$(MeCO)_2]^{2-}$, shows no fold (Adamson, Daly & Forster, 1974).

Problems with small crystals

The difficulties encountered in recording X-ray diffraction data for structure determination from very small crystals arise from the poor crystal quality as well as from the low intensity of the diffraction pattern. The poor crystal quality is indicated by the large mosaic spread, evaluated as *ca* 3° by *MADNES*. The diffraction spots become very extended in reciprocal space, and the derivation of good integrated intensities, especially for the weaker spots is difficult, as already described by Andrews *et al.* (1988). Improvements in the experimental strategy and processing software may allow somewhat better intensity measurements and structure determination in the future, but it is clear that even with the present procedures sufficient data can be recorded from such a small poor crystal to establish the chemical structure and stereochemistry.

We are grateful to SERC for financial support including a studentship (SJM), to SERC Daresbury Laboratory for experimental and computational facilities, and to Drs M. Z. Papiz, J. R. Helliwell, J. W. Campbell, I. J. Clifton, S. M. Clark, I. D. Glover and T. J. Greenough for advice or assistance in using these.

References

ADAMSON, G. W., DALY, J. J. & FORSTER, D. (1974). *Organomet. Chem.* **71**, C17–C19.

ANDREWS, S. J., PAPIZ, M. Z., MCMEEKING, R., BLAKE, A. J., LOWE, B. M., FRANKLIN, K. R., HELLIWELL, J. R. & HARDING, M. M. (1988). *Acta Cryst.* **B44**, 73–77.

CROMER, D. T. & LIBERMAN, D. (1970). *J. Chem. Phys.* **53**, 1891–1898.

JACOB, C. & HEATON, B. T. (1988). Personal communication.

MESSERSCHMIDT, A. & PFLUGRATH, J. W. (1987) *J. Appl. Cryst.* **20**, 306–315.

SHELDRICK, G. M. (1976). *SHELX*. Program for crystal structure determination. Univ. of Cambridge, England.

SMITH, J. H. & WONACOTT, A. J. (1979). *CCP4* program suite. SERC Daresbury Laboratory, England.

Acta Cryst. (1990). **B46**, 195–208

Hydration in Protein Crystals. A Neutron Diffraction Analysis of Carbonmonoxymyoglobin*

BY XIAODONG CHENG

Department of Physics, State University of New York at Stony Brook, Stony Brook, NY 11794, *USA, and Center for Structural Biology, Department of Biology, Brookhaven National Laboratory, Upton, NY* 11973, *USA*

AND BENNO P. SCHOENBORN

Center for Structural Biology, Department of Biology, Brookhaven National Laboratory, Upton, NY 11973, *USA*

(*Received* 25 *April* 1989; *accepted* 9 *November* 1989)

Abstract

In protein crystallography, it has been customary to ignore the contribution of bulk solvent by omitting the low-order diffraction data in refinement procedures. However, these data contain important information on both the structure of the solvent and the gross features of the unit-cell contents. The contribution of the solvent to the low-order structure-factor terms can be evaluated by dividing the solvent volume into shells extending outward from the surface of the protein. Two hydration layers in myoglobin crystals were characterized, allowing a better evaluation of the surface structure of the protein, improved placement of bound water and ion molecules, and a better overall fit to the observed data. A reciprocal-space least-squares refinement program was modified to include restraints on the configuration of water binding in water-to-protein and water-to-water associations. The inclusion of the solvent contribution allows all structure factors to be used in the refinement procedure. Eighty-seven water and five ion molecules were localized in carbonmonoxymyoglobin. All water molecules that are bound to protein bind to polar or charged groups, and the final R factor is 11·5%.

* Research carried out under contract DE-AC02-76CH00016 with the US Department of Energy.

0108-7681/90/020195-18$03.00

1. Introduction

In this paper we attempt to correlate the general information on hydration of proteins with crystallographic analysis of the solvent. On the basis of electron microscopic studies, Kellenberger (1978) suggested that water-soluble proteins are surrounded by a tightly bound layer of water molecules. Small-angle neutron scattering showed that the observed radius of gyration of a protein is larger than that predicted for the protein alone (Ibel & Stuhrmann, 1975; Schoenborn, 1989), suggesting that water is sufficiently closely associated with the protein to 'tumble' with it in solution. However, it has been difficult to demonstrate the existence of water layers in protein crystals. Crystallographic analyses of various proteins only 'see' water molecules that are well localized in three dimensions, and the agreement between water sites found in different experiments is generally poor (Takano, 1977*a*,*b*; Phillips, 1980, 1984; Hanson & Schoenborn, 1981; Kossiakoff, 1985). These bound water molecules represent only a small fraction of the total solvent (Savage, 1986); they are often arranged in clusters, and do not seem to form a layer surrounding the protein. The extent to which a presumed water layer is unexchangeable was investigated by Lehmann *et al.* using ethanol–water (Lehmann, Mason & McIntyre, 1985) and dimethyl sulfoxide–water (Lehmann & Stansfield, 1989) solutions; non-polar ethanol and dimethyl sulfoxide bind to hydrophobic sites in preference to water molecules. Numerous non-crystallographic techniques have been used to estimate the hydration of proteins in attempting to elucidate or improve the structure of the solvent. Analyses of water structure are discussed in recent reviews (Griffin, 1986; Kossiakoff, 1985; Edsall & McKenzie, 1983; Finney, 1979; Kuntz & Kauzmann, 1974; Cooke & Kuntz, 1974).

In crystallographic analysis of proteins, bound water molecules are often treated as though they 'belong' to the protein, but the rest of the bulk solvent is ignored. To minimize the effect of this omission, the low-order diffraction terms are not used in refinement procedures. However, an analysis of the low-order diffraction data can provide information on the general distribution of the solvent and protein within the crystal. To use the low-order crystallographic data in structural refinement, an evaluation procedure for the solvent was developed that uses its average scattering density (Schoenborn, 1988). As demonstrated in §2, this procedure can be extended to evaluate the scattering density of the solvent as a function of distance expanding outward from the surface of the protein. Analysis shows that the water layers surrounding the protein have little mobility, as demonstrated (§3) by a low liquidity factor. This procedure further allows the calculation of phases for the low-order diffraction data that enhances the localization of bound water, so that all data can be included in the refinement calculations in §4 and §5. In §6, a full description of the solvent structure of carbonmonoxymyoglobin (MBCO) is given.

