

# Human liver apolipoprotein B-100 cDNA: Complete nucleic acid and derived amino acid sequence

(low density lipoproteins/low density lipoprotein receptor/mRNA/oligonucleotide probe)

S. W. LAW, S. M. GRANT, K. HIGUCHI, A. HOSPATTANKAR, K. LACKNER, N. LEE, AND H. B. BREWER, JR.

Molecular Disease Branch, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD 20892

Communicated by Donald S. Fredrickson, August 1, 1986

**ABSTRACT** Human apolipoprotein B-100 (apoB-100), the ligand on low density lipoproteins that interacts with the low density lipoprotein receptor and initiates receptor-mediated endocytosis and low density lipoprotein catabolism, has been cloned, and the complete nucleic acid and derived amino acid sequences have been determined. apoB-100 cDNAs were isolated from normal human liver cDNA libraries utilizing immunoscreening as well as filter hybridization with radiolabeled apoB-100 oligodeoxynucleotides. The apoB-100 mRNA is 14.1 kilobases long encoding a mature apoB-100 protein of 4536 amino acids with a calculated amino acid molecular weight of 512,723. apoB-100 contains 20 potential glycosylation sites, and 12 of a total of 25 cysteine residues are located in the amino-terminal region of the apolipoprotein providing a potential globular structure of the amino terminus of the protein. apoB-100 contains relatively few regions of amphipathic helices, but compared to other human apolipoproteins it is enriched in  $\beta$ -structure. The delineation of the entire human apoB-100 sequence will now permit a detailed analysis of the conformation of the protein, the low density lipoprotein receptor binding domain(s), and the structural relationship between apoB-100 and apoB-48 and will provide the basis for the study of genetic defects in apoB-100 in patients with dyslipoproteinemias.

Human apolipoprotein B (apoB) is the principal apolipoprotein on chylomicrons, very low density lipoproteins, intermediate density lipoproteins, and low density lipoproteins (LDLs) (1–4). In human plasma, apoB exists in two forms, designated apoB-48 and apoB-100 that are separable by NaDODSO<sub>4</sub>/PAGE (4). apoB-48 and apoB-100 play several important roles in lipoprotein biosynthesis and catabolism. B-48 and B-100 apolipoproteins are required for lipoprotein particle assembly in the liver and intestine. A defect in apoB-48 and apoB-100 secretion results in abetalipoproteinemia, a disease characterized by a deficiency of the B apolipoproteins as well as all apoB-containing lipoproteins including chylomicrons, very low density lipoprotein, and LDL (4). apoB-100 also plays a pivotal role in lipoprotein metabolism as the ligand on LDL that interacts with the LDL receptor and initiates receptor-mediated endocytosis and LDL catabolism (5). Defects in the LDL receptor are associated with severe plasma elevations of apoB-100 and LDL (5).

The structure and physicochemical properties of apoB-100 have been extensively studied for nearly a decade. The analysis of apoB-100 has been difficult because delipidated apoB-100 is insoluble in aqueous solution, and generally aggregates in buffers containing NaDODSO<sub>4</sub>, urea, or guanidine hydrochloride (1–4). The molecular size of apoB-100 has been controversial, and values ranging from 8 to 400 kDa

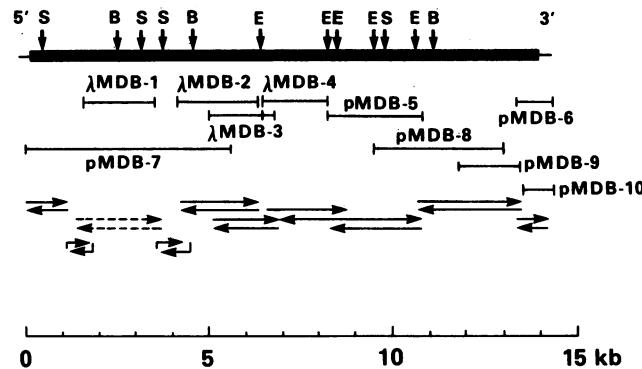


FIG. 1. Restriction endonuclease cleavage map and sequencing strategy for human liver apoB-100 cDNA. The thick solid line represents the coding region of apoB-100 mRNA. The 5'-noncoding and 3'-noncoding sequences are represented by thin lines flanking the coding region. Restriction enzyme sites are represented by a single letter code. (E, EcoRI; B, BamHI; S, Sac I). The regions of the mRNA corresponding to each of the 10 partial cDNA clones are indicated below the map. Arrows below the clones represent direction and length of sequences obtained by the dideoxy procedure on M-13 subclones (solid arrows →) or supercoil plasmid DNAs (bent arrows ↗). Broken arrows (→) represent regions sequenced by the Maxam–Gilbert procedure. The scale indicates the size of the apoB-100 mRNA in kb.

have been reported (1–4, 6). The heterogeneity in apoB-100 molecular size has been attributed to the propensity of delipidated apoB-100 to aggregate and to the reported sensitivity of apoB-100 to protease cleavage (3, 4, 7). We (8, 9) and others (10–16) have reported the partial sequence of human and rat (17) liver apoB-100 cDNA. We now report the complete nucleotide sequence of apoB-100 mRNA and its derived amino acid sequence.

