We have determined the complete sequence of apolipoprotein (apo) B-100 cDNA. It is 14.1 kilobases in length and codes for a 4563-amino acid protein, including a 27-amino acid signal peptide and a 4536-amino acid mature protein. Further, we identified 2366 residues of apoB-100 by direct sequence analysis of apoB-100 tryptic peptides. The mature peptide is characterized by high hydrophobicity (0.916 kcal/residue) and predicted $\beta$-sheet content (21%). Dot matrix analysis revealed the presence of many long internal repeats in apoB-100. The mature peptide contains 25 cysteine residues, 12 of which are in the N-terminal 500 residues. Twenty potential N-linked glycosylation sites were identified, of which 13 were proven to be glycosylated, and 4 were found not to be glycosylated by direct analysis of tryptic peptides. Our findings on apoB structure provide a basis for future experimentation on the role of apoB-100-containing lipoproteins in atherosclerosis.

Apolipoprotein (apo) B-100 is the natural ligand in LDL that binds to the LDL receptor (1). It is the largest of the apoB group of proteins, the other major ones being apoB-48, apoB-74, and apoB-26. In humans, apoB-100 is produced in the liver, and apoB-48 and apoB-26 appear to be constituent fragments of apoB-100 (2,3) and are sometimes detected in LDL.

Structural studies of apoB-100 have been hampered by its enormous size and by its insolubility in aqueous media following delipidation (4). In this study, we have determined the complete sequence of apoB-100 cDNA. Furthermore, we have determined, by direct peptide sequencing, over half of the residues identified matched those deduced from the DNA sequence. The primary sequence of apoB-100 presents some unusual features not found in other apolipoproteins.

**Materials and Methods**

ApoB-100 cDNA Cloning and Sequencing—The complete cDNA sequence of apoB-100 was obtained from sequences determined on 30 overlapped cDNA clones in $\lambda$t11 (8). The clones were identified by antibody screening (5, 6) or by oligonucleotide screening (7). The sequence was determined on both strands in entirety by the method of Sanger et al. (8); both M13 primers and synthetic oligonucleotide primers were used in the sequencing.

Sequence of ApoB-100 Peptides—LDL was purified from the plasma of a patient with familial heterozygous type II hyperlipoproteinemia (9). It contained apoB-100 as its sole protein component. It was reduced and alkylated before tryptic digestion. The peptide mixture was fractionated on a 2.6 × 150-cm Sephadex G-50 column. The void volume fractions were delipidated and redigested with trypsin. The other fractions from the G-50 column, together with the redigested void volume fraction, were purified by high performance liquid chromatography on a Vydac C8 reverse-phase column (4.6 × 250 mm) (10). Some of the peptides were pure at this stage. Those that were still mixtures were refractionated on a Shandon Hypersil ODS 5-μm reverse-phase column (11).

RESULTS AND DISCUSSION

The complete sequence of human apoB-100 cDNA (Fig. 1, Appendix) covers 14,070 bp, plus the poly(A) tail. It includes a 5’ untranslated region of 78 bp, a coding region of 13,689 bp, and a 3’ untranslated region of 303 bp preceding the poly(A). A putative polyadenylation signal, ATTAAA, is located 22 bp upstream from the poly(A). ApoB-100 mRNA is thus one of the largest eukaryotic mRNAs known.

The complete amino acid sequence of apoB-100 is shown in Fig. 1. There are 4563 amino acids in apoB-100, including a 27-amino acid signal peptide and a 4536-amino acid mature peptide. Of the latter, 2366 residues have been confirmed by direct sequencing of apoB-100 peptides (see Fig. 1). ApoB-100 is thus one of the largest monomeric proteins known, the calculated molecular weight being 512,937 daltons for the mature protein. The protein is characterized by high hydrophobicity; the average value of 0.916 kcal/residue is considerably higher than the corresponding values of 0.718, 0.772, 0.806, 0.863, 0.825, 0.838, and 0.752 kcal/residue for apoE, apoA-IV, apoA-I, apoA-II, apoC-I, apoC-II, and apoC-III, respectively, but lower than that of integral membrane proteins like rhodopsin (1.008 kcal/residue) (14).