2. Solvent shell and solvent structure factor

The measured structure factor (F_o) is described by the two crystal components of solvent ($\mathbf{F}_s$) and protein ($\mathbf{F}_p$) (Schoenborn, 1988):

$$F_o = |\mathbf{F}_s + \mathbf{F}_p|. \quad (1)$$

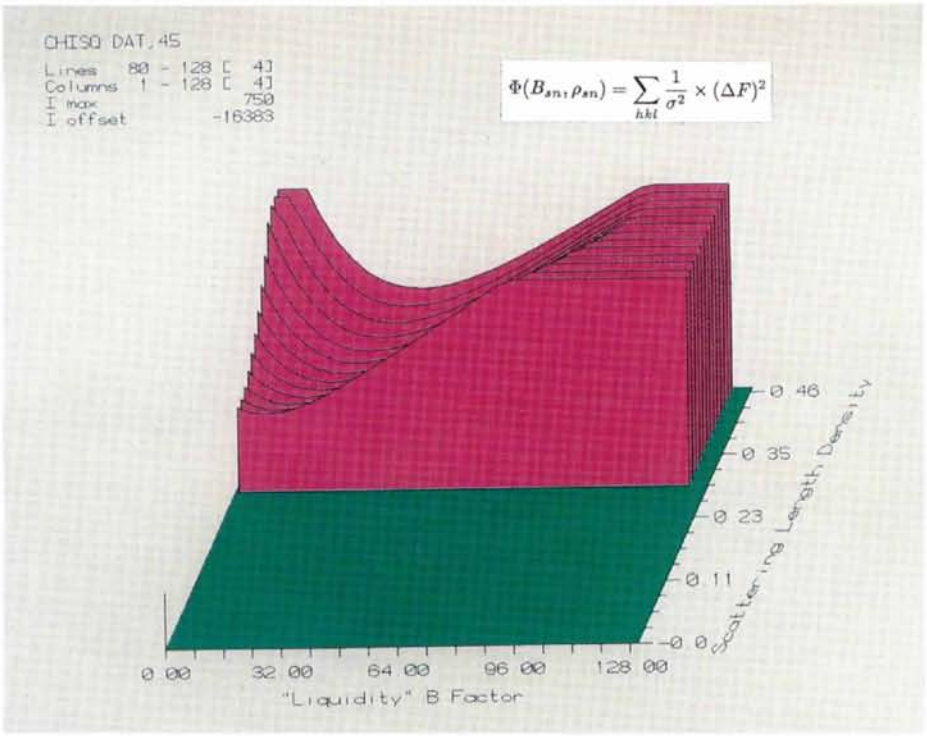
When the initial crystallographic analysis has reached the stage where a good description of the structure of the protein moiety has been obtained, the solvent component of the structure factors is given by the difference between the observed and calculated protein structure factors. Since the overall solvent has a relatively high disorder, here termed the liquidity factor, it contributes only to structure factors with low $\sin\theta$ values and only low-order diffraction data need to be considered. In this analysis, the solvent component ($\mathbf{F}_s$) is expanded into n structure-factor terms ($\mathbf{F}_{sn}$), corresponding to n shells extending outward from the protein's surface:

$$\mathbf{F}_s = \sum_n \mathbf{F}_{sn}. \quad (2)$$

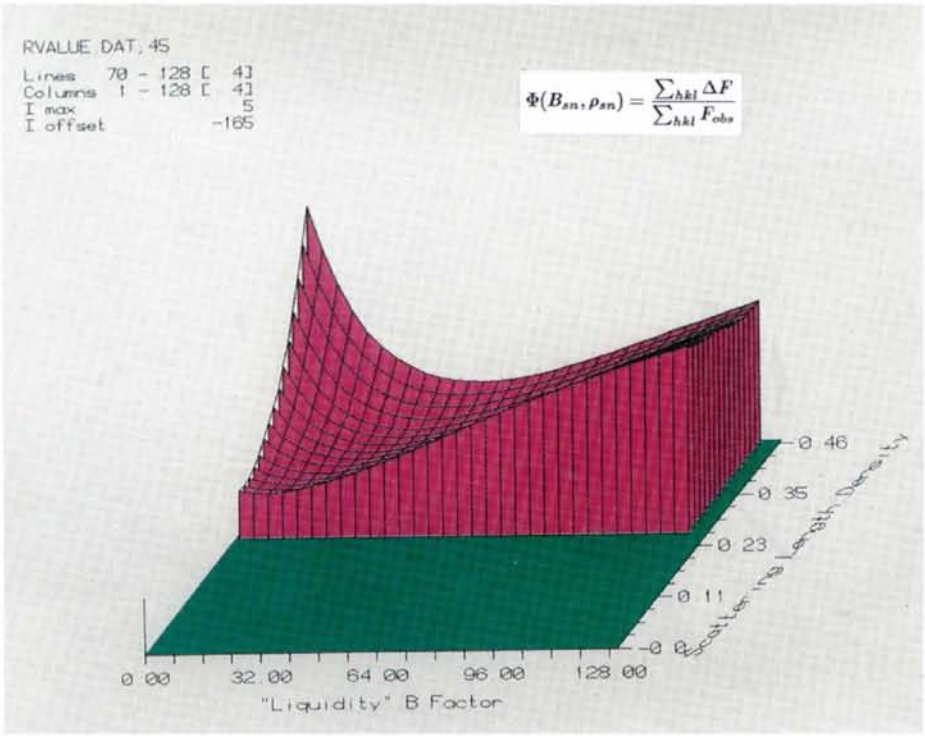
To achieve this the unit cell of the crystal is divided into a three-dimensional grid. A shell is made up of the grid points within a given range of distance from the surface of the protein: (*a*) grid points belong to the protein if the particular point falls within the van der Waals spheres of the atoms of protein; (*b*) grid points external to the protein belong to a given shell, depending on the distance from the surface of the protein. The grid points belonging to a given shell form the coordinate loci used for the calculation of structure factors for that shell. The solvent structure factors ($\mathbf{F}_{sn}$) for each shell and for each Miller index (hkl) are calculated from the given coordinates and the average scattering density of the solvent (ρ_{sn}) for each shell:

$$\mathbf{F}_{sn} = \rho_{sn}\exp(-B_{sn}\sin^2\theta/\lambda^2) \times \sum_{xyz}\exp[-2\pi i(hx + ky + lz)_n] \quad (3)$$

where xyz are the grid solvent coordinates for the nth shell. The scattering length of a solvent grid point is the volumetric scattering length (ρ_{sn}), in fm Å^{-3}, for the particular grid volume used, calculated from the scattering density of the solvent constituents (Table 1); all grid points within a shell have the same scattering densities.



(*a*)



(*b*)

Fig. 1. Plot of (*a*) the χ^2 and (*b*) the conventional R factor *versus* solvent-scattering density (fm Å^{-3}) and solvent liquidity (Å^2). The minimum R value is observed at the same point as the minimum χ^2.

