Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 4983–4986, August 1985
Biochemistry

A partial cDNA clone for human apoliprotein B

(DNA sequence/amino acid sequence/size and tissue distribution of mRNA)

SAMIR S. DEEB*†, ARNO G. MOTULSKY*†‡, AND JOHN J. ALBERS*

Departments of *Medicine and ‡Genetics and †Center for Inherited Diseases, University of Washington, Seattle, WA 98195

Contributed by Arno G. Motulsky, April 10, 1985

ABSTRACT **A human liver cDNA library was screened for sequences coding for apolipoprotein B (apo B), the major protein of human low density lipoproteins. A mixture of synthetic oligonucleotides (26 bases long) coding for an amino acid sequence known to exist in apo B was used as a hybridization probe. A clone was identified that had a cDNA insert of 593 base pairs and that contained sequences coding for a peptide of 24 residues that had earlier been isolated from apo B by limited proteolysis. The entire nucleotide sequence of the cDNA insert consists of one open reading frame coding for 197 amino acids. Apo B-related RNAs were found in human liver, baboon liver, and the human hepatoma cell line HepG2. None were detected in placenta, simian virus 40 (SV40)-transformed fibroblasts, and a lymphoblastoid cell line. The length of the mature apo B mRNA was estimated to be 18 kb, enough to code for a protein with a molecular weight in the neighborhood of 500,000.**

Apolipoprotein B exists primarily in two forms: apo B-48 and apo B-100. Apo B-48 is synthesized by the intestine and is a component of chylomicrons and chylomicron remnants. Apo B-100 is synthesized by the liver and is the primary lipoprotein of very low density lipoproteins (VLDL), VLDL remnants, and low density lipoproteins (LDL) (for reviews, see refs. 1 and 2). Apo B-100 is the protein ligand on LDL that binds to the LDL (apo B, E) receptor, which results in uptake and catabolism of LDL by the liver (3, 4). It is believed that apo B-100 is involved in atherosclerosis, and elevated plasma levels of B-100 have been found in individuals with premature coronary artery disease (5). Furthermore, individuals with familial combined hyperlipidemia and familial hypercholesterolemia have elevated levels of this protein in plasma (5, 6).

There is considerable variation in estimates of the molecular weight of the monomeric unit of apo B. Results from several laboratories suggest that apo B-100 is a multimeric protein with subunits ranging in apparent M_r between 20,000 and 80,000 (1, 7–10). Others have reported apparent M_rs of 250,000–550,000, based on sedimentation-equilibrium and gel-permeation properties in the presence of 6–7 M guanidine (1, 10, 11), NaDodSO$_4$/PAGE (12), and stoichiometric titration of antigenic determinants on LDL particles (13).

Progress in determining the amino acid sequence of apo B has been slow due to aggregation and insolubility of the delipidated protein in aqueous buffers (14, 15). Recently, *Staphylococcus aureus* protease was used to cleave large peptides of human apo B-100 from LDL. Two of these peptides (24 and 25 amino acid residues long) were purified and sequenced (16). We have used a part of the amino acid sequence of one of these peptides to construct an oligonucleotide which we used to probe a human liver cDNA library for apo B sequences. Here we describe the isolation and characterization of recombinant plasmids containing cDNAs coding for a portion of apo B.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

MATERIALS AND METHODS

Preparation of the Oligonucleotide Probe and Screening of the cDNA Library. The oligonucleotide probe used to screen the cDNA library was a mixture of synthetic DNAs 26 nucleotides long (synthesized in B. Reid's laboratory at the University of Washington by the solid-phase triester method, using an Applied Biosystems 380-A DNA synthesizer, and purified by HPLC). These sequences corresponded to the amino acid sequence -Val-Glu-Phe-Val-Thr-Asn-Met-Gly-Ile- belonging to peptide R3-1 (16). The DNA mixture contained the following 64 sequences: 5′ AT(T/G)CCCA-T(A/G)TT(A/G)GT(A/C)AC(A/G)AACTC(A/C)AC 3′. Selection of this subset of sequences out of a total of 2048 possible coding sequences was based on codon usage frequencies in human apo A-I, A-II, C-I, C-II, C-III, and E. The DNA mixture was radiolabeled at the 5′ ends, by transfer of [^{32}P]P$_i$ from [γ-^{32}P]ATP using bacteriophage T4 polynucleotide kinase (Bethesda Research Laboratories), to a specific activity of $\approx 3 \times 10^8$ cpm/μg (17).

