Ringe, D., & Petsko, G. A. (1985) *Prog. Biophys. Mol. Biol. 45*, 197.

Robbins, R. J., Fleming, G. R., Beddard, G. S., Robinson, G. W., Thistlewaite, P. J., & Woolfe, G. J. (1980) *J. Am. Chem. Soc. 102*, 6271.

Rosen, P., Segal, M., & Pecht, I. (1981) *Eur. J. Biochem. 120*, 339.

Ryden, L., & Lundgren, J. O. (1976) *Nature (London) 261*, 344.

Schwartz, R. M., & Dayhoff, M. O. (1978) *Science (Washington, D.C.) 199*, 395.

Silvestrini, M. C., Brunori, M., Wilson, M. T., & Darley-Usmar, V. M. (1981) *J. Inorg. Biochem. 14*, 327.

Solomon, E. I., Hare, J. W., & Gray, H. B. (1976) *Proc. Natl. Acad. Sci. U.S.A. 73*, 1389.

Steiner, R. F., & Kirby, E. P. (1969) *J. Phys. Chem. 73*, 4130.

Strickland, E. H., & Billups, C. (1973) *Biopolymers 12*, 1989.

Szabo, A. G., Stepanik, T. M., Wayner, D. M., & Young, N. M. (1983) *Biophys. J. 41*, 233.

Szabo, A. G., Krajcarski, D., Zucker, M., & Alpert, B. (1984) *Chem. Phys. Lett. 108*, 145.

Takano, T., Dickerson, R. E., Schichman, S. A., & Myers, T. E. (1979) *J. Mol. Biol. 133*, 185.

Tennent, D. L., & McMillin, D. R. (1979) *J. Am. Chem. Soc. 101*, 2307.

Ugurbil, K., & Mitra, S. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 2039.

Ugurbil, K., Maki, A. H., & Bersohn, R. (1977) *Biochemistry 16*, 901.

Ulstrup, J., & Jortner, J. (1975) *J. Chem. Phys. 63*, 4358.

Valeur, B., & Weber, G. (1977) *Photochem. Photobiol. 25*, 441.

van Hoek, A., Vervoort, J., & Visser, A. J. W. G. (1983) *J. Biochem. Biophys. Methods 7*, 243.

Waldeck, D., Cross, A. J., McDonald, D. B., & Fleming, G. R. (1981) *J. Chem. Phys. 74*, 3381.

Wiesenfeld, J. M., Ippen, E. P., Corin, A., & Bersohn, R. (1980) *J. Am. Chem. Soc. 102*, 7256.

Winkler, J. R., Nocera, D. G., Yocom, K. M., Bordignon, E., & Gray, H. B. (1982) *J. Am. Chem. Soc. 104*, 5798.

Yamamoto, Y., & Tanaka, J. (1972) *Bull. Chem. Soc. Jpn. 45*, 1362.

Yamanaka, T., Kijimoto, S., & Okunuki, K. (1963) *J. Biochem. (Tokyo) 53*, 256.

Zweig, A., Lanchaster, J. E., Neglia, M. T., & Jura, W. H. (1964) *J. Am. Chem. Soc. 86*, 4130.

Structure of Porcine Heart Cytoplasmic Malate Dehydrogenase: Combining X-ray Diffraction and Chemical Sequence Data in Structural Studies[†]

Jens J. Birktoft,[‡] Ralph A. Bradshaw,[‡,§,∥] and Leonard J. Banaszak*[,‡]

Department of Biological Chemistry, Division of Biology and Biomedical Sciences, Washington University School of Medicine, St. Louis, Missouri 63110, and Department of Biological Chemistry, California College of Medicine, University of California, Irvine, California 92717

Received July 24, 1986; Revised Manuscript Received November 19, 1986

ABSTRACT: The amino acid sequence of cytoplasmic malate dehydrogenase (sMDH) has been determined by a combination of X-ray crystallographic and chemical sequencing methods. The initial molecular model incorporated an "X-ray amino acid sequence" that was derived primarily from an evaluation of a multiple isomorphous replacement phased electron density map calculated at 2.5-Å resolution. Following restrained least-squares crystallographic refinement, difference electron density maps were calculated from model phases, and attempts were made to upgrade the X-ray amino acid sequence. The method used to find the positions of peptides in the X-ray structure was similar to those used for studying protein homology and was shown to be successful for large fragments. For sMDH, X-ray methods by themselves were insufficient to derive a complete amino acid sequence, even with partial chemical sequence data. However, for this relatively large molecule at medium resolution, the electron density maps were of considerable help in determining the linear position of peptide fragments. The N-acetylated polypeptide chain of sMDH has 331 amino acids and has been crystallographically refined to an *R* factor of 19% for 2.5-Å resolution diffraction data.

The extent to which the function of an enzyme, a protein, or in general any biomacromolecule can be understood depends to a large extent on available structural information. Single-crystal X-ray diffraction has been the most productive method for conformational analysis of proteins, but it requires knowledge of the amino acid sequence as additional input. In the absence of this information, it is still possible to interpret the electron density map of a protein but only in terms of an α-carbon model or, perhaps in somewhat more detail, a polyglycine or polyalanine model. Such models suffice to establish the overall folding patterns, as well as the presence of domain structures, subunit–subunit interactions, and structural relationships to other protein structures, but do not define the nature of the amino acids participating in active sites or other important locations.

Several attempts have been made to interpret electron

[†] This work was supported by NSF Grant PCM-7921864 (L.J.B.), USPHS Grant DK 32465 (R.A.B.) (formerly AM 13362), and a National Institutes of Health biomedical research support grant to Washington University School of Medicine (J.J.B.).

* Author to whom correspondence should be addressed.

[‡] Washington University School of Medicine.

[§] University of California.

[∥] Present address: University of California.

density maps in terms of a complete molecular model without an amino acid sequence, producing a so-called "X-ray sequence". Myoglobin (Watson, 1969), carboxypeptidase (Lipscomb et al., 1969), thermolysin (Matthews et al., 1972), rubredoxin (Herriott et al., 1973), and hexokinase (Anderson et al., 1978a,b) are all examples where structural results were described without a sequence. With the exception of hexokinase, the primary structure of these proteins was subsequently determined by chemical analysis, and in all instances the final results showed that the X-ray sequence was only partially correct. In the most successful case, rubredoxin, 40 of 54 amino acids were correctly identified. That study was carried out at 2.0-Å resolution (Herriott et al., 1973), but even after extensive crystallographic refinement at 1.5-Å resolution, only 50 of 54 amino acids were correctly identified (Watenpaugh et al., 1973). In the cases of myoglobin and carboxypeptidase A, the agreement between chemical and X-ray sequence was considerably less, even though these structures were also analyzed at 2.0-Å resolution.

The determination of the primary structure of proteins larger than 100 residues by chemical methods generally proceeds via the sequencing of shorter fragments, information that will be available well before the complete primary structure is known. If these peptides could be positioned in the electron density map, this important structural information could in turn be used to aid in its interpretation. In the best of circumstances, this approach can facilitate the derivation of the complete amino acid sequence of the protein. Examples of such a strategy include structural studies of wheat germ agglutinin (Wright et al., 1984), tomato bushy stunt virus (Hopper et al., 1984), and lactate dehydrogenase (Taylor et al., 1973).

The three-dimensional structure of porcine heart cytoplasmic or soluble malate dehydrogenase (sMDH; L-malate:NAD$^+$ oxidoreductase, EC 1.1.1.37)[1] was one of the first oligomeric enzymes determined at high resolution. The initial electron density map was obtained by the method of multiple isomorphous replacement at 2.5-Å resolution (Hill et al., 1972; Webb et al., 1973; Tsernoglou et al., 1972). In the absence of knowledge of the primary structure, this map was interpreted in terms of an α-carbon model, which subsequently was expanded into a polyalanine model. Later, information provided by the amino acid sequence of a limited number of peptides was combined with analyses of the electron density resulting in an "X-ray amino acid sequence" or "X-ray sequence", which formed the starting point for restrained least-squares refinement (Birktoft et al., 1982a). At several stages during this process, difference Fourier methods were used to generate electron density maps which were inspected, and the original X-ray sequence assignment was reevaluated. Despite the incomplete nature of the X-ray amino acid sequence, this information was still useful in reaching a number of significant conclusions regarding structure-function properties as well as evolutionary relationships of sMDH (Birktoft et al., 1982a,b; Birktoft & Banaszak, 1983).

With the completion of the amino acid sequence of sMDH by chemical methods,[2] the opportunity to evaluate the success of using electron density maps in the determination of the amino acid sequence for a rather large protein, determined at medium (2.5-Å) resolution, is presented. Such experiences may also have significance in developing methods for partially automating the interpretation of electron density maps.

Materials and Methods

X-ray Diffraction Data. The X-ray diffraction data used in the experiments described here extend to 2.5-Å resolution and were collected by diffractometry as previously described (Tsernoglou et al., 1972). The data set referred to as "native sMDH" data was collected from the so-called "type C" crystals that are grown from 65% saturated ammonium sulfate in the presence of a 10-fold molar excess of NAD at pH 5.1. The crystals belong to space group $P2_12_12$ and have unit cell dimensions of $a = 139.2$ Å, $b = 86.6$ Å, and $c = 58.8$ Å. There is one dimeric molecule in the asymmetric unit. The two chemically identical subunits of a unique dimer will be referred to as subunit 1 and subunit 2.