[*To face p.* 196

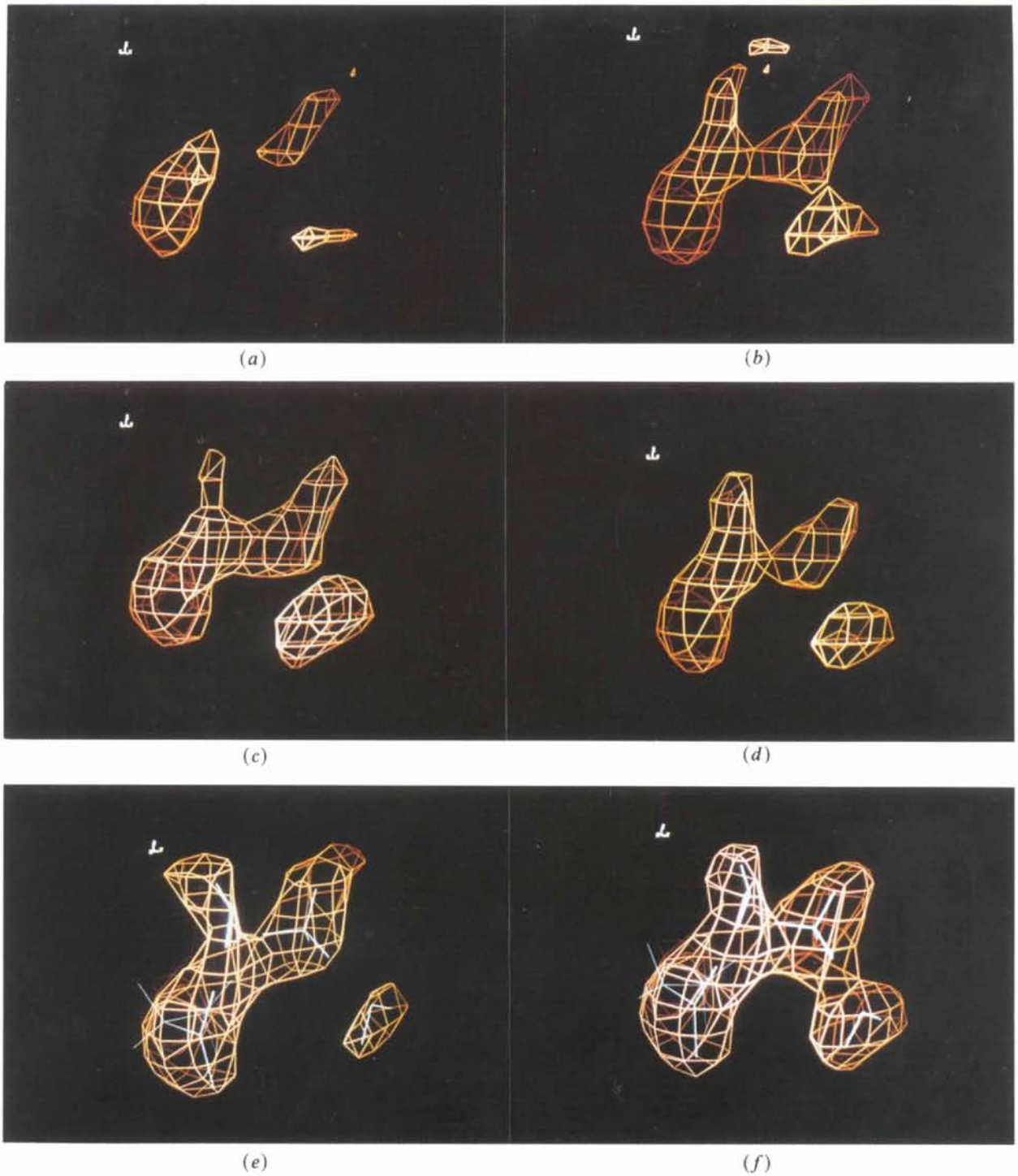


Fig. 4. Detail from the MBCO Fourier map of the surface of the protein, showing a feature on the lower left that corresponds to a COO^- group of the propionic acid attached to pyrrole ring *B* of the heme. (*a*) The Fourier map was calculated with $(2F_o - F_p)$ coefficients with protein phases (Φ_p) and corresponds to the classical case. (*b*) The total phases (Φ_t) [see equation (7.3)] including contribution from the whole uniform solvent were used with the same coefficients $(2F_o - F_p)$ as in (*a*). (*c*) The total phases (Φ_t) using the 15 solvent-shell model were employed with the same coefficients $(2F_o - F_p)$ as in (*a*). (*d*) The coefficients $(2F_o - F_t)$ [see equation (7.2)] were used with the same phases (Φ_t) as in (*c*). The contours depicted above [from (*a*) to (*d*)] were interpreted to contain two water molecules (D_2O) and an ammonium (ND_4^+), and these molecules were added to the protein structure. (*e*) The coefficients $(2F_o - F_p)$ with phases (Φ_p) were used after including the two water molecules and one ammonium ion in the calculation of the structure factors for $\mathbf{F}_p$. (*f*) The coefficients $(2F_o - F_t)$ with phases (Φ_t) for the revised solvent-shell model were used after adding the two water molecules and one ammonium ion to the protein structure.