Fig. 2 is a hydrophobicity plot (15) of apoB-100. The sequence is generally quite hydrophobic. However, long stretches of exclusively hydrophobic residues seen in some membrane spanning peptides have not been found in the sequence (except the signal peptide region, not included in Fig. 2). Nevertheless, subsequences can be identified on the plot that have higher (or lesser) degrees of hydrophobicity. In general, the carboxyl-terminal region is considerably more hydrophobic than the rest of the molecule, residues 4313 → 4536 having an average hydrophobicity of 1.045 kcal/residue. Secondary structure analysis (16) of apoB-100 primary sequence indicates that it contains 43, 21, and 20% $\alpha$-helical, $\beta$-sheet, and random structures, and 16% $\beta$-turns, respectively. The high content of predicted $\beta$-sheets in apoB-100 is consistent with previous experimental observations (17, 18). A notable feature in the sequence is the highly uneven distri-
bution of Cys residues (Fig. 2). Seven of these Cys are in close proximity to each other, being separated from their nearest neighbor by 1 residue or less (Cys-51, -61, -70; -358, -363; and -399, -449 respectively). Furthermore, of the 25 Cys in apoB-100, 12 are located in the N-terminal 500 residues. These and other Cys residues undoubtedly play crucial roles in maintaining the conformation of apoB in apoB-containing lipoproteins. They may also be involved in disulfide linkages with apo(a) in lipoprotein particles highly correlated with cardiovascular diseases (19).

Twenty potential N-linked glycosylation sites are predicted from the sequence (Fig. 2). Of these, 13 were found to be glycosylated and four unglycosylated by direct sequence analysis. Of particular interest is a cluster of six potential sites between residues 3050 and 3450, all of which were found to be glycosylated. These glycosylated sites might be potential antigenic determinants to some anti-apoB-100 sera.

On the basis of a partial cDNA sequence, Knott et al. (20) identified a region of apoB that shows significant homology to apoE (140–150), the putative LDL receptor binding domain. We identified this region of apoB as residues (3345–3381) and have preliminary evidence that it binds to the LDL receptor.2

Since internal repeats were identified in all the other apolipoproteins (21–23), we analyzed the apoB-100 sequence for such repeats by dot matrix analysis (data not shown). There seems to be many long internal repeats. First, the segment of 92 amino acids from residues 1761–1852 has a 25% homology (identity) with the segment from residues 1906–1997. Second, a 30, 24, 24, 27, 24 and 24% similarity with the segments 4207–4851, 745–757, 145, 67–72, 1131–1146, and 205–214 respectively, a 30, 24, 24, 27, 24 and 24% similarity with the apoA-I segment (residues 11–43) to search for homologous segments in apoB-100. We found that segments 581–613, 1318–1350, 1369–1401, 2862–2887, 4043–4075, and 4151–4183 have, respectively, a 30, 24, 24, 27, 24 and 24% similarity with the apoA-I segment. Whether any of these segments are true homologs of the latter can be tested by studying the genomic structure of apoB-100 because the codon encoding the last amino acid of the common block is invariably split by an intron between the second and third positions (23). Of particular interest among the above segments in apoB-100 are the blocks in other apolipoproteins, they can be divided into three repeats of 11 residues, each starting with a hydrophobic residue (23).

Since internal repeats were identified in all the other apolipoproteins, they can be divided into three repeats. For example, the segment 1604–1651 (48 residues) has a 23% homology with the segment 1882–2000 (119 residues) has a 25% homology with the segment 1532–1568 (37 residues) has a 32% homology with the segment 1816–1852 (37 residues) has a 25% homology with the segment 1761–1852 (92 residues) from residues 1761–1852 has a 25% homology with the segment 1906–1997. Second, a 30, 24, 24, 27, 24 and 24% similarity with the segments 4207–4851, 745–757, 145, 67–72, 1131–1146, and 205–214 respectively, a 30, 24, 24, 27, 24 and 24% similarity with the apoA-I segment (residues 11–43) to search for homologous segments in apoB-100. We found that segments 581–613, 1318–1350, 1369–1401, 2862–2887, 4043–4075, and 4151–4183 have, respectively, a 30, 24, 24, 27, 24 and 24% similarity with the apoA-I segment. Whether any of these segments are true homologs of the latter can be tested by studying the genomic structure of apoB-100 because the codon encoding the last amino acid of the common block is invariably split by an intron between the second and third positions (23). Of particular interest among the above segments in apoB-100 are the segments 581–613 and 1318–1350, because, like the common block in other apolipoproteins, they can be divided into three repeats of 11 residues, each starting with a hydrophobic residue.

Acknowledgments—We thank Drs. C. B. Lawrence and D. A. Goldman for assistance with the dot matrix analysis, S. Saltalamacchia for secretarial assistance, and L. Mathis and A. Alkek for their interest and support.

