ApoB-100 Has a Pentapartite Structure Composed of Three Amphipathic α-Helical Domains Alternating With Two Amphipathic β-Strand Domains

Detection by the Computer Program LOCATE*

Jere P. Segrest, Martin K. Jones, Vinod K. Mishra, G.M. Anantharamaiah, David W. Garber

Abstract  Due to the great length of apolipoprotein (apo) B-100, the localization of lipid-associating domains in this protein has been difficult. To address this question, we developed a computer program called LOCATE that searches amino acid sequences to identify potential amphipathic α-helixes and β-strands by using sets of rules for helix and strand terminations. A series of model chimeric protein test datasets were created by tandem linking of amino acid sequences of multiple proteins containing four different secondary structural motifs: motif A (exchangeable plasma apolipoproteins); motif G (globular α-helical proteins); motif C (coiled-coil α-helical proteins); and motif B (β pleated-sheet proteins). These four test datasets, as well as randomly scrambled sequences of each dataset, were analyzed by LOCATE using increasingly stringent parameters. Using intermediate stringent parameters under which significant numbers of amphipathic helixes were found in the exchangeable plasma apolipoproteins. Interestingly, apoB-48 terminates at the N-terminal edge of the middle cluster. By using a similar strategy for analysis of amphipathic β-strands, we discovered that two gap regions between the three amphipathic helix clusters are highly enriched in putative lipid-associating motifs. We propose, therefore, that apoB-100 has a pentapartite structure, representing a globular domain.

Key Words  • apoB-48  • amphipathic helix classes A, Y, G*, G, and B  • globular α-helical proteins  • coiled-coil α-helical proteins  • β-pleated-sheet proteins

Received January 4, 1994; revision accepted July 21, 1994.

From the Departments of Medicine and Biochemistry and the Atherosclerosis Research Unit, UAB Medical Center, Birmingham, Ala.

Correspondence to Jere P. Segrest, MD, PhD, Atherosclerosis Research Unit, Department of Medicine, 630 DREB, No. 3, UAB Medical Center, Birmingham, AL 35294-0012

LOCATE is available on request. It presently runs only on VAX computers.

© 1994 American Heart Association, Inc.
called LOCATE that identifies putative amphipathic α-helices through the application of a set of termination rules to any given protein amino acid sequence. The strategy was to use a number of proteins with known differing secondary structural motifs as controls for the sensitivity of the methods used for identification. The results of the application of the program to the sequence of apoB-100 were then tested against experimentally suggested lipid-associated regions of apoB-100 derived through the use of the principle of releasability of tryptic peptides from trypsin-treated intact LDL.

Various investigators have postulated that amphipathic β-strands may contribute to the high affinity of apoB-100 for the lipid surface of very-low-density lipoprotein and LDL. Additionally, spectroscopic studies have suggested that apoB-100 has a high content of β-structure (up to 40%). To address the issue of amphipathic β-strands, we developed a program option, LOCATE_BETA, which identifies this putative structural motif through the application of termination rules to any given protein amino acid sequence. The set of proteins with known differing secondary structural motifs used as controls for testing the sensitivity of the LOCATE program was used to test LOCATE_BETA.

Methods

Computer Implementations of the Helical Wheel and Helical Net Algorithms

Several simple methods are available to identify the existence of the amphipathic α-helix motif. Two straightforward graphical techniques, the Schiffer-Edmundson helical wheel diagram and the helical net or grid representation by Lim, are the methods of choice for initial analysis. The "helical hydrophobic moment" was introduced by Eisenberg et al as a more quantitative method to describe an amphipathic helical sequence, ie, a numerical way of expressing the helical amphipathicity of a protein segment. This method consists of the vector sum of the hydrophobicity values of the amino acids, taking into account their specific periodic orientation in the α-helix, ie, one residue every 100 degrees, or 3.6 residues in a turn.

To study the location and classification of amphipathic helices, we developed computer implementations of the helical wheel and the helical net algorithms. The helical wheel program (WHEEL) creates a Schiffer-Edmundson helical wheel diagram of a given sequence of amino acids arranged as an ideal α-helix (100° rotation residue) seen down the long axis. The algorithm that locates amphipathic helixes uses the vector sum of the hydrophobic moment per residue as a measure of its amphipathicity or hydrophobicity on the x axis.

The algorithm that locates amphipathic helices uses the notion of the hydrophobic moment. A normalized GES hydrophobicity scale is used to assign a vector magnitude to each amino acid residue of an amphipathic helix, the vector sum of which is the net hydrophobic moment. Since the net hydrophobic moment is dependent on the length of the helix analyzed, the mean hydrophobic moment per residue is used. An amphipathic helix oriented by its hydrophobic moment is aligned so as to place the hydrophobic moment perpendicular to the polar-nonpolar interface. The algorithm can also use the notion of the snorkel orientation of an amphipathic helix postulated previously and recently supported by experimental evidence.

