately to the distance between residues in the two most homologous subsequences of apoLp-III (residues 9–22 and 152–165 as described above). To define more precisely the 14-residue (tetradecapeptide) repeating unit in apoLp-III, we first aligned residues 9–22 and 152–165 as follows to act as landmarks for the arrangement the intervening residues (23–151).

```
 9                    22
 F E E M E K H A K E F Q K T
 . : : . . : . . . : . . . .
 A E E V Q K K L H E A A T K
152                   165
```

The symbol ":" indicates a sequence identity whereas "." signifies a conservative amino acid substitution according to the mutation data matrix (Dayhoff *et al.*, 1983). Next, the amino acid residues of apoLp-III were coded according to hydropathy index and charge (Boguski *et al.*, 1984) to aid in the recognition of conserved physical-chemical properties. Finally, the entire apoLp-III sequence was arranged in blocks consisting, whenever possible, of 14 residues. These blocks were aligned with the landmark sequences based upon conserved structural features as described below.

The repeated sequences of apoLp-III are displayed in Fig. 4. There are approximately twelve repeat units with a considerable degree of length variability. (Residues 1–8 would appear to represent a highly degenerate repeat remnant.) The repeats range in length from 7 to 16 residues, although most consist of 14 or 15 residues. The first 8 residues of each unit are highly conserved with respect to relative hydropathy and/or charge. The repeating motif among these residues is as follows.

```
hydrophobic-acidic-acidic-hydrophobic-
    1        2     3        4

                  hydrophilic-basic-basic-hydrophobic
                       5        6     7        8
```

Residues in the first four positions are most highly conserved (Fig. 4). In positions 1 and 4 of the repeats, 75% of the amino acid residues have hydrophobic side chains. In positions 2 and 3, an impressive 92% of the residues are either acidic amino acids or their amide derivatives. Of the nonconservative substitutions within the first four positions, most are replacements by the small, neutral amino acids glycine, serine, and threonine. Of the 36 residues that comprise positions 5-6-7, 17 residues (47%) are basic with the remainder being equally represented by acidic/amide and hydrophobic residues. Still, this region, considered as a whole, is predominantly basic/hydrophilic in character.

Beyond the initial 8 residues, there is some indication of a more weakly conserved tripeptide sequence with a basic acidic hydrophobic motif. Positions 12-13-14 of the repeat units are more highly variable and only remnants of conserved properties can be discerned. Thus what appears to have been at one time a fundamental tetradecapeptide repeating unit has undergone a combination of insertions, deletions, and other mutational changes. The evolutionary and functional significance of these changes is as yet unknown. It is possible to conclude, however, that natural selection has resulted in a greater degree of sequence conservation among the amino-terminal domains of the repeat units than among their carboxyl-terminal regions.

Based on hydropathy profile shown in Fig. 5*A*, an amphiphilic pattern that alternates regularly between hydrophobicity and hydrophilicity is clearly evident throughout the entire sequence of apoLp-III. Furthermore, prediction of the secondary structure of apoLp-III, using Chou-Fasman rules (Chou and Fasman, 1978), indicates a considerable fraction of the sequence (63%) may exist in alpha-helical conformation (Fig. 5*B*). This prediction agrees well with the value determined by circular dichroism (Kawooya *et al.*, 1986). Thus the paradigm of amphiphilic, helical, lipid-binding domains that has been well established for the mammalian apolipoproteins also may be the structural basis for lipid-binding activity in apolipophorin-III.

What is the precise relationship of the amphiphilic repeat units of apoLp-III to the fundamental undecapeptide repeating unit of the mammalian apolipoproteins? In order to map homologous segments, an intersequence comparison matrix between apoLp-III and human apoA-IV was computed (Fig. 6). ApoA-IV was used because it contains the greatest number of most highly conserved repeats in the apolipoprotein family (Boguski *et al.*, 1984, Elshourbagy *et al.*, 1986; Karathanasis *et al.*, 1986). However the results were essentially the same when apoE was used (data not shown).

Fig. 6 demonstrates that apoLp-III and apoA-IV share many short regions of considerable sequence similarity as evidenced by the numerous diagonals extending throughout the length of both sequences. However, the absence of a main diagonal indicates that apoLp-III and apoA-IV are *not* colinearly related as would be predicted based on the fact that the fundamental repeating units of these two proteins are of different length.