## MATERIALS AND METHODS

**Extraction of Human Liver mRNA.** RNA was isolated from adult human liver obtained from a 34-year-old female automobile accident victim. Frozen liver tissue was pulverized in dry ice by a stainless steel tissue grinder. Liver powder was then homogenized by a polytron in 4 M guanidine thiocyanate (Fluka AG) containing 0.5% sarcosyl, 0.1 M 2-mercaptoethanol, and 0.1% antifoam A followed by centrifugation through a 5.7 M CsCl cushion as described (18, 19). RNA pellets were dissolved in sterile 10 mM Tris-HCl, pH 7.4, containing 0.5% NaDODSO<sub>4</sub> and 1 mM EDTA and then extracted with phenol/chloroform/isoamylalcohol (24:24:1; vol/vol). The RNA in the aqueous phase was precipitated by

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

Abbreviations: LDL, low density lipoprotein; apoB-100, apolipoprotein B-100; kb, kilobase(s).

-27 -25 -20 -15 -10 -5 -1 1 5

ATG GAC CGG CCG ACC GGG CCC GCG CTG CTG SGC CTG GCG CTG CCT GCG CCT CTG CTG CTG CTG GCG GCC AGG GCC GAA GAG GAA ATG CTG  
Met Asp Pro Pro Arg Pro Ala Leu Leu  
Glu Asn Val Ser Leu Val Val Cys Pro Lys Asp Ala Thr Arg Phe His Lys Tyr Asp Ala Arg Lys Tyr Asp Ala Arg Lys Asp Glu Glu Met Leu  
10 15 20 25 30 35 40 45  
GAA AAT GTC AGC CTG TGT TCC CAA AAA GAT GCG ACC CGA TTC AGG AAC CTC CGG AAC TAG  
Gly Ile Asn Val Ser Leu Val Val Cys Pro Lys Asp Ala Thr Arg Phe His Lys Tyr Asp Ala Arg Lys Tyr Asp Ala Arg Lys Asp Glu Glu Met Leu  
50 55 60 65 70 75 80 85  
GCC ACC AGG ATC TCC TAC AAG GTT GAG GCG GAG GTT CCC CAG CGC TCC ACC AGC TTC ATC AGC  
Ala Thr Arg Ile Asn Cys Lys Val Glu Leu Glu Val Pro Glu Leu Cys Ser Phe Ile Leu Lys Tyr Asp Ala Arg Lys Asp Glu Glu Met Leu  
90 95 100 105 110 115 120 125  
CTG AAG AAA ACC AGC AAC TCT GAG GAG TTT GCT GCA GGC ATT TCC AGC TAT GAG CTC AGC  
Leu Lys Ths Lys Asn Ser Glu Glu Phe Ala Ala Met Ser Arg Tyr Glu Leu Lys  
130 135 140 145 150 155 160 165  
ATC CTG AAC ATC TAG AGG GGC ATC ATT TCT GCG CCT CTG GTT CCC CCA GAG ACA GAA GAA  
He Leu Asn Ile Lys Arg Gly Ile Ser Leu Leu Val Pro Pro Glu Thr Glu Glu  
170 175 180 185 190 195 200 205  
AAG ACB AGG AGG AAC ATT GTG GCA ACA GAA ATA TCC ACT GAA AGA GAC CTG GGG GAG TAG  
Lys Thr Lys Tyr Asp Ala Val Aln Thr Asn Ser Thr Glu Arg Asp Leu Lys Cys  
210 215 220 225 230 235 240 245  
CGC CCC TTG TCA ACT CTG ATC AGC AGC AAC TCC TGT TCC TAC TGC AAC TCT GAC AAC TCT  
Arg Pro Leu Ser Thr Leu Ile Ser Ser Ser Glu Cys Gin Tyr Thr Leu Asp Ala Lys  
250 255 260 265 270 275 280 285  
AAG AAT AGG TAT GGG ATT GTG GCA CAA TGA CCA GAG CAC ACT TTT AAA CTT GAA GAC ACA CCA  
Lys Asn Lys Tyr Gly Met Val Aln Val Thr Lys Leu Lys Cys Asp Pro Asp  
290 295 300 305 310 315 320 325  
ACC AAA TCC ACA TCA TCC CCA AAC CAG CGC BCC GAA GCT GTT TTA AAC AGT CTC CAB GAA CTG  
Thr Lys Ser Thr Ser Pro Pys Lys Glu Aln Val Alu Val Thr Leu Glu Leu  
330 335 340 345 350 355 360 365  
ACT GAG CTG CGA CTC AGT GCA GAG TCC AGT GCA TCT GTC TCC CCA CGC CTG ATT AGG  
Thr Glu Leu Arg Gly Leu Ser Asp Glu Val Aln Ser Leu Leu Val Pro Glu Leu  
370 375 380 385 390 395 400 405  
CAC ATC CTC CAG TTG CTG AAA CTT GCA TAT GGC AAC ACC CCC CTT CTG ATA GAT GTC GTC  
His Ile Leu Glu Val Leu Lys Arg Val His Asn Aln Val Pro Leu Ile Asp Val Thr  
410 415 420 425 430 435 440 445  
GCC AGG GAT CAG CGC ACC CGC GGC ACC TCT TTG ATT GTC GCG CTG AAC GAC GTC AAC ATT  