X-ray Methods. Difference Fourier maps were calculated employing fast Fourier transform methods according to Ten-Eyck (1973) and by using as coefficients $(NF_{\text{obsd}} - MF_{\text{calcd}})\exp(i\alpha_{\text{calcd}})$, where F_{obsd} and F_{calcd} were the observed and calculated structure amplitudes, respectively. F_{calcd} and the phase angles α_{calcd} were derived from atomic coordinates, and N and M were used in the following combinations: $(N, M) = (3, 2), (2, 1)$, and $(1, 1)$. Occasionally, selected parts of the model were omitted in the calculation of difference Fourier maps. Such electron density maps will be referred to as partial F_{calcd} maps.

Restrained crystallographic refinement was performed according to the method of Hendrickson and Konnert (1981). Details of this procedure as applied to sMDH have already been described (Birktoft et al., 1982a; Birktoft & Banaszak, 1983). In a few instances where some uncertainty about the exact amino acid sequence remained even after chemical sequence data had been assigned, the side chains of the residues in question were removed from the coordinate list, and the modified structure was subjected to about five cycles of refinement. The resulting coordinates were then used to generate new electron density difference maps.

Peptides Used in Fitting. The methods employed in the preparation and amino acid sequencing of the peptide fragments used in the peptide fitting discussed below will be described elsewhere.[2] Here it suffices to mention that sMDH after being S-carboxymethylated was digested with trypsin, chymotrypsin, thermolysin, or *Staphylococcus aureus* protease. The resulting digests were fractionated with ion-exchange chromatography and gel filtration; the purified peptides were sequenced by manual Edman degradation. The tryptic digest yielded 33 peptides of which 17 were fully and 10 were partially sequenced whereas only the composition and end-group information was available for 6 peptides. Thirty-eight peptides were obtained from the chymotryptic digest, and of these 26 were fully sequenced, 10 were partially sequenced, and 2 had only their composition and/or end groups determined. Of the 72 thermolytic peptides, 27 were fully and 23 were partially sequenced, with only composition data for the remaining 13 peptides. The *S. aureus* protease digest yielded 15 peptides of which 8 were fully and 6 were partially sequenced; for one peptide, only the composition was provided. Finally, a tryptic digest of sMDH that was S-carboxymethylated and maleyated yielded sequence information for two peptides. Independently, Allen et al. (1971) have reported sequence data obtained from a tryptic digest of sMDH that had been aminoethylated. Thirty-nine peptides were described of which 12 were identical with peptides obtained from the above-mentioned tryptic digest. Of the remaining unique 27 peptides, the number of fully, partially, and unsequenced peptides was 16, 2, and 9, respectively.

[1] Abbreviations: sMDH, cytoplasmic malate dehydrogenase; MIR, multiple isomorphous replacement; rms, root mean square.

[2] J.-K. Fang, M. J. Wade, B. E. Glatthaar, M. R. Sutton, G. R. Barbarash, G. A. Grant, and R. A. Bradshaw, manuscript in preparation; R. A. Bradshaw, J.-K. Fang, R. T. Fernley, J. J. Birktoft, and L. J. Banaszak, manuscript in preparation.

The majority of the fully sequenced peptides, 76 of 106, together with a number of the partially sequenced peptides, 30 of 51, and unsequenced peptides, 16 of 30, could be combined into longer segments due to the presence of suggested overlaps. In the end, a total of 19 unique peptide combinations could be deduced from the chemical data. In some instances peptides could be combined in more than one way, but this did not increase the total number since such alternate combinations eliminated other possibilities.

Model Building and Amino Acid Sequence Assignment. During the early phases of this work, model building and map interpretation were accomplished by fitting Kendrew–Watson models to the electron density maps with an optical comparator generating the models denoted model 1 and model 2. All subsequent model building and map interpretation were done on an MMS-X interactive graphics system (Barry et al., 1974). The resulting molecular models are referred to as model 3, model 4, etc. The work on the graphics system was facilitated by the use of the programs NEWNIP (Lederer et al., 1981) and BUILD (Miller et al., 1981).

The initial interpretation of the MIR-phased electron density map at 2.5-Å resolution was done in terms of a polyalanine model, model 1 (Hill et al., 1972). Subsequently, the same map was analyzed with the purpose of estimating the amino acid sequence of sMDH directly from the electron density map. Two experimental protocols were used. One involved the incorporation of side-chain atoms into the model for use in least-squares refinement. The second was the preparation of a list of probabilities for each residue position in order to attempt to position peptides whose sequences were known from chemical studies. The first protocol, which has been described previously (Birktoft & Banaszak, 1983), resulted in model 2, which was used in the initial refinement cycles (Birktoft et al., 1982a; Birktoft & Banaszak, 1983).

The second approach involved a visual estimate of the probability that any one of the known amino acids could be located at a specific position in the electron density map. The probability scoring was done on the scale of 0 (very unlikely) to 5 (highly probable), and each subunit was evaluated independently. Primary consideration was given to the space occupied by the density assumed to enclose the side chain, as well as the shape of the side-chain density. In many instances more than one side-chain type appeared to fit the density volume equally well, and in such instances the scoring values were modified by taking into consideration such features as the relative location of the side chain as well as its proximity to other parts of the protein or to the solvent. Thus the score for polar residues was decreased by one if the side chain is in an internal position, and similarly, the probability scores for hydrophobic residues would be reduced if the side chain is in an external position. Although it is difficult to assess hydrogen bonding without knowledge of the precise amino acid sequence, possible hydrogen bonding between the polypeptide backbone and side chains was taken into consideration if the side-chain density approached densities associated with peptide bond amino and carbonyl groups. Residues for which hydrogen bonding appeared as a possibility had their probability score increased by one. The environment of a residue, in addition to providing clues as to the hydrophobic/hydrophilic nature of the side chain, also suggests its maximum possible size. This was most obvious for internally located residues where surrounding elements of density suggest volume limitations for the side chain. Surface residues on the other hand do not have such restrictions, and for these, longer side chains were also considered a possibility but were only given a low probability score, usually 1. No distinction was made between glutamate and glutamine nor between aspartate and asparagine.

After a residue by residue correspondence between the two subunits was established by the least-squares procedure of Rao and Rossmann (1973), a joint tabulation for each amino acid position was obtained by summation of the scoring for the two subunits with the maximum value being limited to 9. At those positions where the subunits differed in the number of amino acids, the scoring value for the nonobserved residues was entered as 0. A small segment of the scoring matrix, comprising residues 34–60, is given as an example in Table I.

The most probable placements in the polypeptide chain of the chemically sequenced peptides were estimated with a systematic search procedure, PEPTIDFIT, similar to that used for studying amino acid sequence homologies (McLachlan, 1972). By use of the probability scores shown in part in Table I, each known peptide was placed in all possible positions along the polypeptide chain. The scoring value for that peptide in each electron density map position was then the sum of the probability values for the specific amino acids located in these positions. Applications of this protocol resulted in model 3, which incorporated an amino acid sequence that was a mixture of fitted sequenced peptides and assignments based on electron density map appearance. The fitting of the longer and more completely sequenced peptides that resulted in model 4 and model 5 was also facilitated by the PEPTIDFIT program.

Results

First X-ray Amino Acid Sequence (Model 2). After the initial polyalanine model (model 1) of sMDH was completed (Hill et al., 1972), the MIR map was examined further for the purpose of obtaining the initial X-ray sequence as well as coordinates for all fitted atoms including those located in the electron density believed to be associated with side chains, even where the identity of the amino acid was unknown. A description of the history of the sequence assignments for the same residues described in Table I, 34–60, and used in the refinement is presented in Table II. Representative segments of the electron density maps employed during the various revisions tabulated in Table II are shown in Figure 1. A summary description of all the models, including information about chemical sequences, and their refinement is given in Table III. As can be seen from Table III, model 2 was comprised of 4127 non-hydrogen atoms distributed in 321 and 324 amino acids in subunits 1 and 2, respectively. This represents 81% of the total number, excluding hydrogens, of protein atoms found in the sMDH dimer and compares with the 3195 atoms that were included in model 1. Excluded from these numbers are the 98 atoms contributed by bound NAD and sulfate ions (Webb et al., 1973). Of the total of 645 amino acids in the two subunits, 375 were real amino acids, and 270 were pseudo amino acids.

Initial Crystallographic Refinement. The coordinate set representing model 2 was used as the starting point for the restrained crystallographic refinement procedure, which was undertaken with the premise that interpretation of electron density associated with the amino acid side chains would become less ambiguous, leading in the direction of the correct amino acid sequence. Initially, six cycles of refinement with data in the resolution range 9.0–3.5 Å were carried out with the conventional crystallographic R factor decreasing from 45% to 41%. The resulting coordinates were then used to calculate $F_{\mathrm{obsd}} - F_{\mathrm{calcd}}$ and $2F_{\mathrm{obsd}} - F_{\mathrm{calcd}}$ difference electron density maps.