The algorithm starts at the N-terminus of the given amino acid sequence and works towards the C-terminus. It finds the next sequence of five residues that have a hydrophobic moment that is greater than or equal to a specific cutoff value, which by default is 0.20. Also, if this sequence of five residues contains a Pro, the Pro must occur at the N-terminal end. Once the program has a candidate sequence of five residues that meet the above conditions, it begins adding residues to the C-terminus of the candidate sequence. Before a residue is added, the program checks that neither of the following two termination conditions occurs: the residue is a Pro, or the residue begins a cassette of four polar or neutral residues (any of Asn, Gin, Lys, His, Ser, or Thr). If a termination condition

Classification of Amphipathic Helices

We grouped amphipathic helices into seven distinct classes (A, H, L, G, K, C, and M) based on a detailed analysis of their physical-chemical and structural properties. In this classification, class A represents the lipid-associating amphipathic helical domains of the exchangeable apolipoproteins.

The most distinctive feature of the class A amphipathic helix is a unique clustering of positively charged residues at the polar-nonpolar interface and negatively charged amino acid residues at the center of the polar face. The bulk of the van der Waals' surface areas of the positively charged residues are hydrophobic and are thus amphipathic. Therefore, it is not unreasonable to suggest that these amphipathic basic residues, when associated with phospholipid, extend ("snorkel") toward the polar face of the helix to insert their charged moieties into the aqueous milieu. The possibility of snorkeling is one of the major reasons for the interfacial location of the positively charged residues in class A amphipathic helices. The WHEEL program option SNORKEL is based on this premise.

The class A motif has been subdivided into two subclasses: A1 (well-defined class A amphipathic helical domains) and A2 (less well-defined class A amphipathic helical domains). An additional class called class Y has also been defined. The basic features of the class Y motif are two negative-residue clusters on the polar face separating the two arms and the base of the Y motif formed by three positive residue clusters. Both class A and class Y amphipathic helices correspond to the lipid-associating domains of the exchangeable apolipoproteins.

An additional class of amphipathic helix is found in the exchangeable apolipoproteins. This class, termed G*, is present as nine amphipathic helices located in 5 of the 7 exchangeable apolipoproteins and is distinguished by a random radial arrangement of positively and negatively charged residues. These amphipathic helices are similar but not identical to the class G amphipathic helices found in α-helical globular proteins, and thus we call them class G*. Class G* amphipathic helices differ from those of class G in having both a greater hydrophobic moment and a greater nonpolar face hydrophobicity. The class G* amphipathic helices are postulated to function predominantly in protein-protein interactions.

Program for Locating Amphipathic Helices in Protein Sequences (LOCATE)

The LOCATE program finds potential amphipathic helices within a given amino acid sequence. It then creates a diagram showing the relative location of each helix selected on the y axis versus a measure of its amphipathicity or hydrophobicity on the x axis.

The algorithm that locates amphipathic helices uses the notion of the hydrophobic moment. A normalized GES hydrophobicity scale is used to assign a vector magnitude to each amino acid residue of an amphipathic helix, the vector sum of which is the net hydrophobic moment. Since the net hydrophobic moment is dependent on the length of the helix analyzed, the mean hydrophobic moment per residue is used. An amphipathic helix oriented by its hydrophobic moment is aligned so as to place the hydrophobic moment perpendicular to the polar-nonpolar interface. The algorithm can also use the notion of the snorkel orientation of an amphipathic helix postulated previously and recently supported by experimental evidence.

The algorithm starts at the N-terminus of the given amino acid sequence and works towards the C-terminus. It finds the next sequence of five residues that have a hydrophobic moment that is greater than or equal to a specific cutoff value, which by default is 0.20. Also, if this sequence of five residues contains a Pro, the Pro must occur at the N-terminal end. Once the program has a candidate sequence of five residues that meet the above conditions, it begins adding residues to the C-terminus of the candidate sequence. Before a residue is added, the program checks that neither of the following two termination conditions occurs: the residue is a Pro, or the residue begins a cassette of four polar or neutral residues (any of Asn, Gin, Lys, His, Ser, or Thr). If a termination condition

Downloaded from http://ahajournals.org by guest on October 15, 2017
occurs the residue is not added, and no further C-terminal residues are added to the candidate sequence.

Next, the program looks at the C-terminal residue of the candidate sequence. The C-terminal residue is removed from the candidate sequence if any of the following three placement violation conditions occurs: any charged residue is placed equal to or within 40° of the center of the hydrophobic face; any hydrophobic residue is placed equal to or within 40° of the center of the hydrophilic face; or any polar residue is placed equal to or within 20° of the center of the hydrophobic face. This process is repeated until a C-terminal residue is found that violates none of the placement conditions.