Ala Arg Asp Gly Arg Ser Arg Ala Thr Leu Tyr Asp Leu Ser His Ala Val Asn Asp  
450 455 460 465 470 475 480 485  
CAG ATT CAA GAT GAC TSC ACT GGG GAT GAA GAT TAC ATT TTG ATT CTG CGG GTC ATT  
Gin He Gin Asp Cys Thr Gly Asp Tyr Thr Tyr Leu Ile Leu Arg Val Ile  
490 500 505 510 515 520 525 530  
TBT GTC CAA ACT AAC AGG CCA TCA TCA TGT ATC GCA GAA GAT GTC CCT CGT CGG  
Cys Val Gin Ser Thr Lys Pro Leu Met He Gin Lys Asa Ile Gin Alu Leu  
530 535 540 545 550 555 560 565  
TCT CGG GGA GAT CAG CGA CTG GCT CCT TAT ATT TTG ATT GAG AGT CCT TCA GCA  
Ser Pro Gly Asp Lys Arg Leu Ala Asa Tyr Leu Met Leu Met Arg Ser Pro Ser Glu  
570 575 580 585 590 595 600 605  
GTC GCT TCC CAT ATT GGC ATT ATC TGC AAC TAC TGC ATT GCA GAA TTT GTC ATT  
Val Ala Ser His Ile Asn Ile Leu Asn Ser Glu Glu Leu Asp Ile  
610 615 620 625 630 635 640 645  
TTC TCT CGG AAC TCA TAA CCA CTC TAC AAC TGT GTC TCT CTT CCA ATT CCT GAC CCA GTC  
Pho Ser Arg Val Tyr Glu Leu Tyr Lys Ser Val Ser Leu Pro Asp Pro Asp Asn Thr  
650 655 660 665 670 675 680 685  
CTG AAA ACT ACC CTC ATT GCA TTT GCA TGT GCA GTC GAC AAC TCC ATC GAG ATT GGC TTG  
Leu Lys Thr Leu Thr Ala Phe Gly Ala Ser Ala Asp Ile Glu Ile He Glu  
690 695 700 705 710 715 720 725  
GAC AGT CTC AAC AAA GCT TTT TAC TAC TGC TTT ATT GGT CAA GTC CCT GAT GGT CCT TAC  
Asp Ser Val Asn Lys Ala Leu Tyr Trp Val Asn Glu Val Pro Asp Gly Val Ser Lys  
730 735 740 745 750 755 760 765  
GGA ATA ATG CTC ATT GAT GAG AAC CTG ATT AAA GAT TTG AAA TCC AAA GAA GTC CCG GAA  
Gly Ile Met Ser Val Glu Lys Leu Asp Cys Asp Leu Lys Ser Gly Val Glu Leu  
770 775 780 785 790 795 800 805  
CCA CTC CTG GGA AAC TCC CTG CTG ATG GGG CGC CCC ATC CGT GAG GGG ATT CCG AAC  
Gin Leu Leu Gly Leu Leu Leu Met Gly Ala Arg Thr Glu Gin Ile He Pro Gin Met  
810 815 820 825 830 835 840 845  
GAG ATT GCA CTC TTT GCA TGT AAC TGC ATT GCA AAC TCA TGT GCA AAC TCA TGT GCA  
Glu Asn Ala Phe Glu Leu Pro Thr Gly Ala Gly Leu Glu Gin Ile Ser Ser Gly  
850 855 860 865 870 875 880 885  
CTG GTB GCA AAA CTC TCC GTB TCT GTB GAG ATT GTB ACA ATT ATC GGG ATT ATT CTC  
Leu Val Ala Lys Pro Ser Val Ser Glu Phe Val Met Gly Ile He Ile Pro  
890 895 900 905 910 915 920 925  
GCT CAT GTT GCC CTC AAA GCT GGG AGG CTG AAC ATT TTT ATT CTC TCC CCA AAC AGA CCA  
Ala His Val Ile Leu Lys Ala Gly Leu Lys Phe Leu Ile Pro Ser Pro Lys Pro  
930 935 940 945 950 955 960 965  
ATC CCA CCT CTC ATT GAG AAC ACC CGC TCC TCG TCA GTC AAC GAA GTC ATT CCT GGC  
Ile He Leu Glu Asn Arg Gin Ser Thr Pro Val Ser Val Cys Lys Gin Val Phe  
970 975 980 985 990 995 1000 1005  
TAT CGC CTG ACC GGG GAC ACC AGA TAA TGC ATT TCC ATA CCC CCT TTG CAA GCA GAA  
Tyr Pro Leu Thg Asp Thr Arg Leu Glu Leu Glu Arg Pro Thr Gly Glu He Glu  
1010 1015 1020 1025 1030 1035 1040 1045  
CTB AAG ATT TTT GCA ATT CAA GCA GAA GGC CCG AAC AGC ACT GAG CCT ACC ATC TCC AAA  
Leu Lys Thr Leu Thr Glu Ala Glu Gly Ile Asp Thr Val Thr His Thr Phe  
1050 1055 1060 1065 1070 1075 1080 1085  
CTC GGA ACA ATC CTC AGA ATT TAT GAT GAA TCA TCT ACT GAG GGC AAC ACC TCT TAC AGC  
Ley Glu Ile Arg Val Asn Asp Glu Ser Thr Glu Lys Tyr Ser Thr Val  
1090 1095 1100 1105 1110 1115 1120 1125  
GAC AAC AGA GAA GAA AGA ATT AAC AGT GGT ATT ATT TCC ATA CCC CCT TTG CAA GCA GAA  
Asp Thr Lys Glu Glu Arg Lys Ile Lys Gly Val He Ile Ser Pro Arg Leu Glu  
