Primary Structure and Comparative Sequence Analysis of an Insect Apolipoprotein

APOLIPOPHORIN-III FROM MANDUCA SEXTA*

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 mammalian 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 cannot 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, serves as a recycling shuttle that carries fatty acids in the form of diacylglycerol, and in addition cholesterol, carotenoids, 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, apolipopophin-I (apoLp-I) \( M_r = 240,000 \), and apolipopophin-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., 1986).

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 \( \times \) 250 mm). The protein was eluted using a linear gradient formed between 0.2% 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).

Cyanoxy 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
translated 708-base pair EcoRI insert was done in 0.6 M NaCl, 0.08 M Tris, pH 7.8, 4 mM EDTA (4 X SET), 10 X Denhardt's, 0.1% SDS, and 0.1% sodium pyrophosphate at 65°C overnight. The final wash was with 0.5 X SET, 0.1% SDS, and 0.1% sodium pyrophosphate at 65°C for 1 h.

**Sequence Analysis** — The 708-base pair EcoRI inserts and restriction fragments were inserted into either M13, mp 18, or rap 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 Quadruplet program 05-22-85 was used and runs were 8-2 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%.

### 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 EcoRI 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 EcoRI linker additions at both ends, has one large open reading frame starting in Fig. 2 the residues which were confirmed by protein sequencing are indicated by a vertical arrow and underlined and were determined as follows: residues 1-40, the intact protein; residues 13-52, the largest CNBr peptide after treatment with pyro-
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 (Duyhoff et al., 1988; 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-I11 and apoE to 8.754 for apoLP-I11 and apuE. 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., 1986; Boguski et al., 1986a; Karathanasis 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 (Duyhoff et al., 1988; see Table II).
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:

- \( SD = 1.0, \ P = 0.159 \)
- \( SD = 2.0, \ P = 0.227 \times 10^{-4} \)
- \( SD = 3.0, \ P = 0.135 \times 10^{-4} \)
- \( SD = 4.0, \ P = 0.317 \times 10^{-4} \)
- \( SD = 5.0, \ P = 0.287 \times 10^{-4} \)
- \( SD = 6.0, \ P = 0.987 \times 10^{-4} \)
- \( SD = 7.0, \ P = 0.128 \times 10^{-3} \)
- \( SD = 8.0, \ P = 0.622 \times 10^{-4} \)
- \( SD = 9.0, \ P = 0.133 \times 10^{-4} \)
- \( SD = 10.0, \ P = 0.762 \times 10^{-5} \)

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

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 apolipoporphin-III.](image-url)

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 subsequences 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).

\[
\begin{align*}
9 & \quad 22 \\
\begin{array}{cccccccc}
F & E & E & M & E & K & A & E & F & Q & K & T \\
A & E & V & Q & K & K & L & H & E & A & A & T & K \\
152 & 165
\end{array}
\end{align*}
\]

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 the hydrophathy 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 hydrophy and/or charge. The repeating motif among these residues is as follows.

- hydrophobic-acidic-acidic-hydrophobic-
- hydrophilic-basic-basic-hydrophobic

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 12-15-14 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 hydrophathy profile shown in Fig. 5A, an amphiphilic pattern that alternates regularly between hydrophobic and hydrophilic 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. 5B). 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 apolipoprotein-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-
peptides and residues 3-10 of the apolipoprotein undecapeptide 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

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