The program then applies these rules to the N-terminus of the candidate sequence in a slightly different way. It begins by adding residues to the N-terminus of the candidate sequence.

Next the program checks to see if the candidate sequence passes the final three acceptance conditions: the hydrophobic moment of the entire candidate sequence is greater than or equal to a given cutoff value; the average hydrophobicity of the hydrophobic face is greater than or equal to another specified cutoff value; and the number of residues in the candidate sequence is greater than or equal to a specific cutoff value, which by default is 10 (the minimum value is 4). If the candidate sequence passes these conditions, then it is accepted as a tentative β-strand. Otherwise, it is rejected.

The search begins again in the next residue after the accepted or rejected candidate sequence and continues in this manner to the end of the entire amino acid sequence.

**Creation of Model Amino Acid Sequences to Serve as Test Datasets for LOCATE and LOCATE_BETA**

Four test datasets of model chimeric proteins representing four known distinct secondary structural motifs were created by tandem linking (carboxy terminus to amino terminus) of the amino acid sequences of multiple proteins starting at the amino terminus, or in the case of myosin, by the use of a long segment of amino acid sequence. The four test datasets were composed of proteins enriched in the lipid-associating class A (apolipoproteins), the non-lipid-associating class G (globular proteins) and class C (coiled-coil proteins) amphipathic α-helices, and the non-lipid-associating proteins containing predominantly β-strand secondary structure.

For motif A (exchangeable plasma apolipoproteins), seven human apolipoprotein amino acid sequences were combined into one chimeric sequence in the following order: apoA-I, apoA-II, apoA-IV, apoC-I, apoC-II, apoC-III, and apoE. For motif G (globular α-helical proteins), five α-helix-containing globular proteins with a known crystal structure were combined into one chimeric sequence in the following order: hemerythrin, cytochrome b562, myohemerythrin, cytochrome c3, and myoglobin (human). For motif C (coiled-coil proteins), a 701-residue-long segment of the human myosin β-chain was used as a coiled-coil test dataset. Finally, for motif B (β-sheet proteins), three β-sheet-containing globular proteins with a known crystal structure were combined into one chimeric sequence in the following order: carboxypeptidase A (bovine), retinol-binding protein (human), and subtilisin.

**Creation of Random Amino Acid Sequences From the Four Test Datasets and From ApoB-100 to Serve as Additional Test Datasets for LOCATE and LOCATE_BETA**

We developed the program SCRAMBLE to read in a single-letter code sequence of amino acids and write out another single-letter code sequence that is a random permutation of the original one. The random permutation is created by the algorithm presented by Knuth. By using SCRAMBLE, 100 random amino acid sequences were generated from each of the four chimeric dataset sequences (motifs A, G, C, and B) and from the human apoB-100 sequence.

**Results**

LOCATE Analysis of the Four Test Protein Datasets

Table 1 compares the mean hydrophobic moment per amino acid residue and the mean hydrophobicity per amino acid residue of the nonpolar face of lipid-associating classes with non-lipid-associating classes of amphipathic helices. The set of lipid-associating classes of amphipathic helices are from exchangeable human apolipoproteins (classes A, B, and Y) and from insect apolipoprotein-III, the latter made up of five weak class A amphipathic helices as defined by x-ray crystallography data. The set of non-lipid-associating classes of amphipathic helices are from class G (globular α-helix—
TABLE 1. Properties of Amphipathic Helical Domains of Apolipoproteins

<table>
<thead>
<tr>
<th>Class</th>
<th>Apolipoproteins or Proteins</th>
<th>Mean Hydrophobic Moment per AA††</th>
<th>Hydrophobicity per AA of Nonpolar Face§§</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>CON†</td>
</tr>
<tr>
<td>A₂</td>
<td>A-II, C-I, C-II, C-III (n=8)</td>
<td>0.43</td>
<td>0.74</td>
</tr>
<tr>
<td>A₁</td>
<td>A-I, E (n=8)</td>
<td>0.34</td>
<td>0.71</td>
</tr>
<tr>
<td>Y</td>
<td>A-IV, A-I (n=12)</td>
<td>0.37</td>
<td>0.65</td>
</tr>
<tr>
<td>Insect</td>
<td>Lp-III (n=5)</td>
<td>0.38</td>
<td>0.70</td>
</tr>
<tr>
<td>Total</td>
<td>All lipid-associating</td>
<td>0.37</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>amphipathic helixes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>4-Helix bundle (n=16)</td>
<td>0.32</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A-I, A-IV, C-II, C-III (n=4)</td>
<td>0.44</td>
<td>0.71</td>
</tr>
<tr>
<td>Total</td>
<td>All non-lipid-associating</td>
<td>0.37</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>amphipathic helixes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

AA indicates amino acid residue.
††Data derived from CONSENSUS (CON) analyses.
§§Includes only the six residues centered on the nonpolar face.
†Data derived from CONSENSUS/SNORKEL (CON/SNK) analyses.