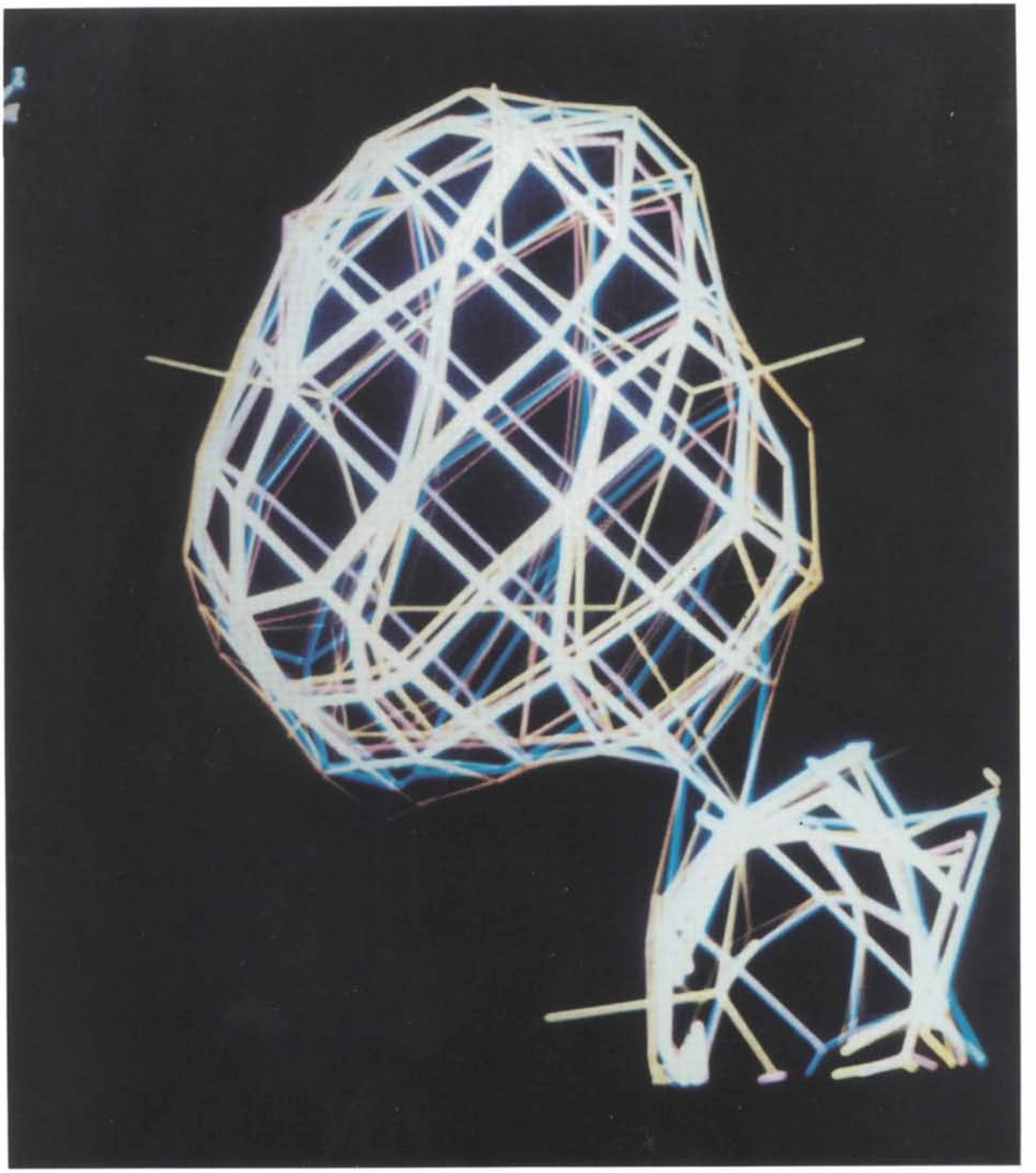


Fig. 5. Fourier sections of a histine (residue 36) in MCBO that are based on different solvent phases corresponding to Figs. 4(*a*) to 4(*f*), showing that the exterior solvent has little effect on the interior protein structure.

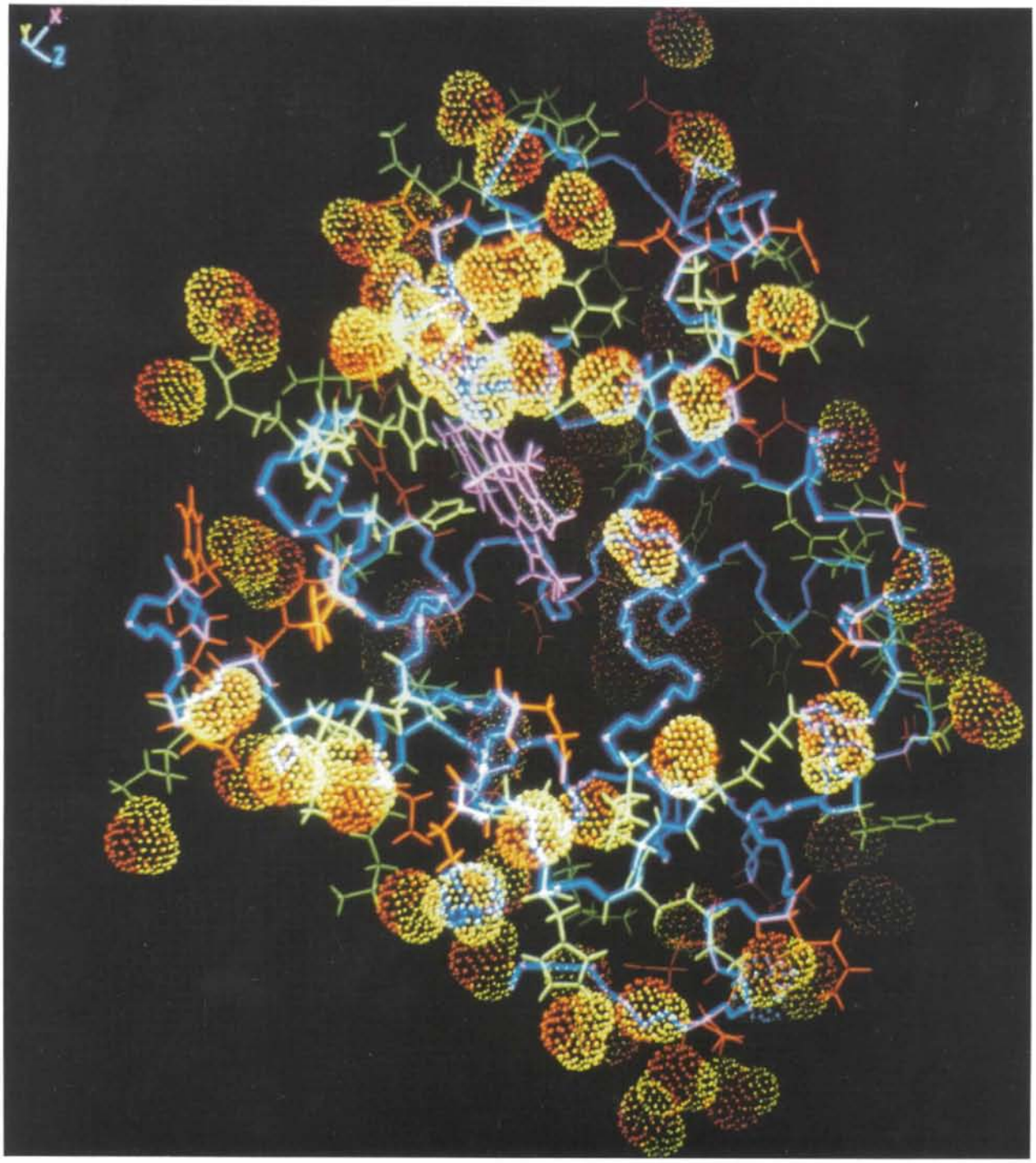


Fig. 7. A model of CO myoglobin showing the surface structure; 87 water and 5 ion sites are shown. Notice that the access path of CO (center of picture) is devoid of bound water. The heme group is in purple, the main chain in blue, acidic side chains in orange and basic side chains in green. The neutral and nonpolar residues are not shown; the purple dots in the main chain are the α carbons. Water and ion molecules are indicated by space-filling dotted clusters.

Table 1. *Average neutron-scattering densities of water, ammonium and sulfate ions*

1 fm = 10^{-5} Å = 10^{-13} cm.