The human liver cDNA library was kindly provided by K. Kurachi and E. Davie (Department of Biochemistry, University of Washington) and contained cDNA inserted into the *Pst* I site of plasmid pUC13. Approximately 10^5 transformants were screened by a modification of the method of Wallace *et al.* (18). Hybridization was performed on colonies transferred to Whatman 541 paper at 37°C. Washing of the hybridized filters was done at 65°C in 6× standard saline citrate (1× is 0.15 M NaCl/15 mM sodium citrate). Strongly hybridizing clones were purified, and plasmid DNA was isolated and purified by centrifugation to equilibrium in cesium chloride/ethidium bromide density gradients (19). DNA inserts from positive clones were isolated by digestion with *Pst* I and characterized by digestion with seven restriction enzymes (Bethesda Research Laboratories), followed by agarose gel electrophoresis.

DNA Sequence Analysis. Restriction fragments from the insert were subcloned in bacteriophage M13 mp11 for sequencing by the dideoxy method (20) using deoxyadenosine 5′-[α-[^{35}S]thio]triphosphate (New England Nuclear) and electrophoresed in gradient gels containing 0.5–2.5× TBE buffer (1× is 89 mM Tris·borate/89 mM boric acid/8 mM EDTA, pH 8.3). Each fragment was sequenced at least twice. T4 DNA ligase and *Escherichia coli* DNA polymerase I (Klenow fragment) were purchased from Bethesda Research Laboratories. Deoxynucleoside triphosphates, dideoxynucleoside diphosphates, and the oligonucleotide M13 primer were purchased from P-L Biochemicals. Part of the sequence was confirmed by the chemical cleavage method of Maxam and Gilbert (21). DNA sequences were stored and analyzed by use of the computer program Gene Pro (Riverside Scientific Enterprises, Seattle, WA).

RNA Analysis. RNA from human liver and the hepatoma cell line HepG2 was isolated by the guanidinium isothiocya-

Abbreviations: apo, apolipoprotein(s); VLDL, very low density lipoprotein(s); LDL, low density lipoprotein(s); kb, kilobase(s).

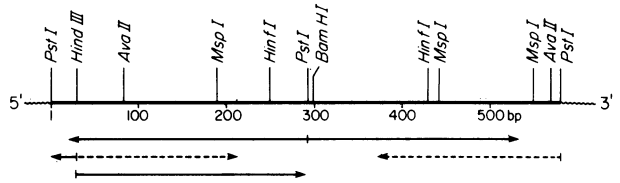

FIG. 1. Partial restriction map and sequencing strategy for the insert of plasmid pHApoB44. The extent and direction of sequencing are indicated by the arrows. Sequence was determined by the dideoxy (solid arrows) or the chemical cleavage (dashed arrows) methods. The 5′ and 3′ termini indicate the orientation of the coding strand. bp, Base pairs.

nate method described in ref. 22. HepG2 cells (23) were grown in minimal essential medium (MEM) supplemented with 5% fetal bovine serum. Preparation of poly(A)$^+$ RNA, its electrophoresis in agarose gels containing formamide, and blot hybridization of RNA were carried out according to published procedures (19, 24). Poly(A)$^+$ RNAs from human placenta and a human lymphoblastoid cell line were kindly provided by B. McConaughy (Genetics Department, University of Washington), baboon liver poly(A)$^+$ RNA was a gift from K. Kurachi, and total RNA from simian virus 40 (SV40)-transformed human fibroblasts was prepared by J. Graves.