The set of lipid-associating classes of amphipathic helices have a mean hydrophobic moment per residue that averages 0.37; the set of non-lipid-associating classes of amphipathic helices have a mean hydrophobic moment per residue that also averages 0.37 (Table 1). The mean hydrophobic moment per residue does not distinguish between the lipid-associating and non-lipid-associating classes. The mean hydrophobicity per amino acid residue of the nonpolar face, however, does a better job of discriminating between the lipid-associating and non-lipid-associating classes of amphipathic helices; the former averages 0.70 and the latter, 0.66.

The four test protein datasets were analyzed by LOCATE by using increasingly more stringent selection parameters (minimal mean hydrophobic moment per residue, minimal mean hydrophobicity per residue of the hydrophobic face, and minimal chain length) for acceptance as an amphipathic helix. Table 2 summarizes some of the results of those analyses. The default parameters for LOCATE (minimal hydrophobic moment per residue, 0.2; minimal hydrophobicity of the nonpolar face, 0.2; minimal helix length, 10).

### Table 2. Analysis of Test Proteins by the Program LOCATE

<table>
<thead>
<tr>
<th>Amphipathic Helices Selected per 100 Residues</th>
<th>Motif A</th>
<th>Motif G</th>
<th>Motif C</th>
<th>Motif B</th>
<th>ApoB-100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default stringency level</td>
<td>3.6</td>
<td>4.3</td>
<td>4.1</td>
<td>3.6</td>
<td>3.5</td>
</tr>
<tr>
<td>Level 1 (length ≥14)</td>
<td>1.6</td>
<td>0.83</td>
<td>1.1</td>
<td>0.83</td>
<td>1.0</td>
</tr>
<tr>
<td>Level 2 (length ≥17)</td>
<td>1.6</td>
<td>0.71</td>
<td>1.1</td>
<td>0.70</td>
<td>0.95</td>
</tr>
<tr>
<td>Level 3 (length ≥20)</td>
<td>1.4</td>
<td>0.50</td>
<td>0.86</td>
<td>0.14</td>
<td>0.53</td>
</tr>
<tr>
<td>Level 4 (length ≥14, snorkel rule)</td>
<td>1.3</td>
<td>0.67</td>
<td>0.86</td>
<td>0.56</td>
<td>0.75</td>
</tr>
<tr>
<td>Level 5 (length ≥20, snorkel rule)</td>
<td>1.0</td>
<td>0.17</td>
<td>0.26</td>
<td>0.14</td>
<td>0.37</td>
</tr>
<tr>
<td>Maximal stringency level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native</td>
<td>1.0</td>
<td>0.17</td>
<td>0.14</td>
<td>0.00</td>
<td>0.26</td>
</tr>
<tr>
<td>Random</td>
<td>0.09</td>
<td>0.10</td>
<td>0.11</td>
<td>0.04</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Apol indicates apolipoprotein. Default stringency parameters: Nosnorkel; minimal hydrophobic moment, 0.2; minimal hydrophobicity of the nonpolar face, 0.2; minimal helix length, 10. Level 1 stringency parameters: Nosnorkel; minimal hydrophobic moment, 0.3; minimal hydrophobicity of the nonpolar face, 0.60; minimal helix length, 14. Level 2 stringency parameters: Nosnorkel; minimal hydrophobic moment, 0.3; minimal hydrophobicity of the nonpolar face, 0.60; minimal helix length, 17. Level 3 stringency parameters: Nosnorkel; minimal hydrophobic moment, 0.3; minimal hydrophobicity of the nonpolar face, 0.60; minimal helix length, 20. Level 4 stringency parameters: Snorkel; minimal hydrophobic moment, 0.3; minimal hydrophobicity of the nonpolar face, 0.60; minimal helix length, 14; snorkel rule. Level 5 stringency parameters: Snorkel; minimal hydrophobic moment, 0.3; minimal hydrophobicity of the nonpolar face, 0.60; minimal helix length, 20; snorkel rule. Maximal stringency parameters: Snorkel; minimal hydrophobic moment, 0.3; minimal hydrophobicity of the nonpolar face, 0.65; minimal helix length, 20; snorkel rule.
FIG 1. LOCATE analyses of the model chimeric protein test datasets. x axis, hydrophobic moment; y axis, test protein amino acid sequence number. Default stringency parameters were used in A and B (minimal hydrophobic moment, 0.2; minimal hydrophobicity of the nonpolar face, 0.2; minimal helix length, 10). Level 3 stringency parameters were used in C-F (minimal hydrophobic moment, 0.3; minimal hydrophobicity of the nonpolar face, 0.60; minimal helix length, 20). A and C, motif A (exchangeable plasma apolipoproteins); B and F, motif B (β-pleated-sheet proteins); D, motif G (globular α-helical proteins); and E, motif C (coiled-coil α-helical proteins).