Visual comparisons of the respective repeats in apoLp-III and apolipoproteins A-I, A-IV, and E (Boguski *et al.*, 1984, 1986a) revealed that residues 1–8 of the apoLp-III tetradeca-

TABLE III

Intrasequence comparison of apoLp-III using RELATE

The frequencies of the top 76 scores with the same displacement are shown. The program RELATE was used to compare the 166 residue mature apoLp-III against itself. A total of 11,175 segments of 17 residues were compared. The scores ranged from −35 to +34 with a mean of −6 and a standard deviation of 11 (for details of how these scores were computed see Dayhoff *et al.*, 1983).

Displacement	Frequency	Average score
−14	15	30.27
−136	11	33.55
−20	9	29.33
−47	9	27.44
−85	8	27.00
−71	5	26.20
−65	5	27.20
−118	3	26.00
−51	3	25.33
−116	2	26.50
−29	2	27.00
−106	2	27.00
−58	1	25.00
−44	1	25.00
−22	1	25.00

Repeat No.	Residue No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
	1							D	A	P	A	G	G	N	A	
I	9	F	E	E	M	E	K	H	A	K	E	F	Q	K	T	F
II	24	S	E	Q	F	N	S	L	V	N	S	K	N			
III	36	T	Q	D	F	N	K	A	L	K	D	G	S	D	S	V
IV	51	L	Q	Q	L	S	A	F	S	S	S	L	Q	G	A	I S
V	67	D	A	N	G	K	A	K	E	A						
VI	76	L	E	Q	A	R	Q	N	V	E	K	T				
VII	87	A	E	E	L	R	K	A	H	P	D	V	E	K	E	
VIII	101	A	N	A	F	K	D	K	L	Q	A	A	V	Q	T	T
IX	116	V	Q	E	S	Q	K	L	A	K	E	V	A	S	N	
X	130	M	E	E	T	N	K	K	L	A	P	K	I	K	Q	A
XI	145	Y	D	D	F	V	K	H								
XII	152	A	E	E	V	Q	K	K	L	H	E	A	A	T	K	Q

FIG. 4. **Alignment of repeated sequences in apolipophorin-III.** This alignment was arrived at as described in the text. Numbers along the left margin represent residue numbers for the mature hemolymph protein. Amino acids have been assigned to groups based on their hydropathy index (Kyte and Doolittle, 1982) and charge as described by Boguski *et al.* (1984). Hydrophobic residues are represented in *green*; acidic residues and their amides in *red*; basic residues in *blue*; and glycine, serine, and threonine, which have hydropathy values near zero, in *black*. Proline residues, which occupy the first position of most of the docosapeptide repeats in the human apolipoproteins AI, AIV, and E, have been colored *yellow* to emphasize their unique structural significance.

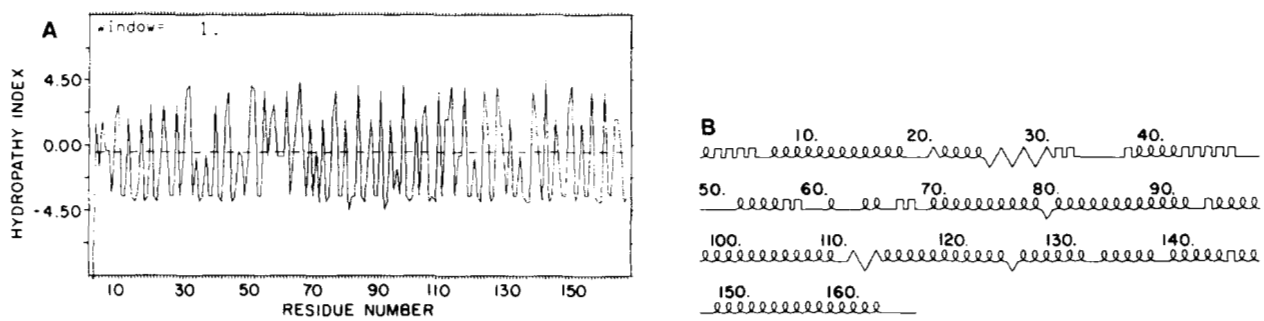

FIG. 5. *A*, hydropathy profile of apolipophorin-III according to Kyte and Doolittle (1982). Hydrophobic amino acids have positive hydropathy indices and hydrophilic residues are represented by negative values. *B*, secondary structure prediction for apolipophorin-III. ⌒, α helix; Λ, β sheet; Π, β bend. A *straight line* indicated that no prediction could be made.