	Scattering density (fm Å^{-3})
H_2O	−0·056
D_2O	0·635
NH_4^+	−0·122
ND_4^+	0·814
SO^{2-}	0·257

Each shell has a liquidity factor (B_{sn}) that describes its mobility (disorder). This liquidity factor is based on the root-mean-square displacement around a solvent grid point and is thus equivalent to the temperature factor. Each shell therefore has two variable parameters, the solvent-scattering density (ρ_{sn}) and the liquidity factor (B_{sn}). It now remains to find the best values for ρ_{sn} and B_{sn} for n shells that minimize ΔF:

$$\Delta F(\rho_{sn}, B_{sn}) = |(\mathbf{F}_o - \mathbf{F}_p) - \mathbf{F}_s(\rho_{sn}, B_{sn})|. \quad (4)$$

$\mathbf{F}_p$ is calculated from the atomic scattering lengths and the coordinates of the protein atoms. $\mathbf{F}_s$ is calculated, as described above, for low-angle data until 3 Å resolution with initial liquidity factor B = 1. The best values for ρ_{sn} and B_{sn} are found by an iterative numerical procedure that changes these values in small increments (a 128 × 128 two-dimensional search) and tests for the lowest ΔF. The behavior of this minimization is measured by a weighting function (Schoenborn, 1988):

$$\chi^2 = \sum(1/\sigma^2)(\Delta F)^2 \quad (5)$$

where σ is the counting statistics of the observed reflections. The conventional R value can be used with similar purpose:

$$R = \sum|\Delta F|/\sum|F_o|. \quad (6)$$

The functions are plotted as a three-dimensional plot (Fig. 1). For the first step, the best overall solvent scattering density ρ_s and the liquidity factor B_s were determined. These overall uniform values for the whole solvent area were used as initial values for all shells. Keeping all except shell l at their initial values of ρ_s and B_s, the values of ρ_{sl} and B_{sl} which gave the minimum χ^2 or R values were determined. With ρ_{sl} and B_{sl} kept for shell l, and with all shells but l and m kept at ρ_s and B_s, the best ρ_{sm} and B_{sm} were determined in the same manner. Continuing in this way, the best values of ρ_{sn} and B_{sn} for each succeeding shell were determined. This optimization is well conditioned, since the number of structure factors that contribute to this function is relatively large (~1000) compared with the few shells (~20).

Table 2. *Average solvent-scattering densities of myoglobin crystals in different solvents*

The crystal solvent D_2O concentrations listed are the initial D_2O/H_2O percentages used for setting up crystallization. The value is an upper limit; the actual D_2O concentration will be less due to exchange. The observed scattering density depends on the actual D_2O concentration and the amount of ammonium sulfate in solution (Schoenborn, 1988).

	Scattering density (fm Å^{-3})	Crystal solvent
MB25	0·23 ± 0·02	~40%D_2O
MB17	0·49 ± 0·02	~80%D_2O
MBCO	0·56 ± 0·02	~90%D_2O

3. Hydration layers in myoglobin crystals

This procedure is particularly powerful for neutron crystallographic analysis of proteins because the average neutron-scattering density of the solvent can be large and easily adjusted by mixing heavy and light water to obtain a given scattering density (Table 1). Neutron diffraction data for three myoglobin crystals (Table 2) with different solvent-scattering densities were collected and used to analyse the solvent liquidity and scattering density as described above. The crystals were grown from saturated ammonium sulfate solutions with MB25 in ~40% D_2O, MB17 in ~80% D_2O, and MBCO in ~90% D_2O (MB25 and MB17 are aquamet-myoglobin crystals). The early crystallographic analysis provided the basic atomic positions of the protein (Hanson & Schoenborn, 1981; Schoenborn, 1971). The coordinates of the solvent grid were determined, using the known atomic positions of the protein. An asymmetric unit of the cell contains a total of 120 000 points at ~0·6 Å grid spacing with 44 978 points in the solvent region; in this case, a hydrogen radius of 1·6 Å was used to determine the surface of the protein, which shows that nearly 38% of the cell volume is solvent. Solvent structure factors ($\mathbf{F}_{sn}$) for fifteen shells with a grid size of ~0·6 Å were calculated to a resolution of 3·0 Å from the coordinates of all the solvent grid points, with initial scattering densities corresponding to the composition of the solvent that was used in crystallization (Table 2).

A minimization calculation of χ^2 and the R factor was carried out by changing the magnitude of the solvent-shell scattering densities and liquidity factors in small steps as discussed above. The best observed scattering densities and liquidity factors of the shells for each of the three data sets were thus determined in this manner. Fig. 2(*a*) shows the results with the best ρ_{sn} and B_{sn} as a function of distance from the surface of the protein. These data show that there is a layer of solvent molecules with low mobility, as given by the low liquidity factors (~25 Å^2), that is located at ~2·3 Å from the center of the surface

atoms (hydrogen/deuterium). The scattering densities are highest at this distance, suggesting that the bulk mass density of the innermost layer is higher than those of solvent layers further out from the protein. The uniform scattering density for each shell, however, includes the scattering density of any ion or fraction of an ion located within that shell.

To test the effect of a finite grid size, calculations were made with a grid size down to ~0·3 Å. The computations related to the structure factors require about 12 h of CPU time on a VAX 11/780 of the 378 205 solvent grid points from the total of 960 000 points. All three data sets yield essentially the same scattering data for the solvent region and are