As the stringency of the LOCATE parameters was increased, the difference between the motif A sequence dataset and the other three sequence datasets also increased. The difference was relatively minor for the level 1 stringency parameters (minimal hydrophobic moment, 0.3; minimal hydrophobicity of the nonpolar face, 0.60; minimal helix length, 14; Table 2). A change in the minimal helix length from 14 to 20 residues (level
3 stringency) produced the largest incremental difference between motif A and the G, C, and B motifs; 1.4 amphipathic helixes per 100 residues was selected for motif A compared with 0.50, 0.86, and 0.14 amphipathic helixes per 100 residues for motifs G, C, and B, respectively (Table 2 and Fig 1C through 1F). These results indicate that one difference between motif A and the G, C, and B motifs may be the length of amphipathic helix domains.

**LOCATE Analysis of the Four Test Protein Datasets Using the Snorkel Rule**

The results of Table 1 illustrate the principle that lipid-associating amphipathic helixes, on average, have a higher mean per residue hydrophobicity of the non-polar face when oriented by the snorkel option than when oriented by the hydrophobic moment. In contrast, the non-lipid-associating amphipathic helixes, on average, have a lower mean per residue hydrophobicity of the non-polar face when oriented by the snorkel option than when oriented by the hydrophobic moment. This is true both for the individual subclasses and classes and for the complete datasets of lipid-associating amphipathic helixes and non-lipid-associating amphipathic helixes. For the complete datasets the snorkel versus nonsnorkel hydrophobicities are 0.71 versus 0.70 and 0.57 versus 0.66 for the lipid-associating (n=33) versus non-lipid-associating (n=25) amphipathic helixes, respectively. This principle will be referred to as the "snorkel rule."

Applying the snorkel rule to LOCATE using level 1 stringency parameters to produce level 4 reduced the number of amphipathic helixes per 100 residues selected for all four motifs approximately equally (Table 2). Applying the snorkel rule to LOCATE using level 3 stringency parameters to produce level 5 began for the first time to differentiate clearly between motif A and the other three motifs (Table 2). Increasing the minimal hydrophobicity of the non-polar face to 0.65 (maximal stringency parameters; minimal hydrophobic moment, 0.3; minimal hydrophobicity of the non-polar face, 0.65; minimal helix length, 20; snorkel rule) did not change the helixes selected by LOCATE for motif A but reduced those selected for motif B to zero (Table 2).

**LOCATE Analysis of Randomized Amino Acid Sequences**

One hundred randomized amino acid sequences were generated from each of the four test protein datasets (as well as the apoB-100 sequence) and analyzed by LOCATE. The mean number of amphipathic helixes per 100 residues selected using maximal stringency parameters are shown in Table 2. This number varied from 0.09 to 0.11 amphipathic helixes per 100 residues for the randomized motif A, apoB-100 motif G, and motif C datasets; the randomized motif B dataset was significantly lower at 0.04 amphipathic helixes per 100 residues, suggesting that the amino acid composition of β-strands is in some way fundamentally different from that of the other types of protein datasets.

**LOCATE Analysis of ApoB-100**

As was true of the test protein datasets, the default parameters for LOCATE selected approximately four amphipathic helixes per 100 residues for apoB-100 (Table 2 and Fig 2A). Further, using level 1 stringency parameters, the number of amphipathic helixes per 100 residues for apoB-100 was indistinguishable from the motif G, C, and B datasets and slightly less than the motif A dataset (Table 1). However, three tightly packed clusters of amphipathic helixes between residues 1 through 1000, 2000 through 2600, and 4000 through 4500 appear on the LOCATE graph (data not shown). Using level 3 stringency parameters to analyze apoB-100 (Fig 2B), two distinct clusters of amphipathic helixes with densities of 2.4 and 2.2 helixes per 100 residues remain between residues 2103 through 2560 and 4061 through 4338, respectively, both greater than the average density of 1.4 in the motif A dataset; the complete lack of amphipathic helixes between residues 1000 through 2000, 2600 through 3200, and 3300 through 3900 is as striking as the presence of clusters. These clusters remain with the use of the snorkel rule (eg, see level 5, Fig 2C) until, using the maximal stringency parameters, the densities of the clusters are reduced to 0.88 and 1.9 amphipathic helixes per 100 residues (data not shown) compared with an average density of 1.0 for the motif A dataset at the same level.

**Chou-Fasman Analysis of ApoB-100**

A Chou-Fasman secondary structural analysis was performed for apoB-100 to locate all α-helical segments predicted to be 15 amino acid residues long. The regions of the three putative amphipathic helical clusters I, II, and III appear to be enriched in predicted α-helical segments compared with the two gap domains between the clusters (data not shown).