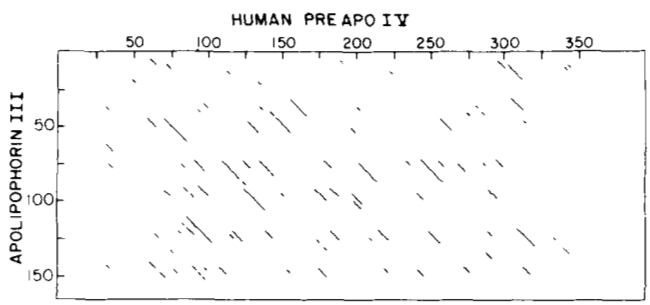

FIG. 6. **Intersequence comparison matrix of *M. sexta* apolipophorin-III and human preaproA–IV.** A span of 17 was used and the threshold score for plotting was set such that the probability of achieving that score by chance alone was less than 1 in 1000.

peptides and residues 3–10 of the apolipoprotein undecapeptides share the same pattern of conserved amino acid residues. In contrast, sequences flanking this 8-residue core region are entirely dissimilar. For example, most of the group A human apolipoprotein undecapeptide repeat units are punctuated by proline residues in position 1. However, of the 2 or 3 proline residues that are found among the repeats in apoLp-III, none occur in positions comparable to position 1 of the human apolipoprotein repeats. Nor do positions 9–14 of the apoLp-III repeats have any convincing counterparts among apolipoprotein sequences.

It is not possible at the present time to determine if the sequence relationships between *M. sexta* apolipophorin-III and the mammalian apolipoproteins represent divergence from a common ancestral sequence. Despite highly significant SD scores and similar core amino acid patterns, the existence of different repeat unit lengths and their lack of colinearity argue against a common ancestor. Indeed, the SD values seem paradoxically high given the vast evolutionary distance between insects and mammals and the fact that other studies have shown the apolipoproteins to evolve very rapidly (Boguski *et al.*, 1986c, Luo *et al.*, 1986). Consistent with this is the fact that whereas the repeats contain very similar types of amino acid residues, the degree of sequence *identity* is actually quite low.

On the other hand, might we reasonably expect that apoLp-III and the human apolipoproteins have acquired such a great degree of similarity by convergent evolution? In other words, is the amphiphilic helix such a useful structure for lipid binding that nature invented it more than once? Although this may seem unlikely at first glance, the amphiphilic helix is widely distributed in nature as a surface covering in the globular proteins (Doolittle, 1981). Thus it is not hard to conceive that natural selection could create the same type of

mechanism for lipid binding independently in insects and vertebrates.

Acknowledgment—We thank Rodney Folz for his assistance in predicting the site of cotranslational processing of apoLp-III.

REFERENCES

Anderson, C. W., Straus, J. W., and Dudock, B. S. (1983) *Methods Enzymol.* **101,** 635–644

Aviv, H., and Leder, P. (1972) *Proc. Natl. Acad. Sci. U. S. A. 69,* 1408–1412

Benson, S. A., and Taylor, R. K. (1984) *Biotechniques* **2,** 126–127

Berent S. L., Mahmoudi, M., Torczynski, R. M., Bragg, P. W., and Bollon, A. P. (1985) *Biotechniques* **3,** 208–220

Boguski, M. S., Elshourbagy, N., Taylor, J. M., and Gordon, J. I. (1984) *Proc. Natl. Acad. Sci. U. S. A.* **81,** 5021–5025

Boguski, M. S., Freeman, M., Taylor, J. M., Elshourbagy, N., and Gordon, J. I. (1986a) *J. Lipid Res.* **27,** 1011–1034

Boguski, M. S., Elshourbagy, N., Taylor, J. M., and Gordon, J. I. (1986b) *Methods Enzymol.* **128,** 753–773

Boguski, M. S., Birkenmeier, E. H., Elshourbagy, N. A., Taylor, J. M., and Gordon, J. I. (1986c) *J. Biol. Chem.* **261,** 6398–6407

Chino, H., Downer, R. G. H., Wyatt, G. R., and Gilbert, L. I. (1981) *Insect Biochem.* **11,** 491

Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J., and Rutter, W. J. (1979) *Biochemistry* **18,** 5294–5299

Chou, P. Y., and Fasman, G. D. (1978) *Annu. Rev. Biochem.* **47,** 251–276

Das, H. K., McPherson, J., Bruns, G. A. P., Karathanasis, S. K., and Breslow, J. L. (1985) *J. Biol. Chem.* **260,** 6240–6247

Dayhoff, M. O., Barker, W. C., and Hunt, L. T. (1983) *Methods Enzymol.* **91,** 524–545