RESULTS AND DISCUSSION

Isolation of Clone. A mixture of 26-base-long oligonucleotides, representing 64 of the 2048 possible coding sequences, was used to probe a human cDNA library for apolipoprotein B-specific sequences. Four positive clones were identified

```
                                                    10                                                  20
      Gly Phe Phe Pro Asp Ser Val Asn Lys Ala Leu Tyr Trp Val Asn Gly Gln Val Pro Asp
5' AA GGA TTT TTC CCA GAC AGT GTC AAC AAA GCT TTG TAC TGG GTT AAT GGT CAA GTT CCT GAT
          |           |           |           |           |           |           |
          10          20          30          40          50          60

                                                    30                                                  40
      Gly Val Ser Lys Val Leu Val Asp His Phe Gly Tyr Thr Lys Asp Asp Lys His Glu Gln
      GGT GTC TCT AAG GTC TTA GTG GAC CAC TTT GGC TAT ACC AAA GAT GAT AAA CAT GAG CAG
          |           |           |           |           |           |           |
          70          80          90          100         110         120

                                                    50                                                  60
      Asp MET Val Asn Gly Ile MET Leu Ser Val Glu Lys Leu Ile Lys Asp Leu Lys Ser Lys
      GAT ATG GTA AAT GGA ATA ATG CTT AGT GTT GAG AAG CTG ATT AAA GAT TTG AAA TCC AAA
          |           |           |           |           |           |           |
          130         140         150         160         170         180

                                                    70                                                  80
      Glu Val Pro Glu Ala Arg Ala Tyr Leu Arg Ile Leu Gly Glu Glu Leu Gly Phe Ala Ser
      GAA GTC CCG GAA GCC AGA GCC TAC CTC CGC ATC TTG GGA GAG GAG CTT GGT TTT GCC AGT
          |           |           |           |           |           |           |
          190         200         210         220         230         240

                                                    90                                                  100
      Leu His Asp Ser Ser Ser Trp Lys Ala Ala Ser His Gly Cys Pro His Ser Ala Gly Asp
      CTC CAT GAC TCC AGC TCC TGG AAA GCT GCT TCA CAT GGG TGC CCG CAC TCT GCA GGG GAT
          |           |           |           |           |           |           |
          250         260         270         280         290         300

                                                    110                                                 120
      Pro Gln MET Ile Gly Glu Val Ile Arg Lys Gly Ser Lys Asn Asp Phe Phe Leu His Tyr
      CCC CAG ATG ATT GGA GAG GTC ATC AGG AAG GGC TCA AAG AAT GAC TTT TTT CTT CAC TAC
          |           |           |           |           |           |           |
          310         320         330         340         350         360

                                                    130                                                 140
      Ile Phe MET Glu Asn Ala Phe Glu Leu Pro Thr Gly Ala Gly Leu Gln Leu Gln Ile Ser
      ATC TTC ATG GAG AAT GCC TTT GAA CTC CCC ACT GGA GCT GGA TTA CAG TTG CAA ATA TCT
          |           |           |           |           |           |           |
          370         380         390         400         410         420

                                                    150                                                 160
      Ser Ser Gly Val Ile Ala Pro Gly Ala Lys Ala Gly Val Lys Leu Glu Val Ala Asn MET
      TCA TCT GGA GTC ATT GCT CCC GGA GCC AAG GCT GGA GTA AAA CTG GAA GTA GCC AAC ATG
          |           |           |           |           |           |           |
          430         440         450         460         470         480

                                                    170                                                 180
      Gln Ala Glu Leu Val Ala Lys Pro Ser Val Ser Val Glu Phe Val Thr Asn MET Gly Ile
      CAG GCT GAA CTG GTG GCA AAA CCC TCC GTG TCT GTG GAG TTT GTG ACA AAT ATG GGC ATC
          |           |           |           |           |           |           |
          490         500         510         520         530         540

                                                    190
      Ile Ile Pro Asp Phe Ala Arg Ser Gly Val Gln MET Asn Thr Asn Phe Phe
      ATC ATT CCG GAC TTC GCT AGG AGT GGG GTC CAG ATG AAC ACC AAC TTC TTC  3'
          |           |           |           |           |           |
          550         560         570         580         590
```

FIG. 2. Nucleotide sequence of the cDNA insert in pHApoB44. The predicted amino acid sequence is shown. Sequences corresponding to the oligonucleotide probe and peptide R3-1 are in boldface.

from amongst $\approx 10^5$ recombinant plasmids. The inserts from three of these clones were found to be identical, based on limited restriction mapping and size (≈600 base pairs). One of these clones, designated pHApoB44, was subjected to further restriction mapping and then sequenced. The restriction map of the insert in this clone and the strategy used to determine the sequence are shown in Fig. 1.