1130 1135 1140 1145 1150 1155 1160 1165  
TCT GCT GCA ATT TAT GGC TCC TCA AAC ATT GGG ATT TCC ATT GTC CAT ATT GAT GTC ATT  
Ser Ala Tha Alt Tyr Gly Ser Thr Val Ser Val Val Val Val Val Val Val  
1170 1175 1180 1185 1190 1195 1200 1205  
TTC CCT GTB GAT CTC TCC GAT ATT CCT AAC AGC ATT CAT ATG TGT CCT ATT AGA AAC CTC CTG  
Phe Pro Val Asp Leu Ser Asp Pro Tyr Pro Leu His Met Tyr Asp Ala Arg Lys  
1210 1215 1220 1225 1230 1235 1240 1245  
GCA ATC TGC TCG CTC CAG AAC TCA CCT GGG AGT CCT CCT ATT AGC AAC CCT ATT GCA  
Ala Met Ser Val Trp Leu Gin Lys Ala Ser Gly Ser Leu Pro Tyr Thr Glu  
1250 1255 1260 1265 1270 1275 1280 1285  
ATC CCA AAC AAC CTC TCA AAC ATT GGC GCG GTC AAC AAC AAC AAC ATT TGC ATT AAC AGC  
He Pro Glu Asn Leu Phe Leu Lys Ser Asp Gly Arg Val Lys Thr Leu Asn Lys  
1290 1295 1300 1305 1310 1315 1320 1325  
TTA GAG ACT TTG AGG ACA CCA CGC CCT CTC AAC TAC ATT GTC GAA TTC CCT CAT CGG CCT  
Ley Glu Ile Arg Thr Pro Leu His Phe Pro Ser Val Ser Val Cys Lys  
1330 1335 1340 1345 1350 1355 1360 1365  
CTG GGT ATT GTC AAC CCT AAC TAC AGC AAC AAC TAC TGC ATT GTC GCG CCT  
Leu Gly Val Leu Ser Ser Thr Val Val Val Val Val Val Val Val Val  
1370 1375 1380 1385 1390 1395 1400 1405  
GCT GAC CCT GTG GTT GAC CCT CTC TCC TAC ATT GTC CAA GCA TCT GCA ATT GGC ATT  
Ala Asp Ser Val Val Val Ser  
1410 1415 1420 1425 1430 1435 1440 1445  
TCG AAT ATC AAA TCT AGT CAT GTC AAA CCT GGA AAC CCA GTC TCA AAA GGT TTA  
Ser Asn He Lys Phe Ser His Val Gly Lys Leu Asp Pro Val Ser Lys  
1450 1455 1460 1465 1470 1475 1480 1485  
TCC AAT AAA AAC CAC TAT CCT GTC AAC ATT GTC GAG AAC GTC ATT GTC GCG CCT  
Ser Lys Lys Val Lys Val  
1490 1500 1505 1510 1515 1520 1525 1530  
CGC CTC AAT GCA TCC AAC CCT AGG ATT AAC TCC CCT CTC AAC GGC ACC AAC CAG  
Arg Leu Asn Glu Ser Asn Leu Ser Ser Tyr Leu Val Glu Thr Asn Val  
1530 1535 1540 1545 1550 1555 1560 1565  
GCC ATC ATT AAA ACT GTC CCT CTA AAC ATT GAT GAG AAC CTC ACT TTA AAA TCT  
Gly Ile He Lys Lys Asn Thr Ala Ser Lys Tyr Glu Leu Thr Leu Lys  
1570 1575 1580 1585 1590 1595 1600 1605  
AAG CAA ATT GCA CTC CCT GTC ATT GAT GTC ATT GAG TCA AAC TGT GAG ATT GTC CCT  
Lys Gly Asn Asa Leu Leu Arg Ser Gly Tyr Glu Asn Ser Leu Arg Phe  
1610 1615 1620 1625 1630 1635 1640 1645  
ACT GAC AAA ATT ATT AGT GGT GCT CAC AAC GGC CGC ATT GGC ATT GGC AAC GAA  
Th Asp Lys Ile Asn Ser Gly Ile Asp Leu Arg Ile Glu Gly Asp Ile  
1650 1655 1660 1665 1670 1675 1680 1685  
GCA GAG CCT TCC TCT GGG GCA TCT ATT GAA ATT AAC TCA CCT AAC GTC CCT  
Ala Glu Leu Gly Leu Ser Ser Gly Asp Met Lys Leu Thr Asn Arg Phe  
1690 1695 1700 1705 1710 1715 1720 1725  
GCT TAT GAG CCT ATT CTG GGT GTC GAC ACC AAA AAC ATT TTC AAC TTC AGC GTC ATT  
Ala Tyr Glu Ala Met Ile Leu Val Asp Ser Lys Ile Phe Glu Val Ser  
1730 1735 1740 1745 1750 1755 1760 1765  
CAC ACA AAC ACT CTG AAC ATT GCA GGC TTA TCA CTC GAC CCT TCT AAC CTC GAC AAC  
His Itr Asn Ser Leu Asn Ile Ala Gly Leu Ser Leu Asp Ser Ser Lys  
1770 1775 1780 1785 1790 1795 1800 1805  
CTG GTA ACT ATT GTC AAC AGT GAC CCT GTC CCT ACC AAC ATT GGG AAA CCT GCA  
Ley Val Thr Thr Leu Asn Ser Asp Lys Tyr Asp Leu Thr Asn Asp  
1810 1815 1820 1825 1830 1835 1840 1845  
AAT ATT GAA ATA AAA AAC CAC ATC TAT GCC ATC TCT TGT GCT GCC TTA TCA GCA  
Asn Asn Glu Ile Lys His Ile Tyr Ala Ile Ser Ser Ala Leu Ser Ala Ser  
1850 1855 1860 1865 1870 1875 1880 1885