Doolittle, R. F. (1981) *Science* **214,** 149–159

Elshourbagy, N. A., Walker, D. W., Boguski, M. S., Gordon, J. I., and Taylor, J. M. (1986) *J. Biol. Chem.* **261,** 1998–2002

Edman, P., and Begg, G. (1967) *Eur. J. Biochem.* **1,** 80–91

Huynh, T. V., Young, R. A., and Davis, R. W. (1985) in *DNA Cloning* (Glover, D. M., ed) Vol. 1 pp. 49–78, IRL Press, Oxford

Johnson, D. A., Gautsch, J. W., Sportsman, J. R., and Elder, J. H. (1984) *Gene Anal. Tech.* **1,** 3–8

Karathanasis, S. K., Zannis, V. I., and Breslow, J. L. (1983) *Proc. Natl. Acad. Sci. U. S. A.* **80,** 6147–6151

Karathanasis, S. K., Yunis, I., and Zannis, V. I. (1986) *Biochemistry* **25,** 3962–3970

Kawooya, J. K., Keim, P. S., Ryan, R. O., Shapiro, J. P., Samaraweera, P., and Law, J. H. (1984) *J. Biol. Chem.* **259,** 10733–10737

Kawooya, J. K., Meredith, S. C., Wells, M. A., Kézdy, F. J., and Law, J. H. (1986) *J. Biol. Chem.* **261,** 13588–13591

Kosak, M. (1984) *Nucleic Acids Res.* **12,** 857–873

Kyte, J., and Doolittle, R. F. (1982) *J. Mol. Biol.* **157,** 105–132

Laemmli, U. K. (1970) *Nature* **227,** 680–685

Luo, C-C., Li, W-H., Moore, M. N., and Chan, L. (1986) *J. Mol. Biol.* **187,** 325–340

Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982) *Molecular Cloning, A Laboratory Manual,* Cold Spring Harbor Laboratory, Cold Spring Harbor, New York

McLachlan, A. D. (1983) *J. Mol. Biol.* **169,** 15–30

Paik, Y-K., Chang, D. J., Reardon, C. A., Davies, G. E., Mahley, R. W., and Taylor, J. M. (1985) *Proc. Natl. Acad. Sci. U. S. A.* **82,** 3445–3449

Pattnaik, N. M., Mundall. E. C., Trambusti, B. G., Law, J. H., and Kezdy, F. J. (1979) *Comp. Biochem. Physiol.* **63B,** 469–476

Prasad, S. V., Ryan, R. O., Law, J. H., and Wells, M. A. (1986a) *J. Biol. Chem.* **261,** 558–562

Prasad, S. V., Fernando-Warnakulasuriya, G. J. P., Sumida, M., Law, J. H., and Wells, M. A. (1986b) *J. Biol. Chem.* **261,** 17174–17176

Ryan, R. O., and Law, J. H. (1984) *Bio-Essays* **1,** 250–252

Ryan, R. O., Prasad, S. V., Henriksen, E. J., Wells, M. A., and Law, J. H. (1986) *J. Biol. Chem.* **261,** 563–568

Sacchettini, J. C., Said, B., Schulz, H., and Gordon, J. I. (1986) *J. Biol. Chem.* **261,** 8218–8223

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

Shapiro, J. P., and Law, J. H. (1983) *Biochem. Biophys. Res. Commun.* **136,** 260–265

Shapiro, J. P., Keim, P. S., and Law, J. H. (1984) *J. Biol. Chem.* **259,** 3680–3685

Southern, E. M. (1975) *J. Mol. Biol.* **98,** 503–517

Steiner, D. F., Quinn, P. S., Chan, S. J., Marsh, J., and Tager, H. S. (1980) *Ann. N. Y. Acad. Sci.* **343,** 1–16

Tarr, G. E., Beecher, J. F., Bell, M., and McKean, D. J. (1978) *Anal. Biochem.* **84,** 622–627

Von Heijne, G. (1983) *Eur. J. Biochem.* **133,** 17–21

Von Heijne, G. (1985) *Nucleic Acid Res.* **14,** 4683–4690

Wells, M. A., Ryan, R. O., Prasad, S. V., and Law, J. H. (1985) *Insect Biochem.* **15,** 565–571

Wells, M. A., Ryan, R. O., Kawooya, J. K., and Law, J. H. (1987) *J. Biol. Chem.* **262,** 4172–4176

Yanisch-Perron, C., Vieira, J., and Messing, J. (1985) *Gene (Amst.)* **33,** 103–119