The complete nucleotide sequence of pHApoB44 and the predicted amino acid sequence are shown in Fig. 2. This clone represented an uninterrupted reading frame coding for 197 amino acids. The amino acid sequence shown in boldface matches that of apo B peptide R3-1 reported by LeBoeuf *et al.* (16), except for residue 184 (Asp predicted from cDNA; Lys determined by amino acid sequencing). A nucleotide sequence coding for the other apo B peptide (R2-5) sequenced by the same authors was not detected in this cDNA clone. The studied cDNA sequence represented an internal segment of the apo B mRNA as it contained neither untranslated flanking sequences nor a 3′ terminal poly(dA).

Amino Acid Sequence and Comparisons. The amino acid sequence shown in Fig. 2 is particularly rich in glycine, valine, leucine, serine, glutamine plus glutamate, and asparagine plus aspartate. Apo B-100 is known to have a high content of these amino acids except for valine (12). A hydrophilic region (residues 33–41) containing six charged amino acids was another interesting feature of this polypeptide.

A computer search for homology between the amino acid sequence of apo B and those of proteins (including apo A-I, A-II, C-I, C-II, C-III, and E) listed in the protein sequence library provided by the Protein Identification Resource (version of November 1984; National Biomedical Research Foundation, Georgetown University, Washington, DC) revealed no significant similarity to any sequence in the data base. A search at the nucleotide sequence level (GenBank, July 1984, Bolt-Beranek and Newman, Cambridge, MA) showed partial homology to a segment of the human proto-oncogene c-*fos*.

```
apo B (bases 5–29)          ATTTTTCCCAGACAGTGTCAACAAA
                            *** **** ***** ****** ***
c-fos (bases 4091–4115)     ATTGTTCCAAGACATTGTCAATAAA
```

This segment of c-*fos* was in the 3′ untranslated region of the gene.

A search for internal homology in the apo B sequence shown in Fig. 2 revealed some similarity between the following two segments.

```
                    Phe Pro Asp Ser Val Asn Lys Ala Leu
(Bases 9–35)        TTC CCA GAC AGT GTC AAC AAA GCT TTG
                     *  **  **   ** ***     **  *   **
(Bases 54–80)       GTT CCT GAT GGT GTC TCT AAG GTC TTA
                    Val Pro Asp Gly Val Ser Lys Val Leu
```