FIG. 2. (Figure continues on the following page.)

FIG. 2. (*Figure continues on the following page.*)

3736           3735           3740           3745           3750           3755           3760           3765  
**T**TT ACA GAT CTT **G**AG GTT CCA TCG TCC AAA CTT GAC TTC AGA GAA ATA CAA ATC TAT AMG AMG CTG AGA ACT TCA TCA TTT GCC CTC ACC CTA CCA ACA CTC CCC GAG STA AAA TTC CCT  
Phe Thr Asp Leu Ile Val Pro Ser Cys 3776 Leu Asp Phe Arg Glu Ile Glu Ile Tyr Ile Lys Leu Arg Thr Val Ser Phe Ala Leu Ile Val Pro Ser Cys 3795 Leu Val Lys Phe Pro  
3770           3775           3780           3785           3790           3795           3800           3805  
**G**AA GTT GAT GTC **T**TA **A**ACA AAA TAT TCT CAA CCA GAA GAC TCC TGA ATT CTC CTT TTT GAG ATA ACC GTG CCT GAA TCT CAG TAA ACT GTG CCT CAG AAC TCC CCA AAA AGT GTT TCA  
Glu Val Asp Val Leu Thr Lys Tyr Ser Gin Pro Glu Asp Ser Leu He Pro Phe Phe Glu Ile Ile Thr Val Pro Glu Ser Gin Leu Thr Val Ser Gin Phe Thr Leu Pro Lys Ser Val Ser  
3810           3815           3820           3825           3830           3835           3840           3845  
**G**AT GGC ATT GCT TTG GAT CTA ATT **G**CA GTC **G**AA AAC ATC GCA GAC TTT GAG TTG CCC ACC ATC ATC GTG CCT CAG AAC ATT TAG ATT CCC TCC ATT AAAG TTC TCT GTA CCT  
Asp Glu Ile Ala Ala Leu Asp Leu Asn Ala Val Ala Asp Lys He Asp Phe Glu Leu Pro Thr Ile Ile Val Pro Glu Gin Thr Ile Glu He Pro Ser Ile Lys Phe Ser Val Pro  
3850           3855           3860           3865           3870           3875           3880           3885  
**G**CT GGA ATT GTC ATT CCT CTC TTT CAA **G**CA CGT ACT GCA CGC CCC TTT GAG GTC GAC TCT CCC GTG TAT ATT GCC ACT TGG AGT GCG ACT TTG AAA AAC GCA GAT TAT GTT GAA ACA GTC  
Ala Glu Ile Val He Pro Ser Phe Gin Ile Leu Thr Ala Arg Phe Glu Val Asp Ser Pro Val Tyr Asn Ala Ile Trp Asn Leu Ser Lys Lys Ala Asp Tyr Val Glu Thr Val  
3890           3895           3900           3905           3910           3915           3920           3925  
**C**TG GAT TCC ACA TGC AGC TCA ACC GTC CAA CTC CTA GAA TAT GAA CTC ATT GAT ATT TTT GGA ACA CAC AAA ATC GAA BGT GTC TAA GTC GCA TGT ACT AAC AAA GCA CTC CCT GAC GAT  
Leu Asp Ser Thr Cys Ser Ser Thr Val Gin Phe Leu Glu Tyr Glu Leu Asn Val Leu Gly 3930           3935           3940           3945           3950           3955           3960           3965  
**G**AC TTC AGT GCA GAA ATT GAA GAA GAT **G**CC AAA ATT GAA GCA CTT GAG GAA TGG GAA GGA AAA GCA CTC ATT ATC AAA ACC GCA GTC CCT ACC GAT CTC CAT CTG CGC TAC CAG AAC  
Asp Phe Ser Ala Glu Tyr Glu Glu Lys Tyr Glu Glu Leu Glu Gin Glu Trp Glu Cys 3970           3975           3980           3985           3990           3995           4000           4005  
**G**AC AAG AAA GGC ATC TCC ACC TCA GCA GGC TCC CCA GGC GTC GGC ACC GTC GGC ATG GAT ATB GAT GAA GAT GAC GAC TTT TCT AAC TGG AGC TTC TAC TAC AAC GTC CCT GAC TCC TCT CCA  
Asp Lys Ile Ser Ser Thr Ser Ala Ile Asp Pro Val Val Ile Arg Phe Glu Val Asp Ser Pro Val Tyr Asn Ala Ile Trp Asn Leu Ser Lys Lys Ala Asp Tyr Val Glu Thr Val  
4010           4015           4020           4025           4030           4035           4040           4045  
**G**AT AAA AAA CCT ACC ATA TTC AAA ACT GAG TTG AGG GTC CGB GAA TCT GAT GAG GAA GAT CAG ATT GAT TGG GAA GAG GCA GCT TCT GGC TTG CTA ACC TCT CTG AAA GAC  
Asp Lys Ile Leu Ile Phe Lys Thr Cys Leu Arg Phe Asp Ser Asp Glu Ile Asp Val Ile Lys 4050           4055           4060           4065           4070           4075           4080           4085  
**A**AC GTG CCC AAC GGC ACA GGG GTC CTT TAT GAT GTC AAC GAG TAC AAC TGG TGG GAA GAC 4090           4095           4100           4105           4110           4115           4120           4125  
Asn Val Asp Ala Thr Gly Val Tyr Asp Tyr Val Asn Lys Tyr His Trp Glu Cys 4130           4135           4140           4145           4150           4155           4160           4165  
**G**AG TGG BTT ATT CAA GGG GGC ATT AGG CAA ATT GAT GTC AAC GTC GAG AGG ATT CAA GTC 4170           4175           4180           4185           4190           4195           4200           4205  
**G**lu Trp Val Tyr Glu Gly Ala Ile Arg Ile Asp Asp Ile Asp Val Arg Phe Glu Ile 4190           4195           4200           4205           4210           4215           4220           4225  
**C**AA GAA CTG TTG ACT GAG GAA GGC CAA GGC ATT GAG GTC CAA ATT GAT GTC GAG 4230           4235           4240           4245           4250           4255           4260           4265  
**G**lu Glu Leu Leu Thr Glu Glu Gly Ala Ile Ser Phe Gin Phe Leu Asp Val Asp Val Ile 4240           4245           4250           4255           4260           4265           4270           4275  
**C**TC ATT GAT TTG CTC AAC TTC CCC AGA ATT CAG TTT CCG GGG AAA CCT GGG ATA TAC ACT 4275           4280           4285           4290           4295           4300           4305           4310  
**G**lu Ile Asp Pro Asp Phe Pro Arg Pro Gin Phe Pro Lys Tyr Pro Ile Tyr Ile Asp 4280           4285           4290           4295           4300           4305           4310           4315  
**A**AC ATT AAA CAG CTC AAA GAG ATG AAA ATT TCT ATT ATT ATT ATT ATC CAA GAT GAB 4320           4325           4330           4335           4340           4345           4350           4355  
**G**lu Ile Lys Leu Lys Glu Met Lys Asp 4325           4330           4335           4340           4345           4350           4355           4360  
**T**CG AAA GTC CAT ATT GGT TCA GAA ATA CTC ATT TCT TAT CAA GCA CTC ATT ATT 4365           4370           4375           4380           4385           4390           4395           4400  
**G**lu Ile Val His Asn Gly Ser Glu Ile Leu Phe Ser Ile Phe Glu Asp Leu Val Ile 4365           4370           4375           4380           4385           4390           4395           4400  
**S**er Lys Val His Asn Gly Ser Glu Ile Leu Phe Ser Ile Phe Glu Asp Leu Val Ile 4370           4375           4380           4385           4390           4395           4400           4405  
**T**TC AAA GAT TTA TCA AAA GAA GGC CAA GAG GTC ATT ATT GAT CTC ATT CTC AAC ACC 4410           4415           4420           4425           4430           4435           4440           4445  
**G**lu Ile Val Ser Lys Glu Ala Glu Glu Ile Ser Phe Gin Ile Ser Phe Gin Ile Ser 4410           4415           4420           4425           4430           4435           4440           4445  
**A**CT TCC CAA CTC TCA AGT CAA GTT GAA CAA ATT CCT CTC CAB AGA ATT GAT GAA ATT 4450           4455           4460           4465           4470           4475           4480           4485  
**G**lu Ile Ser Ser Gin Val Glu Gin Phe Leu His Arg Asn Ile Glu Glu Tyr Ile Asp 4450           4455           4460           4465           4470           4475           4480           4485  
**G**CT CAG GAA ATA ATT AAA AGC CAG GCG ATT GCG AGC AGA AAA ATC ATT ATT TCT GAT CAC 4490           4495           4500           4505           4510           4515           4520           4525  
**G**lu Glu Ile Ile Lys Ser Gin Asp Ile Ile Asn Glu Glu Ile Asn Ser Ile Asn 4490           4495           4500           4505           4510           4515           4520           4525  
**T**TT ATT GCT GAA TCC AAA AGA ATT GTC ATT CCT ATT CAA AAC TAC CAC ACA ATT TCT CTC ATA TAC ATC ACC TCA ATT GAA TCC CAA TCA ACC ACA GTC ATT AGC CCC TAC ATD 4530           4535           4540           4545           4550           4555           4560           4565  
**P**he Ile Ala Glu Asp Ser Lys Arg Leu Asp Leu Ile Glu Asn Tyr His Thr Phe Leu Ile Tyr Ile Thr Glu Leu Lys Lys Leu Glu Ser Thr Thr Val Met Asn Pro Tyr Met 4530           4535           4540           4545           4550           4555           4560           4565  
**A**AG CTT GCT CCA GGA GAA CCT ATT ATC CTC TAA TTTTTTTAAA AGAAATTCTC ATTATTCTT CTTTTCCATTG TGAACCTTCATA CATABCACAGS AAMMAAMATICA AACACTGCTAT ATTBSATAMAA CCATACAGTS 4570           4575           4580           4585           4590           4595           4600           4605  
**A**GGCCAGCCCTT GCGATAGGCCA TGAGACTA AGCAGAGCA CATACTAAC GGCCTCTCACAAAGCTGCC ACCAGGGCTC GGAAGGCTCTC TGAACCTCAGA AGGATGCTCAT TTTTCAAGG TTAAGAAAAA TCAGGATCTC  
AGCTTATTTG CTAATTTG GGGAGGAGGA ACAATAATG GGAGCTCTTAA TTGTGTATCA TAA.....  
**A**as