REFERENCES

1. Brown, M. S., and Goldstein, J. L. (1980) Science 202, 34–47
2. Kane, J. P., Hardman, D. A., and Paula, E. (1980) Proc. Natl. Acad. Sci. U. S. A. 77, 2465–2469
3. Kane, J. P. (1983) Annu. Rev. Physiol. 45, 637–650
4. Bradley, W. A., Rhode, M. F., and Gotto, A. M., Jr. (1980) Ann. N. Y. Acad. Sci. 348, 87–103
5. Young, R. A., and Davis, R. W. (1983) Proc. Natl. Acad. Sci. U. S. A. 80, 1194–1198
6. Wei, C. F., Chen, S. W., Yang, C. Y., Marcel, Y. L., Milne, R. W., Li, W. H., Sparrow, J. T., Goto, A. M., Jr., and Chen, L. (1985) Proc. Natl. Acad. Sci. U. S. A. 82, 7265–7269
7. Chen, L., Moore, M. N., and Tsoo, Y. K. (1986) Methods Enzymol. 128, 745–752
8. Sanger, F., Coulson, A., Barrett, B., Smith, A., and Roe, B. (1980) J. Mol. Biol. 149, 161–178
9. Patash, J. R., Sailer, S., Kostner, G., Sandbofer, F., Holasek, A., and Braunstein, H. (1974) J. Lipid Res. 15, 356–366
10. Yang, C. Y., Pownall, H. J., and Goto, A. M. Jr. (1985) Anal. Biochem. 145, 67–72
11. Yang, C. Y., Pauly, E., Kratzein, H., and Hilschmann, N. (1981) Hoppe-Seyler’s Z. Physiol. Chem. 362, 1131–1146
12. Chang, J. Y., Bruder, D., and Wittman-Liebold, B. (1978) FEBS Lett. 93, 209–214
13. Hewick, R. M., Hunskapper, M. W., Hood, L. E., and Dryer, W. J. (1981) J. Biol. Chem. 256, 7990–8005
14. Marowsky, D., Horgan, D. S., and Gotto, A. M. Jr. (1984) Proc. Natl. Acad. Sci. U. S. A. 81, 4851–4855
15. Kyte, J., and Doolittle, R. F. (1982) J. Mol. Biol. 157, 105–132
16. Garnier, J., Osguthorpe, D. J., and Robson, B. (1978) J. Mol. Biol. 120, 97–120
17. Goto, A. M., Jr., Levy, R. L., and Fredricksson, D. S. (1968) Proc. Natl. Acad. Sci. U. S. A. 60, 1438–1441
18. Chen, C. C., and Kane, J. P. (1975) Biochemistry 14, 3367–3372
19. Berg, K. (1963) Acta Pathol. Microbiol. Scand. Suppl. 59, 166–177: 369–382
20. Knott, T. J., Ball, A. M., Jr., Innerarity, T. L., Jacobson, S. F., Urdea, M. S., Levy-Wilson, B., Powell, L. M., Pease, R. J., Eddy, R., Nakai, H., Byers, M., Prestley, L. M., Robertson, E., Ball, L. B., Bello, C., Shaw, T. B., Mahler, S. W., and Scott, J. (1988) Science 238, 37–45
21. Karathanasis, S. K., Zannis, V. I., and Breslow, L. J. (1983) Proc. Natl. Acad. Sci. U. S. A. 80, 6147–6151

2 C.-Y. Yang, et al., manuscript submitted for publication.
FIG. 1. Nucleotide sequence of apoB-100 cDNA and its deduced amino acid sequence. The numbering of the amino acid starts at the first residue of the mature peptide, whereas that of the cDNA starts at the first nucleotide of the 5' most cDNA clone. The sequences identified by direct peptide sequencing are indicated by lines above the sequences. Compared to three partial cDNA sequences published previously, the following different DNA-deduced amino acid residues were noted: residue 71 was T instead of I in Procter et al. (24); residues 590, 618, and 692 were A, L, and A instead of V, I, and F in Law et al. (25); residues 5292, 5405, 5406, 5705, 5849, 5902, 5907, 4083, 4084, 4605, 4609, 4101, 4106, and 4154 were H, K, E, T, V, I, F, N, Q, V, A, and R instead of D, T, Q, I, A, F, Y, D, H, E, G, and E in Knott et al. (20).
Apolipoprotein B-100 cDNA and Amino Acid Sequence