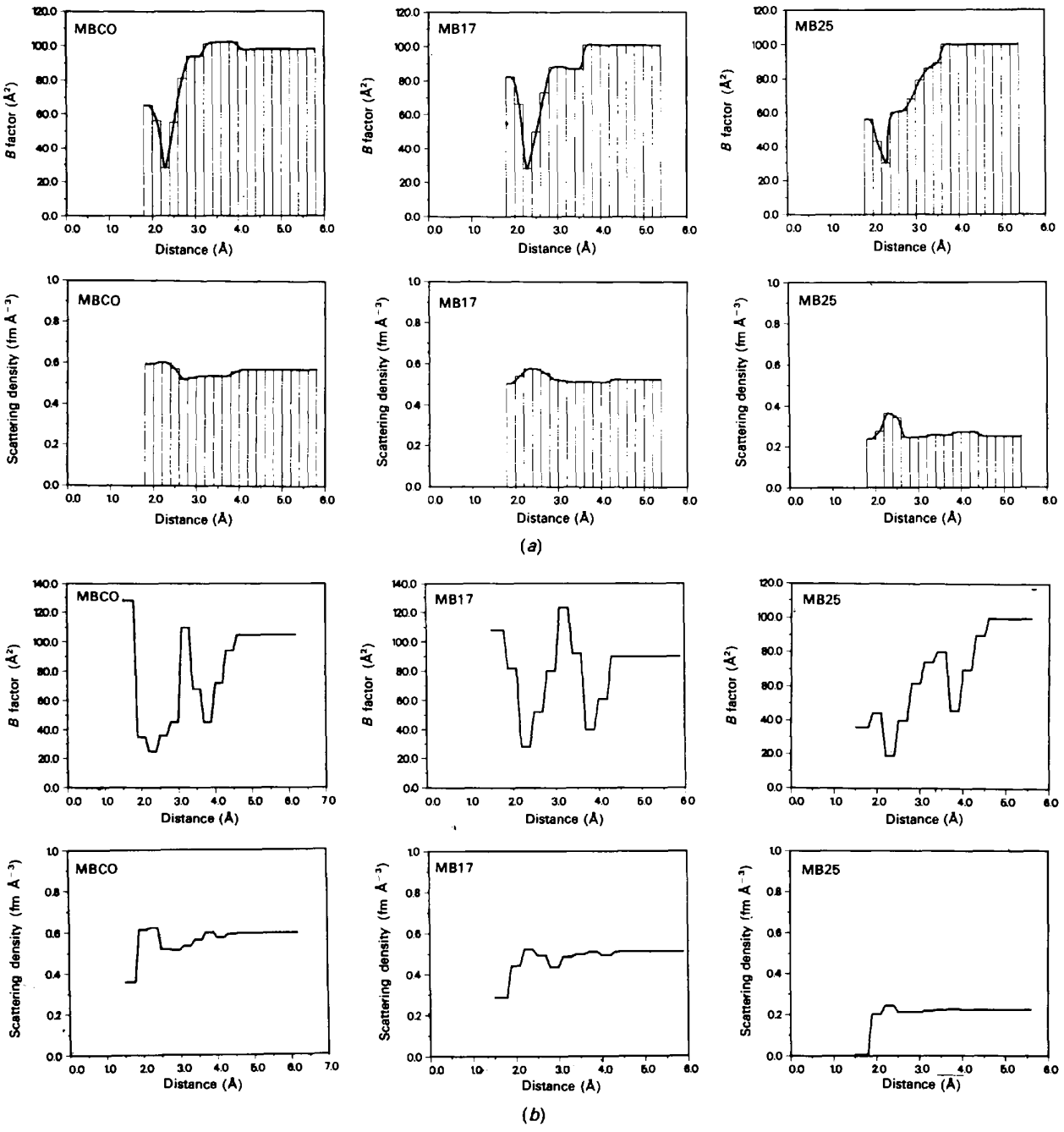


Fig. 2. Graphic presentation of the best scattering densities and *B* factors for three different myoglobin derivatives with different solvent compositions (see Table 2). (*a*) The initial radius for hydrogens is 1·6 Å and the shells extend outward from that radius with thickness 0·2 Å and a grid size of 0·6 Å. (*b*) The results based on a shell thickness of 0·3 Å and a grid size of 0·3 Å.

Table 3. *Shell-solvent analysis on MBCO*

The first row of data (total) are the best values of the overall solvent used to initiate the shell analysis. The content of solvent is about 40%. The shell thickness and the grid size are both 0·3 Å; the volume of one grid point is 0·035 Å^3. The shells (11–15) of the last row of data, at 4·6 to 6·1 Å from the surface of the protein, are treated as one large shell. The radii used in calculations to determine the surface of the protein are C: 1·7, N: 1·5, O: 1·4, S: 1·85, Fe: 1·45, H: 1·6, D: 1·6 Å.

	Distance (Å)	Grid points (number)	Best ρ_{sn} (fm Å^{-3})	Best B_{sn} (Å^2)
Total	1·6–6·1	378205	0·58	86
Shell 1	1·6–1·9	117962	0·36	128
2	1·9–2·2	72212	0·61	35
3	2·2–2·5	52679	0·62	25
4	2·5–2·8	40289	0·52	36
5	2·8–3·1	30754	0·51	45
6	3·1–3·4	23417	0·53	110
7	3·4–3·7	15915	0·56	68
8	3·7–4·0	11468	0·59	45
9	4·0–4·3	7017	0·57	72
10	4·3–4·6	3941	0·58	94
11–15	4·6–6·1	2551	0·59	105

independent of the grid size from 0·6 to 0·3 Å. Fig. 2(*b*) shows the results based on a grid size of ~0·3 Å and a shell thickness of 0·3 Å. Table 3 gives the detailed characteristics of the shell analysis on MBCO. The best overall values for the total solvent are listed in the first line of Table 3. The best average density in the overall solvent region was 0·58 fm Å^{-3}, which agrees well with the D_2O/H_2O ratio of ~90% used for setting up crystallization. The innermost layer of the shell is very well defined and is separated from the second one. A second layer of water molecules is indicated by the second peak at 3·7 to 4·0 Å with liquidity factors of ~45 Å^2 (from Fig. 2*b*).

A possible picture of solvent structure which emerges from this analysis is shown schematically in Fig. 3. There is a well-defined layer of water molecules at the expected van der Waals distance from the protein surface (D/H atoms). Each water molecule in the first layer is hydrogen bonded to water molecules in the second layer. The ideal O—H···O bond length in ice (Pauling, 1967) is 2·76 Å and O—H···O angle 180°. The observed ~3·9 Å distance of the second layer from the protein surface agrees with the predicted value of 4·03 Å, Fig. 3. The scattering densities of the water layers are about the same, but the liquidity factors for the second water layer are higher suggesting that there is less order.

4. Phasing, solvent structure determination and refinement

With the best parameters for the solvent shells known, modified phases for the observed structure factors can be calculated to improve the Fourier map features on the surface of the protein that depict bound water and ion molecules. The phases for F_o's are obtained from the calculated structure factors ($\mathbf{F}_t$) for protein and solvent shells using the expanded structure-factor form:

$$\mathbf{F}_t = \mathbf{F}_p + \mathbf{F}_s(\rho_{sn}^{\text{best}}, B_{sn}^{\text{best}}) \tag{7.1}$$

and,

$$|\mathbf{F}_t| = [(A_p + A_s)^2 + (B_p + B_s)^2]^{1/2} \tag{7.2}$$

where A_p, A_s and B_p, B_s are the real and imaginary terms of the protein and solvent components respectively.

The phase for F_o is given by:

$$\Phi_t = \text{tg}^{-1}[(B_p + B_s)/(A_p + A_s)]. \tag{7.3}$$

The magnitude and phase of the structure factor of the protein and solvent components are:

$$|\mathbf{F}_p| = [(A_p{}^2 + (B_p)^2]^{1/2} \tag{8.1}$$

$$\Phi_p = \text{tg}^{-1}(B_p/A_p) \tag{8.2}$$

and,

$$|\mathbf{F}_s| = [(A_s)^2 + (B_s)^2]^{1/2} \tag{8.3}$$

$$\Phi_s = \text{tg}^{-1}(B_s/A_s). \tag{8.4}$$

The Fourier maps calculated use the observed structure factors (F_o) and the calculated phases (Φ_t) based on all the known features of the particular protein. These Fourier maps are interpreted with a macromolecular graphics modeling program (*PS300 FRODO*). To show the effect of solvent phasing, the Fourier map calculated with coefficients $2|F_o| - |F_p|$ based on phases (Φ_p) from the parent protein structure only, was compared with maps based on the