**FIG. 2.** Nucleotide sequence and derived amino acid sequence of human liver apoB-100 cDNA. Amino acids are numbered below the sequence. Residue 1 is the NH<sub>2</sub> terminus of the mature protein. Negative numbers refer to the prepeptide sequence. Cysteine residues and potential glycosylation sites are printed in bold type. The sequences of apoB peptides that have been published are underlined (22, 23).

ethanol. Poly(A)<sup>+</sup> RNA was isolated by oligo(dT)-cellulose affinity chromatography.

**Preparation of Human Liver cDNA Libraries.** Several human liver cDNA libraries were utilized in this study. A cDNA library established in pBR322 has been described (20). A cDNA library established in λgt11 was kindly provided by G. Ricca. A cDNA library established in pAT153 PvuII/8 was a gift from R. D. Campbell (21).

**Screening the Human Liver cDNA Libraries for ApoB-100 cDNA Clones.** The 5'-end-labeled synthetic oligonucleotides and nick-translated apoB-100 cDNAs were used as hybridization probes to screen the plasmid and the λgt11 cDNA libraries. Monospecific, polyclonal antibodies against apoB-100 as well as monospecific, polyclonal antibodies raised against synthetic apoB-100 peptides (18, 22, 23) were used to screen the λgt11 expression library as described (18).