As a test of the quality of the phases resulting from the incomplete model, the cofactor NAD and the bound sulfate

Table I: Amino Acid Sequence Scoring Matrix Derived from an Electron Density Map[a]

chemical sequence	P	G	A	S	C	T	V	B	I	L	Z	M	K	R	H	F	Y	W
34 Pro	⟨4⟩	0	6	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35 Ile	0	0	0	0	0	8	8	0	⟨3⟩	0	0	0	0	0	0	0	0	0
36 Ile	0	0	0	0	0	5	4	2	⟨2⟩	3	4	0	0	0	2	0	0	0
37 Leu	0	0	0	0	0	0	0	3	5	⟨6⟩	2	0	0	0	0	0	0	0
38 Val	0	1	4	4	1	5	⟨4⟩	0	2	0	0	0	0	0	0	0	0	0
39 Leu	0	0	5	5	0	3	4	0	3	⟨1⟩	0	0	0	0	0	0	0	0
40 Leu	0	0	0	0	0	2	4	0	3	⟨0⟩	4	3	0	3	0	0	0	0
41 Asp	0	0	0	2	2	0	0	⟨8⟩	0	7	3	0	0	0	3	0	0	0
42 Ile	4	0	4	5	0	2	2	0	⟨0⟩	0	0	0	0	0	0	0	0	0
43 Thr	0	0	0	0	0	⟨4⟩	3	5	0	0	2	0	1	1	0	0	0	0
44 Pro	⟨6⟩	0	4	4	0	2	1	2	0	0	1	0	0	0	0	0	0	0
45 Met	0	0	0	2	0	0	0	5	0	4	4	⟨0⟩	3	5	2	0	0	0
46 Met	0	0	1	2	0	2	2	2	0	1	1	⟨1⟩	1	1	0	0	0	0
47 Gly	0	⟨8⟩	5	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
48 Val	4	0	0	2	0	6	⟨6⟩	5	4	0	0	0	0	0	0	0	0	0
49 Leu	0	0	0	6	0	4	4	5	0	⟨3⟩	0	0	0	0	0	0	0	0
50 Asp	0	0	0	0	0	0	0	⟨4⟩	0	4	7	4	2	2	0	0	0	0
51 Gly	0	⟨2⟩	9	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
52 Val	0	0	0	4	4	6	⟨8⟩	2	0	0	0	0	0	0	0	0	0	0
53 Leu	0	0	0	0	0	0	0	4	0	⟨3⟩	7	5	5	4	0	0	0	0
54 Met	0	0	0	0	0	0	0	0	0	0	7	⟨9⟩	2	2	4	3	3	0
55 Glu	0	9	6	3	0	0	0	4	0	4	⟨2⟩	0	0	0	0	0	0	0
56 Leu	0	5	2	0	0	0	0	2	3	⟨3⟩	2	1	0	0	0	0	0	0
57 Gln	0	0	0	0	0	0	0	2	0	2	⟨8⟩	3	6	6	0	0	0	0
58 Asp	0	0	0	0	0	0	0	⟨3⟩	0	4	0	0	0	0	6	5	5	4
59 Cys	0	0	3	3	⟨5⟩	6	6	1	1	1	1	0	0	0	0	0	0	0
60 Ala	0	4	⟨8⟩	6	0	0	0	2	0	0	0	0	0	0	0	0	0	0

[a] The table shows the joint scoring matrix for a segment of the sMDH polypeptide chain that includes residues 34–60. The scoring value for each amino acid, ranging from 0 to 9, is entered under the single-letter amino acid code at the top of the table. The chemically determined sequence is listed in the left column, and the scoring value for the correct amino acid is enclosed in brackets.

Table II: Summary of Sequence Assignments of Residues 34–60 of sMDH[a]

residue no.	model 1	first revision	model 2	second revision	model 3	third revision	model 4	fourth revision	model 5
34	Ala	Ala	Ala	pept fit	Pro	pept fit	Pro	pept fit	Pro
35	Ala	Thr, Val	Thr	pept fit	Ile	pept fit	Ile	pept fit	Ile
36	Ala	Thr	Ava	pept fit	Leu	pept fit	Leu	pept fit	Ile
37	Ala	Leu	Abu	Leu	Leu	good fit	Leu	good fit	Leu
38	Ala	Thr	Ser	good fit	Ser	good fit	Glu	Ser better	Val
39	Ala	Ala, Ser	Ala	Leu	Leu	no side chain, Gly?	Gly	Leu, Glu(?)	Leu
40	Ala	Val, Glx	Val	Met	Met	Leu better	Leu	good fit	Leu
41	Ala	Asx	Asp	good fit	Asp	good fit	Asp	pept fit	Asp
42	Ala	Ser	Ala	Val	Val	Pro or Ile better	Ile	pept fit	Ile
43	Ala	Asx	Ava	Val	Val	Ser or Thr better	Thr	pept fit	Thr
44	Ala	Pro	Ala	Pro	Pro	good fit	Pro	pept fit	Pro
45	Ala	Arg, Asx	Arg	Lys	Lys	not Lys, Met better	Met	pept fit	Met
46	Ala	Ser, Thr, Asx	Ava	Glx	Gln	aromatic, Met	Met	pept fit	Met
47	Ala	Gly	Gly	Glx	Gln	no side chain	Gly	pept fit	Gly
48	Ala	Thr	Abu	Thr, Val	Thr	Val or Ile	Val	pept fit	Val
49	Ala	Ser	Ser	good fit	Ser	Leu better	Leu	pept fit	Leu
50	Ala	Glx	Ala	Glx	Glu	Asx better	Asp	pept fit	Asp
51	Ala	Ala	Ala	good fit	Ala	Gly better	Gly	pept fit	Gly
52	Ala	Val	Abu	Val, Asx	Val	good fit	Val	pept fit	Val
53	Ala	Glx	Ava	Asx	Asn	Leu better	Leu	pept fit	Leu
54	Ala	Met	Met	good fit	Met	good fit	Met	pept fit	Met
55	Ala	Gly	Gly	pept fit	Glu	pept fit	Glu	pept fit	Glu
56	Ala	Gly	Gly	pept fit	Leu	pept fit	Leu	pept fit	Leu
57	Ala	Glx	Aca	pept fit	Gln	pept fit	Gln	pept fit	Gln
58	Ala	His	Phe	pept fit	Asp	pept fit	Asp	pept fit	Asp
59	Ala	Thr	Cys	pept fit	Cys	pept fit	Cys	pept fit	Cys
60	Ala	Ala	Ala	pept fit	Ala	pept fit	Ala	pept fit	Ala

[a] The residue numbers in the left-most column of the table are based on the model 5 structure. The initial sequence assignment, which was based on the assessment of the MIR electron density map, is listed in the "first revision" column and corresponds to the highest probability values as is shown in Table I. The residue names Abu, Ava, and Aca are the names used for the nonstandard amino acids incorporated into model 2 and describe linear aliphatic α-amino acids with two, three, and four carbon atoms, respectively, in the side chain. The abbreviation "pept fit" indicates that a sequenced peptide has been fitted at this location, and "good fit" indicates a good agreement between the listed amino acid and the difference electron density map calculated from the refined model being analyzed. Electron density maps calculated from the refined model 3 and model 4 are shown in panels b and c of Figures 1, respectively.

ion (Webb et al., 1973) were not included in the refinement nor in the structure factor calculations. If the refinement and partial model phases were aiding the analysis, electron density at the known cofactor binding sites should be more easily interpretable, a situation that was indeed observed with model 2 coordinates (Birktoft et al., 1982a).