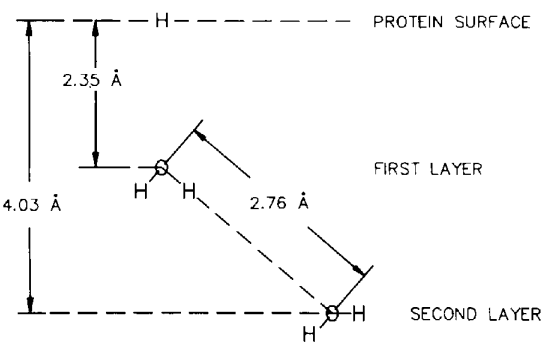


Fig. 3. Water molecules in the first and second hydration layers at the surface of the protein. 2·35 Å is the sum of van der Waals radii for the H atom (protein) and the O atom (water). 2·76 Å and the linear bond angle are the hydrogen-bond distance and angle in ice. The bisector of the H—O—H bond angle of 105° is taken perpendicular to the protein surface. (Not shown in this sketch are salt ions which can also be bound at the surface of the protein.)

total phases including the solvent structure contribution (Φ_t). The neutron density map ($2|F_o| - |F_p|$, Φ_p) corresponds to the classical case (no water molecules are included in the calculated structure factors), and the map ($2|F_o| - |F_p|$, Φ_t) corresponds to the solvent-phased density. The small Fourier section in Fig. 4 (see §6) shows the effects of including solvent phasing on the information obtainable, and demonstrates how the improved phasing model greatly facilitates the localization and orientation of water and ion molecules. The solvent-phased Fourier map shows features that are completely absent in the conventional map and, generally, yields a greatly improved topography of the surface of the protein.

These new calculated structure factors ($\mathbf{F}_s$) that represent the solvent-shell model can also be used to modify the observed structure factors (F_o) so that they exclude the contribution of the solvent and represent only protein. These modified terms (F_{om}) are then used to refine the protein structure in a modern reciprocal-space least-squares refinement (*PROLSQ*) that includes all structure factors:

$$F_{om} = |\mathbf{F}_o - \mathbf{F}_s(\rho_{sn}^{\text{best}}, B_{sn}^{\text{best}})|. \quad (9)$$

The inclusion of the low-order data in the refinement is advantageous, since these data are generally the strongest with the best counting statistics and they condition the least-squares refinement well. $|F_o|$ is scaled to match $|\mathbf{F}_p|$ for the high-angle data (< 3 Å), the same scale factor is used for low-angle data (> 3 Å) to scale $|F_{om}|$ to match $|\mathbf{F}_p|$.

Through an inspection of the Fourier map, bound-water and ion molecules were localized and added to the protein coordinates. The D and O atoms of water were placed at stereochemically reasonable positions as indicated by the Fourier features. Another set of modified F_{om} were calculated with all of the newly located solvent molecules treated as a part of the protein. The refinement of water molecules poses some new problems. Water molecules do not always have full occupancy and, in addition, hydrogen atoms are exchangeable. In our case, the occupancy factor (Q) for all three atoms in a water molecule is determined by the oxygen atom only, while the scattering length of an exchangeable hydrogen is given by the relation $b = (6{\cdot}67 + 3{\cdot}74)f - 3{\cdot}74$, where 6·67 and −3·74 are the scattering lengths, in fm, of deuterium and hydrogen, respectively, and f is the fractional hydrogen-exchange rate. The permissible range of f is from 0 (hydrogen) to 1 (deuterium). The refinement is initiated with zero scattering ($b = 0{\cdot}0$ when $f = 0{\cdot}355$). On completion of several cycles of refinement of the protein and water structure, the whole solvent-analysis procedure is repeated, followed by a Fourier synthesis and calculation of another set of modified F_{om} which are then used in the next round of refinement.

5. Water restraints in the least-squares refinement

The restrained least-squares refinement (Hendrickson & Konnert, 1981; Hendrickson, 1985) was modified to allow the refinement of the water structure by including restraints on the binding configurations in water-to-protein and water-to-water associations. The *PROLSQ* procedure of Hendrickson & Konnert (1981) was expanded to include a set of water-repulsive terms. These restrictions on water structure are based on results of Savage & Finney (1986) that describe stereochemical restraints in ices and crystalline hydrates. These interactions are characterized by interatomic potential-energy functions $U(d)$ which can be approximated (Hendrickson, 1985) by:

$$U(d) - U(d_{\min})|_{d < d_{\min}} \simeq (1/\sigma_V^4)(d - d_{\min})^4 \quad (10)$$

where $d_{\min}$ locates the minimum energy, only repulsive contacts ($d < d_{\min}$) are considered, and σ_V is the standard deviation to be permitted. σ_V varies in principle with the type of contact, but in practice the values are quite uniform. The value of $d_{\min}$ depends on which atomic elements are in contact and what type of contact they form:

(*a*) O···O repulsion of hydrogen-bonded structures (O—H···O)

$$d_{\min}|_{\alpha \geq 120} \simeq -(\alpha/120) + 4{\cdot}0$$

where α = O—H···O angle (°), and $\sigma_V = 0{\cdot}65$ for water-to-protein and 0·80 for water-to-water contacts.

(*b*) H···O repulsion of hydrogen bonds (O—H···O)

$$d_{\min}|_{\alpha \geq 120} \simeq -(\alpha/70{\cdot}0) + 4{\cdot}0$$

where α = O—H···O angle (°) and $\sigma_V = 0{\cdot}65$.

(*c*) H···H′ non-bonded repulsion (O—H···H′—O′)

$$d_{\min}|_{180 \leq \chi \leq 250} \simeq -(\chi/220) + 3{\cdot}1$$

where $\chi = \alpha + \beta$, α = O—H···H′ angle (°), β = H···H—O′ angle (°) and $\sigma_V = 0{\cdot}65$.

(*d*) $H_2 \cdots O_1$ oxygen–remote-neighbor hydrogen non-bonded interaction ($H_2—O_2 \cdots H_1—O_1$)

$$d_{\min}|_{80 \leq \varphi \leq 113} \simeq -(3\varphi/200) + 4{\cdot}2$$

where $\varphi = H_2—O_2 \cdots H_1$ angle (°) and $\sigma_V = 0{\cdot}65$.

Beyond the range of angles ($\alpha \geq 120$, $180 \leq \chi \leq 250$, $80 \leq \varphi \leq 113°$), $d_{\min}$ will take the value of the van der Waals contact distance. Restraints have been applied to possible bonds formed by water characterized by distance limits.

6. Application to carbonmonoxymyoglobin crystals

The structure of MBCO was originally determined from neutron crystallographic data (Norvell, Nunes & Schoenborn, 1975) using the Diamond real-space refinement (Hanson & Schoenborn, 1981). The CO myoglobin crystal was grown from a 70% saturated

Table 4. *Summary of the refinement*

Every run involved several cycles. Each cycle refines either atomic coordinates *xyz*, or together with the refinement of one of the three factors, *B* (individual isotropic thermal parameter), *f* (hydrogen-exchange rate of H/D atom) and *Q* (occupancy factor of solvent molecule). Each cycle reads parameter shifts from the previous cycle and uses a conjugate-gradients procedure to solve the new parameter shifts. Finally, it tests the new shifts on the *R* value, searching for an optimal shift-damping factor against a sample of data. In our case, we simply alternate the application of *xyz*, *B*, *f* and *Q* shifts in successive cycles. The total number of cycles of each run is shown in parentheses in column 1; the number of cycles on each parameter is shown in columns 3 to 6. Only occupancy factors of solvent molecules are refined.