**DNA Sequence Determination.** The entire apoB-100 mRNA sequence was included in 10 overlapping cDNA clones. Both the Maxam and Gilbert chemical modification procedure (24) and the Sanger's dideoxynucleotides chain-termination procedure (25) were used. In the latter procedure, single-stranded DNA templates were generated by either subcloning into M13mp18 and M13mp19 or by denaturation of supercoil plasmid DNA by NaOH (26). Universal sequencing primers were purchased from Bethesda Research Laboratories. Synthetic oligonucleotide primers were purchased from OCS Labs (Denton, TX).

## RESULTS

**Identification of Human Liver ApoB-100 cDNA Clones.** The λgt11 human liver cDNA library was screened with a mono-

specific antibody to human apoB-100 (GR-22) and with synthetic oligonucleotides SN-R3-1 (22) and SN-MDB-18 (23) that were based on the corresponding peptides present in apoB-100. Four overlapping apoB-100 cDNA clones were isolated. Clone λMDB-1 is 1.9 kilobases (kb) long and located 1.5 kb from the NH<sub>2</sub>-terminal end of apoB-100 (8, 9). The other three overlapping clones (λMDB-2, λMDB-3, λMDB-4) are located 0.7 kb away from the 3' end of clone MDB-1 and have a combined length of 4.2 kb (Fig. 1). To obtain the entire sequence of apoB-100 mRNA, 30-base synthetic oligonucleotides homologous to the 5' and 3' end of our apoB cDNA inserts were employed as hybridization probes to screen several other cDNA libraries. One of the clones, pMDB-7 (Fig. 1) was isolated from a human liver cDNA library, which used the Pvu II site of pAT153/Pvu II/8 (21). This clone hybridized to oligonucleotide probes at the 5' end of λMDB-1 and λMDB-2 and thus contained a minimum of 4 kb of the 5' end of the apoB-100 mRNA. Restriction enzyme analysis revealed an insert size of >6.0 kb indicating this clone included sequences bridging the gap between clones λMDB-1 and λMDB-2. Clones containing the 3' half of the apoB-100 mRNA were obtained by screening the plasmid cDNA library (18, 20) with a nick-translated cDNA probe of λMDB-4 and synthetic oligonucleotides based on the sequence of the apoB peptide MDB-18. Several clones were isolated, clone pMDB-5 (Fig. 1) contained a 3.0-kb insert overlapping the 3'-untranslated region of apoB-100 mRNA and 750 base pairs of the coding region. Several clones containing the rest of the apoB-100 mRNA sequence were obtained from a plasmid cDNA library using the 30-base synthetic oligonucleotides homologous to clones pMDB-5 and pMDB-6.

Each of the putative apoB-100 cDNAs was verified by both RNA gel and Southern hybridization analysis. Oligonucleotides of each clone were also synthesized and used as hybridization probes to human liver poly(A)<sup>+</sup> RNA by RNA gel blot analysis. All cDNA and synthetic probes hybridized to the 14.1-kb apoB-100 mRNA band. In addition, monospecific antibodies prepared against synthetic peptides based on the derived amino acid sequence of the cDNA clones immunoblotted to apoB-100 separated by NaDODSO<sub>4</sub>/PAGE (23).

**Restriction Endonuclease Cleavage Map and Sequence Determination of ApoB-100 mRNA.** The restriction endonuclease cleavage map of apoB-100 mRNA is illustrated in Fig. 1. cDNA inserts were isolated and subcloned into M13mp18 and M13mp19. In some instances, a modification of the dideoxy sequencing protocol using end-labeling was employed (27). Deletion subcloning (28) was used when sequencing large cDNA inserts as in the case of clone pMDB-7. Both the plus and minus strands were sequenced, and sequence data were edited with the aid of a computer.

**Structural Analysis of ApoB-100 mRNA and the ApoB-100 Protein.** The results of our sequence studies established that the apoB-100 mRNA is 14.1 kb long encoding a plasma B-100 apolipoprotein of 4536 amino acid residues with an amino acid molecular weight of 512,723. The completed sequence of apoB-100 mRNA is in good agreement with the published partial sequences of apoB-100 (8–16), except for the 3'-untranslated region that is different from that reported by Knott *et al.* (10) in many places. Computer analysis of the nucleic acid and derived amino acid sequence revealed several interesting features of apoB-100. Twelve out of a total of 25 cysteine residues are located within the first 500 amino acids at the amino terminus of apoB-100 (Fig. 2). There are 20 potential glycosylation sites, the majority of which are located in the middle of the protein.

## DISCUSSION

In the present report we have described the complete nucleic acid and derived amino acid sequence of human liver mature apoB-100. At 4536 amino acids, to our knowledge, apoB-100 is the largest protein cloned and sequenced. The structure and physicochemical properties of apoB-100 have been extensively studied for nearly a decade. However, the analysis of this protein has been difficult because delipidated apoB-100 is insoluble in aqueous solution and forms aggregates even in buffers containing NaDODSO<sub>4</sub>, urea, or guanidine hydrochloride. Our cloning of the entire apoB-100 cDNA permitted the definitive determination of the molecular weight of apoB-100 protein as 512,723. The first 500 residues of apoB-100 contains 12 cysteines that may form intramolecular disulfide bridges resulting in the NH<sub>2</sub>-terminal region of apoB-100 being folded into a globular structure. In this respect, this region of apoB-100 is similar to that of human serum albumin (29, 30) that has 34 cysteine residues forming 17 pairs of intramolecular disulfide bonds. It is known that the introns of the serum albumin gene are located in a characteristic manner at the 5' end and reflect the domain structure of the protein. Based on the similarity in structure of the two proteins, we can anticipate that there are multiple introns in the first 500 amino acid region of apoB-100.