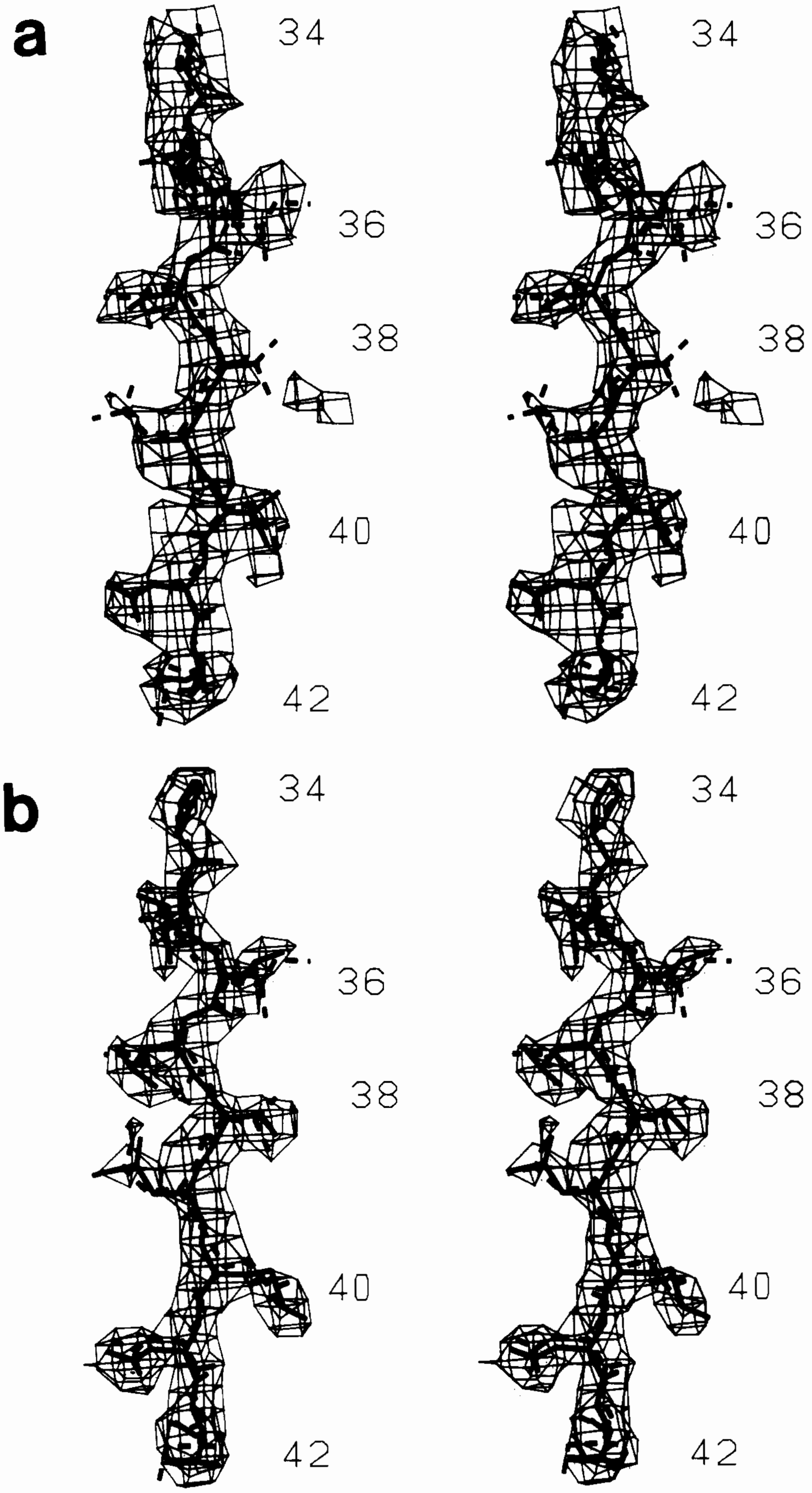
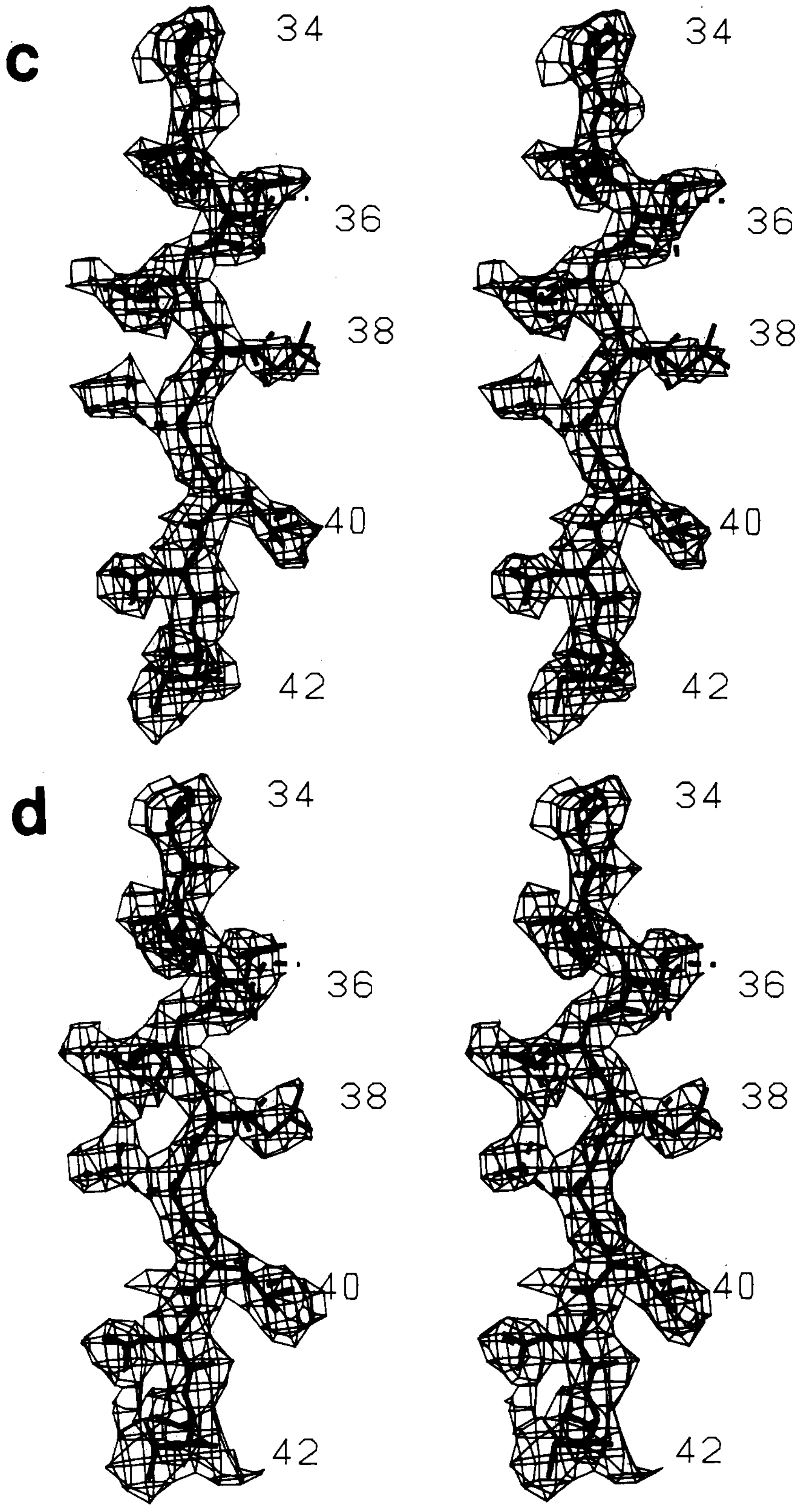


FIGURE 1: Molecular models and corresponding electron density maps used in the X-ray sequencing of sMDH. Each stereodiagram shows residues 34–42 of model 5, drawn in stippled lines, together with the same residues from an earlier model version, drawn in heavy lines. The modes are superimposed on electron density maps that correspond to the model drawn in a heavy line. The maps were contoured such that the polypeptide backbone appeared similar in all panels. (a) Original MIR-phased electron density map: (—) model 2; (- -) model 5. (b)

Fitting of Sequenced Peptides. At this point, the first attempt to incorporate the information from sequenced peptides was made. With the same small segment of the polypeptide chain as an example, residues 34–60 in Tables I and II, a number of correct and incorrect placements were obtained by the probability method described under Materials and Methods. A large peptide (e.g., Pep 3 in Table IVa) containing 35 residues could be fitted at the N-terminal end of sMDH. Residues 34–36 in Table II are part of this peptide. The last six residues included in this table, 55–60, are contained in another successfully fitted peptide (e.g., Pep 12 in Table IVa), but this placement was clearly aided by the presence at position 59 of a binding site for the heavy atom reagent *p*-(chloromercuri)benzoate used for MIR phasing (Hill et al., 1972;



Partial F_{calcd} map. The figure is a composite of two different $2F_{obsd} - F_{calcd}$ maps. The first was calculated by omitting residues 28–37 from the refined model 3 and the second by omitting residues 38–47. The narrow waist between residues 37 and 38 is the dividing point between the two maps: (—) model 3; (--) model 5. (c) $2F_{obsd} - F_{calcd}$ map based on the refined model 4: (—) model 4; (--) model 5. (d) $2F_{obsd} - F_{calcd}$ map based on a refined model where the side chains of residues 37–40 have been omitted: (—) model 4; (--) model 5.

Tsernoglou et al., 1972). This amino acid was presumed to be cysteine. The correctness of the placement of these two peptides was verified upon completion of the chemical determination of the amino acid sequence.

The example given in Table II can also be used to illustrate the difficulties and ambiguities encountered in the fitting of shorter peptides. As eventually became evident upon completion of the sequence, a pentapeptide from the chymotryptic digest of sMDH (e.g., Pep Ch27 in Table IVb) with the sequence BITPM should also have been fitted in this segment with its first residue, Asx, at position 41. The score for the placement of this peptide can be derived from the amino acid sequence scoring matrix exemplified in Table I. The total score for this peptide in this position is 18 or, in normalized form,

Table III: Summary of the Five Models of sMDH Used during the X-ray Sequencing[a]

		no. of amino acids		correctly assigned amino acids (%)		crystallographic R factor (%)		
	no. of atoms	Sub 1	Sub 2	group 1	group 2	9.0–3.5	6.0–2.5	comments
model 1	3195	314	325			nd	nd	no chemical data incorporated; model was a polyalanine structure
model 2	4127	321	324	20	40	41.4	nd	model based on a few peptides of known amino acid sequence, pseudo amino acids, and estimates of MIR electron density map
model 3	4650 (4748)	324	325	52	23	27.2	31.8	improved X-ray sequence based on fitted peptides and refinement of coordinates from a partial model
model 4	5048 (5222)	330	330	97	2	nd	19.8	complete amino acid sequence determined by chemical methods; few residues still in error
model 5	5076 (5250)	331	331	100		nd	19.1	complete amino acid sequence known; two residues still uncertain

[a] The abbreviations Sub 1 and Sub 2 refer to subunits 1 and 2, respectively. Group 1 refers to residues that are identical in the intermediate model and the final structure, model 5. (Valine and threonine, glutamate and glutamine, and aspartate and asparagine are assumed identical.) Group 2 refers to residues that are nearly identical (i.e., leucine, isoleucine, aspartate, and asparagine) or that differ from each other by no more than one atom. The number of atoms does include the atoms in the N-terminal acyl group but not those in the bound NAD and sulfate ions. The numbers in the parentheses are the total number of atoms in the asymmetric unit included in the refinement of the given model. nd = not determined.

$18/(5 \times 9) = 0.40$. To illustrate how the scoring value varies, if this peptide were placed at position 40 or 42, the normalized scores would be 0.04 and 0.07, respectively. An inspection of all possible placements for this peptide identified seven locations that produced higher scores than the correct one. The largest score for these incorrect placements was 0.53. Thus even pentapeptides had a multiplicity of locations in the electron density map. In retrospect, perhaps such ambiguities could have been resolved by calculating a correlation function between a given location and the electron density, but this would require prior fitting of the model at each suggested site in the map. Finally, in order to obtain a complete version of the X-ray sequence, those residues that could not be matched with any sequenced peptide were estimated from the electron density map alone. The most likely residue is listed under the column second revision in Table II.