**Analysis of Amphipathic Helices Located in the ApoB-100 Clusters**

Fig 3 represents a COMBO analysis of the three amphipathic helix clusters, 58 through 476 (cluster I), 2103 through 2560 (cluster II), and 4061 through 4338 (cluster III), identified by a level 3 LOCATE analysis of human apoB-100. Cluster I appears to be formed predominantly of class G* amphipathic helices (Fig 3A). COMBO analysis of cluster II shows a relatively good class Y pattern, although it differs from the class Y of the exchangeable apolipoproteins in having some reasonable clustering of negative charges at the polar-nonpolar interface (Fig 3B). Finally, COMBO analysis of cluster III shows some clustering of charged residues in an irregular pattern (Fig 3C). Individual WHEEL analyses of the different amphipathic helices in the three clusters shows the presence of four basic amphipathic helical motifs, class G* in cluster I (data not shown) and class Y, class A, and a previously undescribed motif that we term class B (for bipolar) in clusters II and III.

Fig 4A shows one example of a class Y amphipathic helix, 2142 through 2164; Fig 4B, a class A amphipathic helix, 2197 through 2217; and Fig 4C, a class B amphipathic helix, 4262 through 4293. The class B motif has the unusual feature of a severe asymmetry in negative and positive amino acid residue clusters, the negative residues being almost entirely clustered on the left half of the polar face and the positive residues on the right half.

Fig 4D shows a COMBO analysis of the class Y amphipathic helices in clusters II and III. While the Y motif is readily recognizable, it is not as well defined as the class
LOCATE_BETA analysis of the Four Test Protein Datasets and ApoB-100

We noted with interest that a recent Fourier analysis of amphipathic structures in apoB-100 by Atkinson and coworkers (personal communication and Reference 23) indicated that the two gap regions between the three amphipathic helix clusters might be highly enriched in amphipathic \( \beta \)-strands. We decided to test this possibility by designing a version of LOCATE (LOCATE_BETA) to identify amphipathic \( \beta \)-strand regions.

The four test proteins and apoB-100 were analyzed for amphipathic \( \beta \)-strand structure by using various stringency parameters (Table 3). At level 1 stringency (minimal hydrophobic moment, 0.4; minimal hydrophobicity of the nonpolar face, 0.6) and a minimal strand length of 6, there was little difference between the five sequences, other than the fact that motif G had a lower number of amphipathic \( \beta \)-strands than the other four sequences. At level 2 stringency (minimal hydrophobic moment, 0.5; minimal hydrophobicity of the nonpolar face, 0.65) and a minimal strand length of 11, apoB-100 began to be differentiated from the four test proteins, even from motif B. With the addition of only one more residue at level 2 stringency (to a minimal strand length of 12), there was complete discrimination between apoB-100 and the four test proteins: LOCATE_BETA selected 0.55 amphipathic \( \beta \)-strands for apoB-100 and zero for all four test proteins, including the \( \beta \)-sheet motif B. Further, native apoB-100 contained amphipathic \( \beta \)-strands up to 20 residues in length when analyzed at level 2 stringency.

LOCATE_BETA analysis of apoB-100 shows two rather distinct clusters of amphipathic \( \beta \)-strands at level 2 stringency and a minimal strand length of 11 (Fig 5A). These two clusters of amphipathic helices are located in the gap regions between the three clusters of amphipathic helices (residues 827 through 1961 and 2611 through 3867) and have densities of 1.4 and 1.0 amphipathic \( \beta \)-strands per 100 residues, respectively. The distinctly separate linear relations between the three clusters of amphipathic helices and the two clusters of amphipathic \( \beta \)-strands are well illustrated by constructing linear plots (LOCATE_LINEAR) of amphipathic helices (stringency level 3) and amphipathic \( \beta \)-strands (stringency level 2; minimal length, 14) (Fig 5B and 5C, respectively).

LOCATE_BETA analysis of Randomized Amino Acid Sequences

LOCATE_BETA analyses (level 2 stringency; minimal lengths, 12 to 20) were performed on 100 randomized amino acid sequences generated from each of the four test protein datasets and the apoB-100 sequence (Table 3). In spite of clustering, at every length native apoB-100 contained a considerably greater number of amphipathic \( \beta \)-strands than the average for the randomized...
sequences. For example, analysis of native apoB-100 resulted in 0.11 amphipathic $\beta$-strands per 100 residues when the minimal length was set at 18 residues, while an average of only 0.01 amphipathic $\beta$-strands per 100 residues were selected under the same stringency conditions for apoB-100 randomized 100 times. Thus, apoB-100 may prove to be unique among known proteins in regard to the length of its individual amphipathic $\beta$-strands.