The secondary structure of apoB-100 based on computer analysis revealed 40%  $\alpha$ -helix, 25%  $\beta$ -structure, and 35% random coil but very little amphipathic helical structure. Thus, unlike other apolipoproteins, amphipathic helices are

not a characteristic feature of apoB-100. ApoB-100, however, has a greater percent of  $\beta$ -structure when compared to other apolipoproteins, and segments of the  $\beta$ -structure contain distinct hydrophobic and hydrophilic faces.

The elucidation of the complete amino acid sequence of apoB-100 will permit a detailed analysis of the conformation of the protein, the LDL receptor binding domain(s), the structural relationship between apoB-100 and apoB-48, and the ability to study the genetic defects in apoB structure in patients with dyslipoproteinemias.

- Jackson, R. L., Morissett, J. D. & Gotto, A. M., Jr. (1976) *Physiol. Rev.* **56**, 259–316.
- Scanu, A. M. & Teng, T. (1979) in *The Biochemistry of Atherosclerosis*, ed. Scanu, A. M. (Dekker, New York), pp. 107–122.
- Osborne, J. C., Jr. & Brewer, H. B., Jr. (1977) *Adv. Protein Chem.* **31**, 253–337.
- Kane, J. P. (1983) *Annu. Rev. Physiol.* **45**, 637–650.
- Goldstein, J. L. & Brown, M. S. (1977) *Annu. Rev. Biochem.* **46**, 897–930.
- Elovson, J., Jacobs, J. C., Schumaker, V. N. & Puppione, D. L. (1985) *Biochemistry* **24**, 1567–1578.
- Cardin, A. D., Witt, K. R., Chao, J., Margolis, H. S., Donaldson, V. H. & Jackson, R. L. (1984) *J. Biol. Chem.* **259**, 8522–8528.
- Law, S. W., Lee, N., Monge, J. C., Sakaguchi, A. Y., Naylor, S. L. & Brewer, H. B., Jr. (1985) *Biochem. Biophys. Res. Commun.* **131**, 1003–1012.
- Law, S. W., Lackner, K. J., Hospattankar, A. V., Anchors, J. M., Sakaguchi, A. Y., Naylor, S. L. & Brewer, H. B., Jr. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 8340–8344.
- Knott, T. J., Rall, S. C., Innerarity, T. L., Jacobson, S. F., Urdea, M. S., Levy-Wilson, B., Powell, L. M., Pease, R. J., Eddy, R., Nakai, H., Byers, M., Priestley, L. M., Robertson, E., Rall, L. B., Betsholtz, C., Shows, T. B., Mahley, R. W. & Scott, J. (1985) *Science* **230**, 37–43.
- Proter, A. A., Hardman, D. A., Schilling, J. W., Miller, J., Appleby, V., Chen, G. C., Kirsher, S. W., McEnroe, G. & Kane, J. P. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 1467–1471.
- Wei, C.-F., Chen, S.-H., Yang, C.-Y., Marcel, Y. L., Milne, R. W., Li, W.-H., Sparrow, J. T., Gotto, A. M., Jr. & Chan, L. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 7265–7269.
- Mehrabian, M., Schumaker, V. N., Fareed, G. C., West, R., Johnson, D. F., Kirchgessner, T., Lin, H. C., Wang, X., Ma, Y., Mendiaz, E. & Lusis, A. J. (1985) *Nucleic Acids Res.* **13**, 6937–6953.
- Carlsson, P., Olofsson, S. O., Bondjers, G., Darnfors, C., Wiklund, O. & Bjursell, G. (1985) *Nucleic Acids Res.* **13**, 8813–8826.
- Deeb, S. S., Motulsky, A. G. & Alberts, J. J. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 4983–4986.
- Shoulders, C. C., Myant, N. B., Sidoli, A., Rodriguez, J. C., Cortese, C., Baralle, F. E. & Cortese, R. (1985) *Atherosclerosis* **58**, 277–289.
- Lusis, A. J., West, R., Mehrabian, M., Reuben, M. A., LeBoeuf, R. C., Kaptein, J. S., Johnson, D. F., Schumaker, V. N., Yuhasz, M. P., Schotz, M. C. & Elovson, J. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 4597–4601.
- Law, S. W. & Brewer, H. B., Jr. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 66–70.
- Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. & Rutter, W. J. (1979) *Biochemistry* **24**, 5294–5299.
- Law, S. W., Gray, G. & Brewer, H. B., Jr. (1983) *Biochem. Biophys. Res. Commun.* **112**, 257–264.
- Tertin Belt, K., Carroll, M. C. & Porter, R. R. (1984) *Cell* **36**, 907–914.
- LeBoeuf, R. C., Miller, C., Shively, J. E., Schumaker, V. N., Balla, M. A. & Lusis, A. J. (1984) *FEBS Lett.* **170**, 105–108.
- Hospattankar, A. V., Fairwell, T., Meng, M., Ronan, R. & Brewer, H. B., Jr. (1986) *J. Biol. Chem.* **261**, 9102–9104.
- Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 560–564.
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
- Yanish-Perron, C., Vieira, J. & Messing, J. (1985) *Gene* **33**, 103–119.
- Brunner, A. M., Schimenti, J. C. & Duncan, C. H. (1986) *Biochemistry*, in press.
- Dale, R. M. K., McClure, B. A. & Houchins, J. P. (1985) *Plasmid* **13**, 31–40.
- Law, S. W. & Dugaiczyk, A. (1981) *Nature (London)* **291**, 201–205.
- Minghetti, P. P., Ruffner, D. E., Kuang, W. J., Dennison, D. E., Hawkins, J. W., Beattie, W. G. & Dugaiczyk, A. (1986) *J. Biol. Chem.* **261**, 6747–6757.