Tables I and II contain only a partial account of the complete peptide assignments. A more complete accounting of the peptide fitting procedure is presented in Table IV. The first of these, Table IVa, is a summary for those peptides that were fitted into the model (model 3) whereas Table IVb describes the results obtained with the peptides that were not fitted. Some of the peptides listed in Table IVb were only partially sequenced. Of the nine fitted peptides, which ranged from 4 to 36 amino acids in length, six could be placed with a high degree of confidence and three with somewhat less certainty. An additional three peptides (i.e. Pep 5a, 6, and 9a) could also have been assigned (see Table IVb for further details). One fitted peptide, Pep 15, does not appear to belong to sMDH. It was incorporated into model 3 because it had been assumed that this peptide contained the active site histidine and further that of all the histidine-containing peptides this one gave the best overall fit near the putative active site histidine.

Pep 4a (Table IVb) illustrates what can go wrong when peptides with sequence errors are combined into a larger peptide. The first 15 residues of Pep 4a does not belong to sMDH, and the last 13 residues are the result of an incorrect combination of two shorter peptide fragments. The AZIALK sequence is located at position 164, and although no attempts had been made to fit the peptide, it could have been placed there. The other fragment had too many sequence errors to be usable.

The acceptance or rejection of a peptide location was generally done by reinspection of the probability table (Table I). Recall that this table contains a subjective numerical estimate of the electron density map itself. Therefore, the placement of a peptide leading to the positioning of a small side chain into a location observed to be of sufficient volume to accommodate a larger residue was a cause for rejection. Conversely, a placement would be rejected if the side chain was much larger than the density it was supposed to fit into. It was, however, kept in mind that side chain associated electron densities may appear truncated and particularly so for surface residues. Placement of charged residues in internal locations was considered to be unlikely, as was the placement of several nonpolar residues from the same peptide in external positions.

In retrospect, wherever a peptide could not be fitted to the map, most often the difficulties were associated with errors in the chemical sequence or errors in the preliminary model. The latter were particularly troublesome if main-chain atoms were badly placed in the model, causing an incorrect number of residues to be incorporated.

The sequence data used here are associated with a considerable amount of errors, which is a reflection of the relatively primitive methodology employed during their collection. Inadequate deduction (subtractive Edman dansyl techniques) and insensitive purity analysis resulted in the accumulation of substantial errors; these were, of course, subsequently eliminated by the requisite redundancy in assignment. Vastly improved automated methodology also served to "refine" these data. However, in the interest of accurately assessing the "fitting" technology, the data base of Table IV is presented without retrospective refinement.

Two types of sequence errors were notable. In some cases residues were interchanged, which caused a lowering of the scoring value and more importantly placed residues in positions where they clearly were in conflict with the electron densities. Of more severe consequences was the omission or inclusion of extra residues, which in general caused a substantial reduction of the scoring as well as poor correlation with the electron density (e.g., Pep 8 and Pep 4 in Table IVb).

Table IV also gives the scoring values obtained when the correct chemical sequence was utilized. It should be mentioned that this information was not available until the refinement of model 3 was completed. Table IV also demonstrates that when the correct peptide sequence is employed, the highest score corresponds to the correct position for the peptide and that the difference between the highest and next highest score is larger than when an incorrect sequence is used.

Had the correct sequence been available most of the longer peptides listed in Table IV most probably could have been fitted. The placement of many of the shorter peptides on the other hand may have remained uncertain due to the high number of high-scoring possibilities. However, the placement of the longer peptides would have eliminated many such locations. Finally, in retrospect, it can be seen that the criteria used for accepting potential peptide placements were probably too restrictive. Pep 5a, 6, and 9a were needlessly rejected as being unacceptable.

Overall, of the 331 amino acids in sMDH, 172 or 52% were correctly identified in model 3. Of these, 123 residues were located by using peptides of known chemical sequence, and 49 were estimated from electron density maps alone. This summary is based on the isologous amino acids being grouped together: asparagine and aspartic acid, glutamine and glutamic acid, and valine and threonine. An additional 18 residues (5%) were correct with respect to the number of atoms, that is, leucine for asparagine or isoleucine, etc., and another 61 residues (18%) were correct within one atom, such as tyrosine vs. phenylalanine, valine vs. isoleucine, etc.

Crystallographic Refinement of the Second X-ray Sequenced Model (Model 3). The structure resulting from the second sequence revision, model 3, contained 4650 atoms distributed into 324 amino acids in subunit 1 and 325 amino acids in subunit 2. When the 98 atoms located in the NAD and sulfate ions bound to each subunit are included, the total is 4748 atoms per asymmetric unit (Table III). With utilization of X-ray data between 6.0-Å and 2.45-Å resolution, the restrained least-squares refinement was continued for an additional 30 cycles, at which point the conventional crystallographic R factor was 31.8%. The refinement was followed by a refitting of the molecular model with partial F_{calcd} maps, which were calculated by omitting 10 residues from the structure and using the remaining atoms to generate structure factor amplitudes and phases. The fit of the refined structure to these partial F_{calcd} maps was evaluated by inspecting the regions of the omitted residues. The X-ray sequence was modified as suggested by the difference maps. Examples of this reevaluation are shown in Table II under the heading third revision.

Incorporation of Complete Chemical Amino Acid Sequence. At this point, the determination of the amino acid sequence by chemical means began to yield additional overlaps and longer peptides that could be incorporated into the X-ray sequence with less difficulty. The original sequence scoring matrix based on the MIR map was again used but was revised such as to take into consideration deletions and in particular additions of residues as suggested by new electron density maps and models. The fitting procedure suggested that some of these peptides were created by an erroneous combination of two shorter peptides. Not only did the fitting procedure indicate incorrect peptide combinations but equally importantly alternate and more plausible combinations were also suggested. Subsequently, all turned out to be correct.

The refined model 3 and the essentially completed amino acid sequence were combined yielding model 4. This was accomplished by maintaining the polypeptide backbone structure of the refined model 3 and making the necessary replacements of the amino acid side chains. The generation of model 4 was followed by a regularization, both procedures utilizing the Hermans–McQueen program REFINE (Hermans & McQueen, 1974). In those instances where changes in the length of the polypeptide were required, deletions and insertions were made without attempting to maintain proper peptide bond stereochemistry. Instead, the model adjustments were done by using the electron density map in the same manner that was used for the adjustment in orientation of any replaced amino acid side chain.

Each subunit of model 4 contains 2524 atoms distributed among 331 amino acids. Together with the atoms in the cofactor NAD and the bound sulfate ions, there is a total of 5146 atoms in the sMDH dimer contained in the asymmetric unit (Table III). By use of the partial F_{calcd} maps derived from the refined model 3, the model 4 structure was then adjusted on the graphics system, followed by refinement with the restrained least-squares procedure. At several stages the model was regularized and then readjusted against difference Fourier maps based on the model prior to the most recent regularization. During the latter refinement cycles, the dyad symmetry between the subunits was included as a restraint. Finally, in the last two refinement cycles, 76 solvent molecules, assumed to be water, were included in the model. Individual temperature factors were also assigned to each atom and allowed to refine isotropically. At this stage the R factor is 19.8% for a model in which the rms deviations from canonical values for bond length and bond angles are 0.06 Å and 10.1°, respectively.

Resolution of Conflicts between the Chemically Determined Amino Acid Sequence and the Electron Density Maps. Even though the chemically determined amino acid sequence was generally in good agreement with the X-ray results, some uncertainty was still associated with three segments of the polypeptide chain. The first region in question is at residues 87–88 (Table V), where the chemical data suggested a Gly-Ser sequence, whereas a Ser-Gly or less likely an Ala-Gly sequence is more compatible with the X-ray data. The Ser-Gly sequence was incorporated into model 4, and a comparison between the model and electron density map indicated that this was correct.

The second region involving lack of correlation between X-ray and chemical results involved the residues located between Ile-35 and Asp-41, which is part of the protein used as an example in Tables I and II and illustrated in the stereo panels in Figure 1. The initial amino acid sequence from chemical analysis was not unambiguous in this region but suggested the following sequence: Leu-Gly-Gly-Leu. Both the original MIR map (Figure 1a) and the electron density maps based on the refined model 2 indicated that an additional residue was needed in this region of the model and further that a leucine was the most probable residue to be accommodated at position 36. Partial F_{calcd} maps based on the refined model 3 (Figure 1b) suggested that the three residue types suggested by the chemical sequence data for positions 36–40, leucine, glycine, and glutamate, best could be accommodated as a Leu-Leu-Glu-Gly-Leu sequence, which was built into model 4. However, as can be seen in Figure 1c, difference electron density maps based on the refined model 4 suggested that residue 38 was smaller than glutamate and residue 39 larger than glycine. The sequence uncertainty for residues 36–40 was further explored by crystallographic least-squares refiement of a modified molecular model. The side chains of the four residues in questions were were removed from the refined model 4, and after 5 cycles of refinement the resulting coordinates were used to calculate difference electron density maps (Figure 1d). Analysis of these showed that the sequence Leu-Leu-Gly-Leu-Glu was most compatible with both X-ray and chemical data. The sequence Leu-Leu-Ser-Leu-Leu did, however, appear more compatible with the electron density map, as can be seen in Figure 1d. The final chemical analysis showed that the sequence for residues 36–40 is Ile-Leu-Val-Leu-Leu, which is in excellent agreement with the difference electron density maps.