**Discussion**

Knowledge of the primary sequence of apoB-100 has allowed the location of several functional regions, including the probable location of the LDL receptor-binding region. By using the principle of nondissociability of tryptic peptides from trypsin-treated intact LDL, two major apoB-100 lipid-associating domains have been reported at residues 1701 through 3070 and 4101 through 4536. Analyzing these data in another way (Fig 6), two continuous stretches of residues, 2100 through 2700 and 4100 through 4500, contain fewer than 10 trypsin-releasable residues per 100 amino acid residues. The clusters of putative lipid-associating amphipathic helices identified by LOCATE between residues 2103 through 2560 and 4061 through 4338 correspond closely to the two presumed lipid-binding domains defined by the absence of trypsin-releasable residues.

The hydrophobic moment analysis, usually performed in a windowing fashion, is most often combined with the calculation of average hydrophobicities. A plot of both
values for every \( n \)-residue-long segment in a protein, often referred to as an Eisenberg plot,\(^4\) enables the detection of different kinds of helices as they cluster into specific regions of the plot. Transmembrane helices have a low helical hydrophobic moment and high hydrophobicity, surface-seeking helices have an average hydrophobicity and high helical hydrophobic moment, and the helices of most globular proteins have both average hydrophobicity and helical hydrophobic moment characteristics. We have modified the Eisenberg plot idea of analysis of amphipathic helices by a combination of the hydrophobic moment with a series of computer programs based on the helical wheel; the average hydrophobicity of the nonpolar face and the snorkel rule are of particular importance to our approach, forming the central basis for LOCATE.

The data presented here suggest that proper use of LOCATE can identify potential lipid-associating amphipathic helical domains of apolipoproteins with an acceptable degree of certainty. Three lines of evidence support this conclusion. First, by using high stringency levels along with the snorkel rule, LOCATE clearly distin-
TABLE 3. Analysis of Test Proteins by the Program LOCATE_BETA

<table>
<thead>
<tr>
<th>Amphipathic β Strands Selected per 100 Residues</th>
<th>Motif A</th>
<th>Motif G</th>
<th>Motif C</th>
<th>Motif B</th>
<th>ApoB-100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1 (length ≥6)</td>
<td>3.2</td>
<td>1.8</td>
<td>3.1</td>
<td>3.0</td>
<td>3.5</td>
</tr>
<tr>
<td>Level 2 (length ≥11)</td>
<td>0.42</td>
<td>0.17</td>
<td>0.0</td>
<td>0.28</td>
<td>0.73</td>
</tr>
<tr>
<td>Level 2 (length ≥12)</td>
<td></td>
<td>0.13</td>
<td>0.14</td>
<td>0.10</td>
<td>0.13</td>
</tr>
<tr>
<td>Native</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.55</td>
</tr>
<tr>
<td>Random</td>
<td>0.13</td>
<td>0.14</td>
<td>0.10</td>
<td>0.13</td>
<td>0.17</td>
</tr>
<tr>
<td>Level 2 (length ≥14)</td>
<td></td>
<td>0.07</td>
<td>0.06</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>Native</td>
<td>0.07</td>
<td>0.06</td>
<td>0.06</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>Random</td>
<td></td>
<td>0.07</td>
<td>0.06</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>Level 2 (length ≥16)</td>
<td></td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>Native</td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>Random</td>
<td>0.07</td>
<td>0.06</td>
<td>0.06</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>Level 2 (length ≥18)</td>
<td></td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>Native</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>Random</td>
<td></td>
<td>0.07</td>
<td>0.06</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>Level 2 (length ≥20)</td>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Native</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Random</td>
<td>0.07</td>
<td>0.06</td>
<td>0.06</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>Level 2 (length ≥21), native</td>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

*Apo* indicates apolipoprotein. Level 1 stringency parameters: Minimal hydrophobic moment, 0.4; minimal hydrophobicity of the nonpolar face, 0.6. Level 2 stringency parameters: Minimal hydrophobic moment, 0.5; minimal hydrophobicity of the nonpolar face, 0.65.

guished between the known lipid-associating dataset, motif A, and the three non–lipid-associating datasets, motifs G, C, and B; at the maximal stringency level, motif B, made up of known β-pleated-sheet proteins, fell to 0 helixes per 100 residues (Table 1). Second, LOCATE analyses of scrambled sequences at maximal stringency were used to define the random-noise baseline. Using this criterion, the occurrence of selected amphipathic helixes in motifs G, C, and B is equal or very close to baseline (0.1), while motif A and the two
clusters of amphipathic helices in apoB-100 are 10 times or more the random-noise baseline (1.0 or greater). Third, clusters II and III in apoB-100 fit the experimentally derived lipid-associating domains of this apolipoprotein extremely well (Fig 5).

ApoB-48 represents approximately 48% of the apoB-100 sequence truncated at residue 2152. Thus, it is interesting that of a total of 11 lipid-associating amphipathic helices, the second lipid-associating amphipathic helix of the middle cluster (cluster II) ends at residue 2137 and the third begins at residue 2170. ApoB-48 apparently differs from apoB-100 in that the former is missing the entire C-terminal cluster (cluster III) and most but not all of the middle cluster of putative lipid-associating amphipathic helices.