RNA Analysis. RNA from various tissue types was exam-

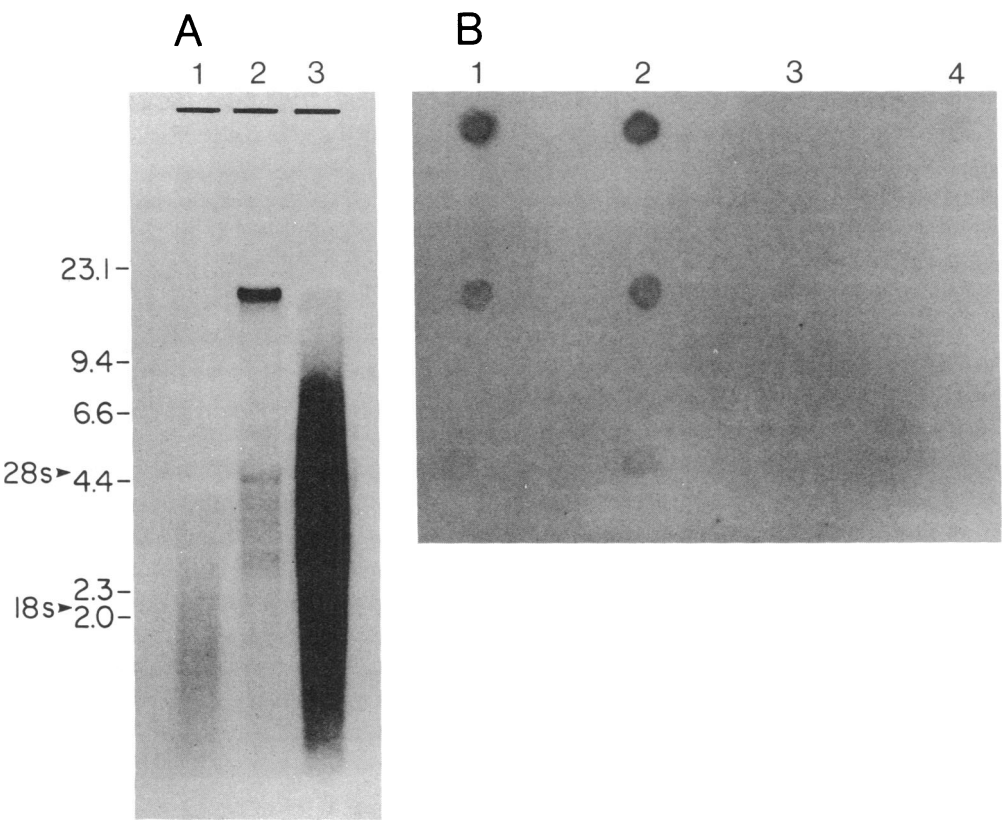

FIG. 3. Analysis of RNA from various tissues for apo B-specific sequences. (*A*) Blot hybridization of electrophoretically fractionated poly(A)$^+$ RNA. ^{32}P-labeled pHApoB44 insert DNA (10^7 cpm; specific activity 1.5×10^8 cpm/μg) was hybridized to RNA blotted onto nitrocellulose after electrophoresis in a 1% agarose gel containing formaldehyde. Unhybridized probe was removed by washing in 0.2× standard saline citrate/0.1% $NaDodSO_4$ at 50°C for 1 hr. Each lane contained 12 μg of poly(A)$^+$ RNA. Lanes: 1, human liver; 2, baboon liver; 3, hepatoma cell line HepG2. The positions to which human 18S and 28S RNA migrated are indicated. The other size markers (given in kb) were *Hin*dIII–digested bacteriophage λ DNA (Bethesda Research Laboratories). The filters were exposed to x-ray film (Kodak XAR-2) with an intensifying screen at −70°C for 13 hr. (*B*) Dot blot hybridization. Poly(A)$^+$ RNA samples from human liver (column 1), baboon liver (column 2), human placenta (column 3), and a human lymphoblastoid cell line (column 4) were spotted (top row, 5 μg; middle row, 1.3 μg; bottom row, 0.32 μg) onto nitrocellulose paper and hybridized to ^{32}P-labeled pHApoB44 insert DNA. The filters were washed as described in *A* and exposed to x-ray film for 16 hr.

ined for the presence of apo B-specific sequences. Radiolabeled insert DNA from clone pHApoB44 was used to probe dot blots of total cellular poly(A)$^+$ RNAs from human placenta and liver, baboon liver, and two cultured cell lines of human origin: HepG2 and a lymphoblastoid cell line. The probe hybridized only to RNA from liver and from the hepatoma cell line HepG2 (Fig. 3*B*). This hepatoma cell line has been shown to synthesize apo B-100 (25). Total RNA from simian virus 40 (SV40)-transformed human fibroblasts was also analyzed and shown not to contain apo B mRNA (data not shown). The labeled pHApoB44 insert was also used to probe blots of electrophoretically fractionated poly(A)$^+$ RNAs (Fig. 3*A*). The size of apo B mRNA in baboon liver was estimated to be 18 kilobases (kb); four minor bands were observed at 2.4–4.4 kb. The size distribution of apo B-related RNAs in HepG2 cells was quite broad: the majority of RNAs hybridizing to the probe were 2.5–4.5 kb long, with minor fractions reaching up to 18 kb in length. For human liver, RNA hybridizing to the probe was even smaller (0.5–3.0 kb). The apo B-specific RNA species smaller than 18 kb may represent fragments generated by nuclease activity during preparation. A mature mRNA of 18 kb could code for a protein of $M_r \approx 500{,}000$.