The third region, where a sequence revision was necessary, consists of residues 199–214. These residues are located on the molecular surface where they are folded into two hairpin loop structures. When the chemical sequence was incorporated

Table IV: Summary of the Placements of Peptides in the Electron Density Map of sMDH

ID	top scores for original sequence[a]	top scores for correct sequence[b]	sequence of peptides[c], errors in sequence if any, and comments about fitting of peptides
			(a) peptides that were fitted.
Pep 1	0.53 0.30 0.27 294 4 254	0.49 0.27 0.26 294 252 151	KTWKIVEGLPIBDFSREKMNLTAKELAZEKETAFEKBBA (41 residues); the last seven residues are incorrect, but the error did not influence the fitting
Pep 2	0.51 0.38 0.35 125 6 153 297	0.48 0.41 0.38 125 297 6	VIVVGBPATNNCLTASK (17 residues); two residues were inverted in the peptide sequence; the TN sequence should have been NT
Pep 3	0.36 0.27 0.26 2 147 1 279 250 295	0.46 0.26 0.25 1 146 188	ISZPRVLVTGAAGZIAYSLLFTIGDGSVFGKNZPIL (36 residues); two changes were made to the original chemical sequence in order to get better fit to the density; AGZ was changed to AZ and YS to FT; the last residues are PIIL rather than PIL
Pep 5	0.43 0.41 0.40 296 46 153 170	0.51 0.43 0.42 170 268 296	LGVTBBVSKB (10 residues); fitting was uncertain due to the large number of possible placements; position 170 was considered most likely for placement; however five residues were changed to provide better fit to density; four residues were permutated; the BBVS sequence should have been SBBV.
Pep 5b	0.31 0.28 0.26 156 97 205	0.33 0.31 0.28 97 156 205 250 301	RKDLLKABVKIFKCQGAALBKYWKKSVKW (29 residues); the three highest placements all appeared equally likely, but two had conflicts with other fitted peptides; the first four residues were changed to get a better density fit as were the last two residues; the first of the two Trp's is Ala, but Trp fits the density better; the other Trp is Val as had been fitted into the model
Pep 12	0.44 0.38 0.35 55 304 159	same	ELZBCALPLLK (11 residues); only Pro did not fit well
Pep 13	0.64 0.56 0.51 157 305 250	same	LBHBR (5 residues)
Pep 14	0.78 0.72 0.64 181 52 266	same	VSMG (4 residues); the peptide was not placed at highest scoring position, due to conflict with placement of the presumed active site histidine peptide, Pep 13; other possible placements resulted in poor agreement between model and electron density; peptide was placed at position 266 but VS was replaced with AZ to get better fit to density
Pep 15	0.68 0.46 0.43 183 180 253	n/a	IGZHGQ (6 residues); the peptide was incorporated at position 183, since it had been assumed to be the peptide containing the active site histidine; the Ile gave a poor fit and was replaced with Tyr; however, this peptide cannot be located in the final chemical sequence, and does not seem to belong to the protein
			(b) Peptides that were not fitted.
Pep 4	0.36 0.32 0.31 124 179 211	0.46 0.36 0.29 179 289 271	BVIHWGNPSSYZBBVHTAK (19 residues); one residue was missing and 10 were incorrect, four due to the omitted residue; all suggested placements gave poor density agreement, particular at the location for the active site histidine which is located in this peptide
Pep 4a	0.31 0.30 0.29 283 142 134	n/a	RTPVDPSAIVPLNKKQAMDLTKAZIALK (28 residues); the first half of this peptide does not seem to belong to the protein; see text for further details
Pep 5a	0.36 0.33 0.32 65 295 286	0.38 0.34 0.32 65 156 281	KDVIATNQKEIAFKBLBVAIL (21 residues); the peptide should have been fitted. In the original model two residues were missing in this region, and the sequence for a couple of residues is wrong; these errors do not affect the scoring values noticeably; the missing residues are located in a loop region on the protein surface
Pep 6	0.37 0.32 0.30 139 244 237	0.42 0.31 0.30 139 284 242	KKASAPSIPKEBFSCLTRL (19 residues); the peptide should have been fitted; residues at the beginning and in the middle of the peptide gave poor fits for the top score, and other possible placements also seemed unlikely; the sequence for the first three residues is wrong
Pep 7	0.39 0.38 0.36 23 277 153 128	0.40 0.37 0.36 277 258 128 153	LGVPBBLLYSB (11 residues); the peptide was not fitted due to a large number of possible placements, all with near equal scores; all gave relatively poor agreement with the density for several residues; the first and last residues had the wrong sequence
Pep 8	0.44 0.36 0.32 287 247 164	0.49 0.31 0.30 247 206 146 289	KAICVHBRBWK (11 residues); the peptide was not fitted due to poor agreement for several residues; the sequence for five residues is wrong and one residue is missing
Pep 9a	0.46 0.40 0.39 219 150 178	same	KGZFITTVZZRG (12 residues); the peptide should have been fitted; the Phe did not agree too well, but not bad enough to cause rejection; all other placements gave poor fits to the density
Pep 9b	0.52 0.46 0.37 288 181 16 26 50 112 274	0.54 0.52 0.46 287 246 288	GAVIWAK (7 residues); all the high scoring possibilities showed poor agreement with the density for one or more residues; the sequence of the peptide is quite wrong and should have been GAAVIK; even with the correct sequence the score is only 0.22 for placement at the correct position, at 230

Pep 10	0.37 41	0.35 20	0.33 2 288	0.36 208	0.35 199	0.34 39	LQAKEVGVYEAVKBBSWLK (19 residues); the original model was uncertain in this region, and later turned out to be missing two residues; the correct placement for the peptide shown would be at 201 with a score of 0.31; all suggested placements gave too many poor fits; the chemical sequence is still uncertain in this region; see text for further details
Pep 19	0.52 179	0.42 25	0.40 50	n/a			BGVLLGBZGR (10 residues); the peptide was created incorrectly from three shorter peptides; only one of these was sequenced; the first four residues are located at position 50 (see under peptide Ch 29 in this table), but the remainder could not be placed anywhere
Th 54	nd			0.49 296	0.44 321	0.39 258	FGTFPZ(Z,G) (8 residues); the sequence had numerous errors, and no reasonable placement was suggested; the correct sequence is FGTPEGEF; two permutations of the Z,G composition are possible and the top scores for these are 0.40, rank 32nd, and 0.44, rank 11th, respectively
Sp 21	0.51 151	0.48 296	0.41 152	same			FPVTIKD (7 residues); the score at the correct location, at 287, is 0.39 which ranks 4th; of the top 10 scores, 7 would be acceptable
Th 41	nd			0.72 80	0.48 2	0.43 121	LBVAI(G,B) (7 residues); the sequence was in error and should be LBVAILVG; for either permutation of the two unsequenced residues the score for the correct placement ranked second; however these as well as all other reasonable placements were rejected due to density conflicts; when the first five residues only were used the top score is 0.62 at position 80
Th 5	nd			0.51 48	0.43 80	0.33 300	VLBG(Z,M) (6 residues); the sequence was in error and should have been VLDGVLME; the top scores for the two possible permutations of the sequence are 0.48, rank 15th, and 0.54, rank 9th; all suggested placements were rejected because of density conflicts
Th 36	0.61 155 295	0.46 277	0.44 276	0.53 212	0.50 291	0.48 168 299	VWKBSB (6 residues); the sequence was in error and should have been VKDDSW; all of the suggested placements resulted in poor density correlation
Tr 20	0.57 266 316	0.48 285	0.46 52 247	0.43 296	0.42 239	0.41 170	SAMASK (6 residues); the sequence was quite wrong and should have been LSSAMSAAK; the score for the correct location is 0.37 and ranks 14th; all suggested placements gave some density conflicts
Ch 27	0.53 179	0.44 223	0.42 7 50 153 224	same			BITPM (5 residues); the fit of this peptide has been described in detail in the text; the correct location is at 41 with a score of 0.40 which ranks 8th; of the top 10 possibilities all but 2 were questionable due to conflicts with longer fitted peptides; ignoring this 7 of the top 10 could have been fitted
Sp 12	0.49 268	0.47 84	0.44 169	0.56 217 268 301 328	0.52 32	0.48 63 80 175 257	FLSS(A) (5 residues); the last two residues do not belong and the sequence should be FLS; the correct position is at 329, which is located at the C-terminus of sMDH; the MIR map and scoring assignments are poor for this part of sMDH; 5 of the top 10 placements could have been fitted
Th 38	0.75 157	0.56 299	0.53 170	0.70 168	0.67 31 155 177	0.59 166	VRBB (4 residues); the sequence was wrong and should have been VRB; the score for the peptide at the correct position, at 155, is 0.44 and ranks 9th; of the top 10 positions 9 were acceptable
Th 48	0.67 242	0.50 317	0.47 236	same			AKKS (4 residues); peptide belongs to Pep 5b (part a) but was not included due to the wrong inclusion of a Trp in that sequence; peptide could have been placed in numerous positions; of the top 10 positions 6 were acceptable; the right location is at 119, but the score is only 0.22 which ranks 85th
Th 53	0.53 162 187	0.50 160 197 270	0.46 246 246	same			AKAZ (4 residues); numerous acceptable placements; peptide fitted at location 162, but sequence was changed to AKSM to get better fit to the electron density
Ch 29	0.69 82	0.53 25 46 206 302	0.50 15 128	same			BGVL (4 residues); peptide was included in Pep 19 which could not be fitted (see above for further details); the score at the correct location, at 50, is 0.47 and ranks 9th; the top 10 positions 7 were acceptable
Th 12	0.74 134 135	0.59 158 79	0.56 80 222	same			LLB (3 residues); the score at the correct location at 39, is 0.33 and ranks 40th; most of the high scoring possibilities were acceptable, but none were fitted due to their multitude

[a] The top three scoring values, in normalized form, are given for for each peptide, with the sequence location listed underneath. More than one location for a given score may be given. The sequence numbers correspond to those used in Table V. The correct location for the peptide is underlined. "nd" refer to those peptides, that were only partially sequenced. For these all possible permutations were tried. Their scoring is discussed in the last column. [b] The "correct sequence" is the one incorporated into model 5 and is listed in full in Table V. "same" means that the original sequence was correct, and "n/a" means that the peptide could not be located in the final chemical sequence (Table V). [c] The sequences listed are those derived from the "raw" peptide data base, as described under "Peptides Used in Fitting". In some instances the peptide sequence was changed when incorporated into model 3, and the listed sequence differ from the X-ray sequence in Table V. Such cases are discussed for each peptide.