The amphipathic helices of apoB-100 differ from those found in the exchangeable apolipoproteins A-I, A-II, A-IV, E, C-I, C-II, and C-III in the diversity of classes defined by the distribution of charged clusters on their polar faces. ApoB-100 has class A, class Y, and class G* amphipathic helices but also a variety of ill-defined amphipathic helix classes as well as a previously undescribed class we call class B; we have also observed the amino-terminal amphipathic helix in serum amyloid A to be class B. The biological role of the great diversity of amphipathic helix classes in apoB-100, and particularly the role of class B, is presently unknown.

Amphipathic helices have been noted in apoB-100 before. By using computer analysis of internal repeats, De Loof et al. noted two clusters of potential 22-mer amphipathic repeats between residues 2079 through 2426 and 4150 through 4484 that are similar to the positions of amphipathic helix clusters II and III as identified by LOCATE. Moreover, by using comparison matrix analysis, De Loof et al. were able to show that the regions between residues 2035 through 2506 and 4002 through 4527 contained high sequence homology to the exchangeable apolipoproteins.

Our LOCATE studies provide several pieces of information not provided by De Loof et al. First, use of the SNOKE_RULE option strongly suggests that amphipathic helix clusters II and III represent lipid-associating domains. Second, the SNOKE_RULE option further suggests that cluster I at the N-terminus is involved in protein-protein interactions and thus may represent a folded α-helical globular domain. Electron microscopic studies have suggested the presence of one or more globular domains on the surface of LDL particles. Third, WHEEL and COMBO analyses of the amphipathic helices identified by LOCATE show that, while apoB-100 contains class A, Y, and G* amphipathic helices, overall the amphipathic helices of this apolipoprotein are considerably more diverse than those found in the exchangeable plasma apolipoproteins. Finally, our study identifies a unique class of amphipathic helix, class B, that is present in multiple copies in apoB-100.

Based on the results reported here, we suggest that there is a highly probability that the two dense clusters of amphipathic helices located precisely in the middle and at the carboxy-terminal end of its amino acid sequence represent major lipid-associating domains of apoB-100. However, three lines of evidence suggest that these two amphipathic helix-containing domains are not the only lipid-associating regions of apoB-100. First, LDL-like particles are secreted with N-terminal apoB fragments as small as B-31; this construct, which ends at approximately residue 1406 (indicated by the arrow in Fig 5C), does not contain amphipathic helix clusters II and III but does contain a significant portion of the first amphipathic β-strand cluster. Second, recombinant DNA constructs of apoB-100 progressively truncated from the C-terminal end produce an apparent linear decrease in size of the LDL particles secreted, which suggests that lipid-associating domains are rather uniformly distributed along the B-100 molecule at least as far the C-terminus of B-31. Finally, Herscovitz et al. have shown that a B-17 construct, while largely secreted in a water-soluble form, will form discoidal complexes with phospholipid. Since this construct has its C-terminus at approximately residue 771 (indicated by the arrow in Fig 5), all of the first amphipathic β-strand cluster is deleted but the amphipathic helix cluster I is retained; these latter results are compatible with the G* nature of amphipathic helix cluster I. The results of the LOCATE_BETA analyses of apoB-100 reported here thus strongly support the concept of amphipathic β-strands serving as lipid-associating domains in this apolipoprotein.

What then is the function of the three amphipathic helix and two amphipathic β-strand clusters we have identified and characterized, and what is their implication for the overall structure of the protein? We suggest the following working hypotheses. First, amphipathic helix cluster I, composed of class G* amphipathic helices and containing many disulfide bonds, represents one or more folded α-helical globular domains that can, under certain circumstances, associate with lipid. Second, the two lipid-associating amphipathic helix clusters, II and III, may represent reversible, and thus flexible, lipid-associating domains, perhaps accounting for the presence of several subclasses of LDL of different sizes. Finally, the gap regions that contain extensive amphipathic β-strands represent irreversible lipid-associating domains.

We propose, therefore, that apoB-100 has a pentapartite structure, NH₂-α₁-β₁-α₂-β₁-α₂-α₁-COOH. Such a model for apoB-100 in LDL is schematically illustrated in Fig 7. This model is supported by experimental thermodynamic data from Walsh and Atkinson that indicate a five-domain, folding organization for apoB-100.
Acknowledgments

This work was supported in part by National Institutes of Health grants HL-34343 and AI-28928.

References

apoB-100 has a pentapartite structure composed of three amphipathic alpha-helical domains alternating with two amphipathic beta-strand domains. Detection by the computer program LOCATE.

J P Segrest, M K Jones, V K Mishra, G M Anantharamaiah and D W Garber