The isolation of a cDNA probe for apo B has allowed the screening of a human genomic library for the gene, and several clones containing apo B sequences have been identified. This approach will allow determination of the complete amino acid sequence of the protein. The apo B probe would also be useful in assigning the gene to a human chromosome, in identifying DNA polymorphisms for use in linkage analysis, and in studying regulation of transcription.

We thank Drs. E. Davie and K. Kurachi for providing the human liver cDNA library, Betty McConaughy for the poly(A)$^+$ RNAs from human placenta and a lymphoblastoid cell line, J. Brunzell and A. Failor for help in obtaining human liver and preparation of RNA, and Sigrid Hornung for excellent technical assistance. This work was supported by National Institutes of Health Grant HL30086 and Center Grant GM15253.

1. Kane, J. P. (1983) *Annu. Rev. Physiol.* **45,** 637–650.
2. Mahley, R. W., Innerarity, T. L., Rall, S. C., Jr., & Weisgraber, K. H. (1984) *J. Lipid Res.* **25,** 1277–1294.
3. Goldstein, J. L. & Brown, M. S. (1977) *Annu. Rev. Biochem.* **46,** 897–930.
4. Brown, M. S. & Goldstein, J. L. (1983) *J. Clin. Invest.* **72,** 743–747.
5. Brunzell, J. D., Sniderman, A. D., Albers, J. J. & Kwiterovich, P. O., Jr. (1984) *Atherosclerosis* **4,** 79–83.
6. Brunzell, J. D., Albers, J. J., Chait, A., Grundy, S. M., Groszek, E. & McDonald, G. B. (1983) *J. Lipid Res.* **24,** 147–155.
7. Socorro, L. & Camejo, G. (1979) *J. Lipid Res.* **20,** 631–638.
8. Pollard, H., Scanu, A. M. & Taylor, E. W. (1969) *Proc. Natl. Acad. Sci. USA* **64,** 304–310.
9. Bradley, W. A., Rohde, M. F., Gotto, A. M., Jr., & Jackson, R. L. (1978) *Biochem. Biophys. Res. Commun.* **81,** 928–935.
10. Smith, R., Dawson, J. R. & Tanford, C. (1972) *J. Biol. Chem.* **247,** 3376–3381.
11. Steele, J. C. H., Jr., & Reynolds, J. A. (1979) *J. Biol. Chem.* **254,** 1633–1638.
12. Kane, J. P., Hardman, D. A. & Paulus, H. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 2465–2469.
13. Milne, R. W. & Marcel, I. L. (1982) *FEBS Lett.* **146,** 97–100.
14. Cardin, A. D., Witt, K. R., Barnhart, C. L. & Jackson, R. L. (1982) *Biochemistry* **21,** 4503–4511.
15. Lee, D. M., Valente, A. J., Kuo, W. H. & Maeda, H. (1981) *Biochim. Biophys. Acta* **666,** 133–146.
16. LeBoeuf, R. C., Miller, C., Shively, J. E., Schumaker, V. N., Balla, M. A. & Lusis, A. J. (1984) *FEBS Lett.* **170,** 105–108.
17. Wallace, R. B., Shaffer, J., Murphy, R. F., Bonner, J., Hirose, T. & Itakura, K. (1979) *Nucleic Acids Res.* **6,** 3543–3557.
18. Wallace, R. B., Johnson, M. J., Hirose, T., Miyake, T., Kawashima, E. H. & Itakura, K. (1981) *Nucleic Acids Res.* **9,** 879–894.
19. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
20. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
21. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65,** 499–560.
22. Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. & Rutter, W. J. (1979) *Biochemistry* **18,** 5294–5299.
23. Knowles, B. B., Howe, C. C. & Aden, D. P. (1980) *Science* **209,** 497–499.
24. Thomas, P. S. (1983) *Methods Enzymol.* **100,** 255–266.
25. Zannis, V. I., Breslow, J. L., SanGiacoma, T. R., Aden, D. P. & Knowles, B. B. (1981) *Biochemistry* **20,** 7089–7096.