Table V: Amino Acid Sequence of Cytoplasmic Malate Dehydrogenase[a]

```
Residue No:         1       5         10        15        20        25        30        35        40        45        50        55

X-ray Seq:       ac s p z i r v l v t g a a - q l a f t l l y s i g b g s v f g k b z p i l l s l m d v v p k z z t s z a v n m z
                    : : : : : : : : : : : :   : : :     : : : : : : : : : : : : : : : : :   :   :   :   : :       :       :   : :
Chem Seq:        Ac S P E I R V L V T G A A G Q I A Y S L L Y S I G N G S V F G K D Q P I I L V L L D I T P M M G V L D G V L M E
Enviroment:                   I I I               I     I       I I       I I           I   I I
Dimer:                                        Q     Q Q   Q Q Q     Q                                           Q Q   Q Q Q   Q Q
Substrates:                         N N   N N N N                                                 N N N     N
                       |---------------|   |---------------------------|             |-------------|       |----------------------
                             beta A                 alpha B                              beta B                   alpha C


Residue No:                 60        65        70        75        80        85        90        95        100       105       110

X-ray Seq:          l z b c a l p l l k s z f g k b s g b y a - - s z b v g v l l a g z s a k b - - - a a k b l k a b v k i f k c z
                    : : : : : : : : : :           :         :         : :     :     :         :               : : : : : : : : : : :
Chem Seq:           L Q D C A L P L L K D V I A T D K E E I A F K D L D V A I L V S G M P R R D G M E R K D L L K A N V K I F K C Q
Enviroment:                                                 I I     I   I I I I                                     I I     I
Dimer:              Q Q Q Q Q Q Q     Q
Substrates:                                               N                         N N N N N N                               N
                    -------|     |---------------|       |-------------|-----------|                               |----------------
                                      beta C                 alpha C’      beta D                                         alpha D


Residue No:               115       120       125       130       135       140       145       150       155       160       165

X-ray Seq:          g a a l b k y w k k s v i v i v v g b p a t b b c l t a s k b s a z l b k a k z v b s v k l b h b r a k s m l s
                    : : : : : : :   : : : :   : : : : : : : :     : : : : : : : :           :             :   : : : : : : :
Chem Seq:           G A A L D K Y A K K S V K V I V V G N P A N T N C L T A S K S A P S I P K E N F S C L T R L D H N R A K A Q I A
Enviroment:         I   I I       I       I   I I I         I I   I I   I I       I     I         I   I   I     I       I       I
Dimer:                                                                                                        Q     Q Q   Q Q
Substrates:                                         N N N   N                                           N     N       N
                    -------------|         |---------|   |-------------------------|         |---------|   |------------------------
                       alpha E               beta E               alpha 1F                     beta F              alpha 2F


Residue No:             170       175       180       185       190       195       200       205       210       215       220

X-ray Seq:          z k l g b s p k l s k b v i l y g z h g z s z f g t l i z l z l - z b k z s a g v r - a s k b z s w k t s i y b
                      : : :             : : : :     :   :       :                 :   :     : :   : :     :   : :   : :
Chem Seq:           L K L G V T S D D V K N V I I W G N H S S T Q Y P D V N H A K L K Q A A K E V G V Y E A V K D D S W L K G E F I
Enviroment:                           I     I I I I I I             I I       I   I                   I                 I       I
Dimer:                          Q Q Q
Substrates:                                       N
                    -------|     |-----------------|             |-----|       |---------|   |---------|           |----------------
                                       beta G                     beta H                      beta J                    alpha 1G


Residue No:           225       230       235       240       245       250       255       260       265       270       275

X-ray Seq:          b v i z z g g v v h v z a r t a b b s m k t g f a l b l y v k h l w k g i - s z k l a z m g l i a h g - k a a a
                          : :   :           : :           :         :         :       :   :                 : :   :     :
Chem Seq:           T T V Q Q R G A A V I K A R K L S S A M S A A K A I C D H V R D I W F G T P E G E F V S M G I I S D G N S Y G V
Enviroment:             I                                           I I     I I     I       I             I I I I   I
Dimer:                        Q     Q Q     Q Q Q     Q Q Q Q   Q Q     Q
Substrates:                     N N     N         N N N       N
                    -----------|---------------|       |-------------------------------|             |-------------|
                      alpha 1G     alpha 2G                       alpha 3G                               beta K


Residue No:         280       285       290       295       300       305       310       315       320       325       330

X-ray Seq:          s p k z b f s c v t r l z n k t w k i v e g l p i b d f s r e k m n l t a k e l a z z z k t e f a e k b s n a
                                :     :       : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :     :   :
Chem Seq:           P D D L L Y S F P V T I K D K T W K I V E G L P I N D F S R E K M D L T A K E L A E E K E T A F E F L S - - -
Enviroment:                 I I I I I I   I             I                   I             I                     I
Dimer:
Substrates:
                       |---------------------|   |-----|             |-----------------------------------------------------|
                              beta L              beta M                                 alpha H
```

[a] The table lists both the X-ray-refined chemically determined amino acid sequence and the X-ray amino acid sequence used in the refinement of model 3. The numbering system is based on assigning the first amino acid residue to position 1. The symbol ":" is used to indicate residues that are the same in the two amino acid sequences. The following residues cannot be distinguished and were assumed to be identical for this purpose: valine and threonine; aspartate and asparagine; glutamate and glutamine. The symbols below the sequence data are flags that indicate special interactions for certain positions as follows: "I" for residues whose side chains in the isolated subunit are removed from contact with the external solvent; "Q" for residues that are involved in subunit–subunit interactions; "N" for residues that are in contact with substrates or coenzyme. Elements of secondary structure are also indicated, according to the customary nomenclature for this protein (Hill et al., 1972; Banaszak & Bradshaw, 1975).

into model 3, it was noted that several of these residues did not fit the electron density of the partial F_{calcd} maps particularly well and that the agreement with the X-ray sequence was much worse than in most other regions. Despite these warning signs, the suggested chemical sequence was incorporated into model 4 because the electron density in this region was particularly troublesome and did not suggest any obvious alternatives. This region was analyzed further by removing all side chains from residues 200–215 of the refined model 4, carrying out five cycles of restrained least-squares refinement and analyzing the difference electron density map calculated from the resulting coordinates. The best overall agreement between the proposed chemical sequence and difference electron density maps is obtained if two additional residues are incorporated into the model. They have tentatively been identified as lysine and alanine and occupy positions 200 and 203 in the sequence listed in Table V. Considering the external location of both residues and the marginal quality of the electron density map in this region, their identification is somewhat uncertain. The chemical sequence in this area is now under investigation.

These three revisions, summarized above, were incorporated into the refined model 4, resulting in model 5. Several more cycles of refinement have been performed on this model, and currently the *R* factor is 19.1%. This model still possesses acceptable stereochemistry with the rms deviation from standard values of bond length and bond angles being 0.06 Å and 8.4°, respectively.

The amino acid sequence of cytoplasmic malate dehydrogenase from pig heart is given in Table V. Included in this table is the X-ray sequence derived by fitting sequenced peptides into the electron density map and used in the model 3 refinement. A full discussion of the structural features of the molecular model of sMDH will be published elsewhere.[3] However, we have indicated in Table V in symbolic form the location of elements of secondary structure, using the nomenclature introduced previously for this enzyme (Hill et al., 1972; Banaszak & Bradshaw, 1975). Also marked in the table are the residues that are involved in subunit–subunit interactions as well as those that are in contact with the cofactor NAD or the anion believed to be located at the substrate binding site. The structure–function aspects of sMDH have been discussed in detail in previous publications (Birktoft et al., 1982a; Birktoft & Banaszak, 1983). However, the amino acid numbering system has changed as a result of the chemical sequence determination of sMDH. In order to provide a point of reference to previous publications, the important active site residues are as follows (with the previously used residue numbers in parentheses): histidine-186 (180), aspartate-158 (152), and arginine-161 (155).

Discussion

Like other investigators, we have found that X-ray crystallographic data alone are insufficient in most instances to identify the amino acids belonging to an otherwise easily traceable polypeptide chain (Anderson et al., 1978b; Watenpaugh et al., 1973). However, it is possible to produce a subjective table of probabilities for each location in the electron density map. Chemical sequence information on peptide fragments may aid in identifying side chains in an electron density map, but often several locations in the map can be assigned to a given partial sequence. At 2.5-Å resolution, partial models, refinement procedures, and electron density maps based on the incomplete model help and, at least for sMDH, were useful in resolving some of the amino acid sequence ambiguities.

It was our experience that longer peptides, Table IVa, could be fitted into the model with less difficulty, whereas shorter peptides, Table IVb, i.e., five residues or fewer, in general yielded too many possibilities of near equal probability. In a few cases for longer peptides, a probable location included a few residues placed in a position that was in poor agreement with the scoring matrix and the electron density maps. In such instances, if the peptide otherwise could be accommodated at this location in the electron density map, preference was given to the residue suggested by the electron density rather than by the chemical sequence.

For brevity sake, the history of this study was illustrated with a 27 amino acid segment, spanning residues 34–60, that is presented in Tables I and II, and a portion of this segment of the polypeptide chain is presented in Figure 1. These stereodiagrams show the models and the electron density maps as they progressed through various stages of the fitting study. At either end of this segment, sequenced peptides were placed in model 2 with the probability tables. However, the chemical identity of the middle 18 residues was not confirmed through chemical sequencing until the complete amino acid sequence was known. Of these residues, six were correctly identified in model 3, and this number increases to eight if valine and threonine are assumed to be identical. The third revision suggested the correct amino acid for all but one of the remaining 10 residues, although sometimes more than one choice was still possible. The lone incorrect assignment was serine for valine at position 38, an error of only one atom. The identification process for this particular segment of sMDH was clearly aided by the partial model refinement and model phases. Even though we have chosen just a short segment as an example, the experience described for these 27 residues is fairly representative for the sMDH structure as a whole.

As has been discussed above, the subsequent analysis of this region did indeed improve the sequence assignments. For the sMDH structure as a whole, it is our feeling that subsequent cycles of revision of the sequence probability table on the basis of the shape of the side-chain density and refinement would have improved the sequence, but the process was tedious without further automation.

Some conclusions can be reached about the procedures used to fit partial chemical data and employment of refinement methods. The overall quality of electron density maps calculated from even a partial structure appeared higher than the original MIR map as can be seen in comparing panel a with panel b of Figure 1. This was most apparent when electron density maps were analyzed at the two cofactor positions, noting that the coordinate set used in the refinement of model 2 and subsequent phase calculations did not include contributions from the NADs. This can be seen in Figure 1 in Birktoft et al. (1982a).

The probability matrix scoring procedure that we employed, part of which is given in Table I, facilitated examination of chemically sequenced peptides as they became available. In the studies described here, many short peptides had to be compared with the X-ray results at various stages of the study. However, the same method might facilitate the fitting of a complete amino acid sequence to an electron density map. When attempting to interpret a new electron density map, it is not unusual to have incorrect chain segment connectivity, breaks in electron density, frame shifts of one or two residues, or even, in rare instances, errors in the chemical sequence. Many of these would probably be obvious in the comparison matrix, and even more importantly the corrective action should be apparent.

Acknowledgments

We thank Drs. E. Hill, D. Tsernoglou, and L. E. Webb for collecting the X-ray diffraction data and for their contribution to the initial interpretation of the sMDH electron density map. We thank Drs. W. A. Hendrickson and P. Briley for supplying the Konnert–Hendrickson least-squares refinement programs and their advice in their use. We gratefully acknowledge the

[3] J. J. Birktoft and L. J. Banaszak, submitted for publication to *J. Biol. Chem.*

excellent assistance of Suzanne Winkler and Sophie Silverman in the preparation of the manuscript.

REFERENCES

Allen, L., Vanecek, J., & Wolfe, R. G. (1971) *Arch. Biochem. Biophys. 143*, 166–174.

Anderson, C. M., McDonald, R. C., & Steitz, T. A. (1978a) *J. Mol. Biol. 103*, 1–13.

Anderson, C. M., Stenkamp, R. E., & Steitz, T. A. (1978b) *J. Mol. Biol. 123*, 15–33.

Banaszak, L. J., & Bradshaw, R. A. (1975) *Enzymes (3rd Ed.) 11A*, 369–396.

Barry, C. D., Bosshard, H. E., Ellis, R. A., & Marshall, G. R. (1974) *Fed. Proc., Fed. Am. Soc. Exp. Biol. 33*, 2368–2372.

Birktoft, J. J., & Banaszak, L. J. (1983) *J. Biol. Chem. 258*, 472–482.

Birktoft, J. J., Fernley, R. T., Bradshaw, R. A., & Banaszak, L. J. (1982a) in *Molecular Structure and Biological Activity* (Griffin, J. F., & Duax, W. L., Eds.) pp 37–55, Elsevier/North-Holland, New York.

Birktoft, J. J., Fernley, R. T., Bradshaw, R. A., & Banaszak, L. J. (1982b) *Proc. Natl. Acad. Sci. U.S.A. 79*, 6166–6170.

Hendrickson, W. A., & Konnert, J. H. (1981) in *Biomolecular Structure, Function, Conformation and Evolution* (Srinivasan, R., Ed.) Vol. 1, pp 43–57, Pergamon, Oxford.

Hermans, J., & McQueen, J. E. (1974) *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr. A30*, 730–739.

Herriott, J. R., Watenpaugh, K. D., Sieker, L. C., & Jensen, L. H. (1973) *J. Mol. Biol. 80*, 423–432.

Hill, E., Tsernoglou, D., Webb, L., & Banaszak, L. J. (1972) *J. Mol. Biol. 72*, 577–591.

Hopper, P., Harrison, S. C., & Sauer, R. T. (1984) *J. Mol. Biol. 177*, 701–713.

Lederer, F., Glatigny, A., Bethge, P. H., Bellamy, H. B., & Mathews, F. S. (1981) *J. Mol. Biol. 148*, 427–448.

Lipscomb, W. N., Hartsuck, J. A., Quiocho, F. A., & Reeke, G. N. (1969) *Proc. Natl. Acad. Sci. U.S.A. 73*, 2619–2623.

Matthews, B. W., Colman, P. M., Jansonius, J. N., Titani, K., Walsh, K. A., & Neurath, H. (1972) *Nature (London), New Biol. 238*, 41–43.

McLachlan, A. D. (1972) *J. Mol. Biol. 64*, 417–437.

Miller, J. R., Abdel-Meguid, S. S., Rossmann, M. G., & Anderson, D. C. (1981) *J. Appl. Crystallogr. 14*, 94–100.

Rao, S. T., & Rossmann, M. G. (1973) *J. Mol. Biol. 76*, 241–256.

Taylor, S. S., Oxley, S. S., Allison, W. S., & Kaplan, N. O. (1973) *Proc. Natl. Acad. Sci. U.S.A. 70*, 1790–1794.

TenEyck, L. F. (1973) *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr. A29*, 183–191.

Tsernoglou, D., Hill, E., & Banaszak, L. J. (1972) *J. Mol. Biol. 69*, 75–87.

Watenpaugh, K. D., Sieker, L. C., Herriott, J. R., & Jensen, L. H. (1973) *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem. B29*, 943–956.

Watson, H. C. (1969) in *Progress in Stereochemistry*, Vol. 4, pp 229–333, Butterworth, London.

Webb, L. E., Hill, E. J., & Banaszak, L. J. (1973) *Biochemistry 12*, 5101–5108.

Wright, C. S., Gavilanes, F., & Peterson, D. L. (1984) *Biochemistry 23*, 280–287.

Selective Isolation of Human Plasma Low-Density Lipoprotein Particles Containing Apolipoproteins B and E by Use of a Monoclonal Antibody to Apolipoprotein B[†]

Eugen Koren,[*] Petar Alaupovic, Diana M. Lee, Nassrin Dashti, Hans U. Kloer,[‡] and Gary Wen

Lipoprotein and Atherosclerosis Research Program, Oklahoma Medical Research Foundation,
Oklahoma City, Oklahoma 73104

Received August 21, 1986; Revised Manuscript Received January 8, 1987

ABSTRACT: A monoclonal antibody to human plasma apolipoprotein B was used in a single-step immunoaffinity chromatography procedure to isolate a subpopulation of low-density lipoprotein particles from normolipidemic human plasma. The isolated particles were homogeneous in terms of size (20 nm), flotation coefficient ($S_f = 9.5$), and electrophoretic mobility (β band). Their protein moiety consisted of apolipoproteins B and E in a molar ratio close to 2. The lipid moiety consisted of 47.3% cholesterol, 4.7% triglycerides, and 48.0% phospholipids. To indicate its characteristic apolipoprotein composition and hydrated density properties, this family of particles was named LP-B:E_{L2}. In most normolipidemic subjects, LP-B:E_{L2} particles accounted for less than 10% of the total plasma apolipoprotein B content. The LP-B:E_{L2} particles bound to the membranes of the human hepatoma HepG2 cells in a specific and saturable manner indicative of receptor-mediated binding. Their binding was significantly higher than that of low-density lipoprotein particles containing only apolipoprotein B.

Human plasma low-density lipoprotein (LDL) are considered to be potentially the most atherogenic of all plasma lipoproteins (Fredrickson et al., 1978). In fact, the recently released results of the Lipid Research Clinics trial have shown that reduction of LDL–cholesterol is associated with a significant decrease in the morbidity and mortality of coronary heart disease (Lipid Research Clinics Program, 1984). Physicochemically, LDL represent a system of particles heterogeneous with respect to hydrated density, size, and lipid

[†] This work was supported by USPHS Grant HL-23181.

[*] Correspondence should be addressed to this author.

[‡] Present address: Medizinsche Poliklinik, Universitaet Giessen, Rodthohl 8, D-63 Giessen, West Germany.

0006-2960/87/0426-2734$01.50/0