Run No.	*R* (%)	*xyz*	*B*	*f*	*Q*	Atoms	Comments
1 (5 cycles)	29·3→26·7	2	1	2	—	2543	MBCO structure from real-space refinement (Hanson & Schoenborn, 1981), original observed structure factor F_o used.
2 (8)	25·4→19·3	4	1	3	—	2543	Overall solvent contribution included, a modified observed structure factor F_{om} used.
3 (16)	19·0→16·3	5	4	4	3	2780	Solvent-shell analysis performed, 79 water molecules added, a new F_{om} used.
4 (16)	17·6→14·8	5	3	3	5	2941	Solvent-shell analysis performed, a total of 127 water and 3 ammonium ions included, a new F_{om} used.
5 (28)	15·2→12·7	7	7	7	7	2821	Solvent-shell analysis performed, a total of 86 water, 3 ammonium and one sulfate ions included, a new F_{om} used.
6 (8)	12·9→11·5	4	4	—	—	2829	With a relaxed weighting of the stereochemical conditions.

Table 5. *Reflections to a resolution of* 12 Å *of MBCO*

The definitions of each column are given in formulas (7) to (9).

				Protein		Solvent		Total		
H	*K*	*L*	F_o	F_p	Φ_p	F_s	Φ_s	F_t	Φ_t	F_{om}
0	2	0	332.	972.	340.	758.	177.	331.	298.	973.
0	0	1	661.	1397.	14.	902.	207.	560.	352.	1492.
0	1	1	690.	857.	102.	941.	251.	484.	186.	901.
0	0	2	562.	881.	0.	1395.	180.	514.	180.	833.
1	2	−1	295.	352.	56.	237.	300.	327.	15.	328.
1	2	0	113.	267.	311.	176.	136.	94.	301.	286.
1	2	1	291.	504.	77.	233.	280.	304.	60.	492.
1	1	−2	392.	323.	100.	176.	9.	367.	71.	346.
1	1	−1	695.	639.	48.	1352.	225.	715.	222.	659.
1	1	0	1906.	76.	7.	1865.	141.	1814.	139.	70.
1	1	2	270.	396.	41.	379.	250.	193.	330.	428.
1	0	−2	478.	408.	0.	101.	180.	307.	360.	579.
1	0	−1	1270.	542.	0.	1781.	180.	1239.	180.	511.
1	0	2	346.	27.	0.	264.	180.	237.	180.	82.
2	2	0	228.	63.	91.	312.	266.	249.	265.	84.
2	2	1	588.	868.	204.	323.	17.	550.	208.	906.
2	1	−2	273.	470.	75.	230.	274.	264.	58.	479.
2	1	−1	196.	954.	187.	1317.	17.	412.	40.	1141.
2	1	0	369.	164.	269.	321.	335.	416.	314.	135.
2	1	1	448.	894.	317.	617.	166.	468.	277.	879.
2	1	2	426.	740.	34.	492.	204.	268.	52.	890.
2	0	−1	993.	248.	0.	1196.	180.	948.	180.	203.
2	0	0	1305.	330.	180.	1575.	0.	1246.	360.	270.
2	0	1	794.	97.	0.	747.	180.	650.	180.	47.
2	0	2	582.	421.	180.	168.	180.	589.	180.	414.
3	2	−1	142.	438.	349.	287.	148.	199.	21.	390.
3	2	0	484.	713.	242.	260.	74.	462.	235.	735.
3	1	−2	324.	42.	100.	346.	89.	387.	91.	22.
3	1	−1	150.	707.	79.	795.	256.	96.	234.	658.
3	1	0	431.	691.	313.	510.	146.	227.	282.	871.
3	1	1	393.	508.	128.	197.	320.	318.	120.	582.
3	0	−2	212.	846.	180.	724.	360.	122.	180.	937.
3	0	−1	209.	654.	0.	900.	180.	246.	180.	691.
3	0	0	147.	146.	180.	294.	360.	148.	360.	147.
3	0	1	283.	644.	180.	259.	360.	385.	180.	541.
4	1	−2	176.	343.	282.	261.	111.	95.	256.	417.
4	1	−1	473.	706.	244.	273.	39.	474.	258.	706.
4	1	0	347.	148.	325.	351.	19.	454.	3.	92.
4	0	−1	229.	1029.	0.	1329.	180.	301.	180.	1101.
4	0	1	545.	709.	0.	86.	180.	623.	360.	631.
5	0	−1	269.	274.	180.	50.	360.	224.	180.	320.
5	0	0	417.	572.	180.	155.	360.	418.	180.	572.

ammonium sulfate solution at pH 5·7, and was soaked in D_2O mother-liquor solution for several months. We re-examined this structure using the solvent-modified structure-factor terms in the expanded least-squares refinement (*PROLSQ*), as described above. The starting coordinates of the protein resulted in the following *PROLSQ* parameters: 2543 atoms, 8397 distances, 213 planes, 182 chiral centers, 763 torsion angles, 47 018 possible contacts, and 3156 non-bonded contacts in the protein itself. The total number of variables is 10 173 and there are 6520 structure-factor observations ($>2\sigma$ for ∞ to 2 Å and $>7\sigma$ for 2 to 1·8 Å). Five cycles of refinements (two cycles on *xyz*, one cycle on individual isotropic thermal *B* factors, and two cycles on exchange ratios of H/D atoms) with the original observed F_o reduced the *R* from 29·3 to 26·7% (Table 4). This produced an r.m.s. shift in atomic positions for all atoms of 0·36 Å. Comparison of the coordinates with the starting set shows that most of the large shifts were in surface side-chain atoms.

The overall solvent contribution to the structure factors was then calculated using a ~0·6 Å grid spacing and data to 3·0 Å resolution (2245 reflections). The protein structure was further refined after the overall solvent contribution was included in the observed structure factors; *i.e.*, modified observed structure factors were calculated using the best scattering densities and liquidity factors of the total solvent, and then used in the refinement. The *R* value improved rapidly to 19·3% after eight cycles of refinement (run No. 2). A solvent-shell analysis was then performed. Solvent structure factors ($|\mathbf{F}_s|$, Φ_s) were calculated, based on the results listed in Table 3 and equations (7)–(9). A comparison of F_o, $\mathbf{F}_p$, $\mathbf{F}_s$, $\mathbf{F}_t$ and F_{om} for low-resolution data (>12 Å) is given in Table 5, showing the effect of the solvent's contribution. For the low-order data, the contributions from the protein and the solvent are of the same order of magnitude and show large phase shifts. F_t, which includes contributions from both the protein and solvent parts, is now directly comparable with the observed total structure factor (F_o).

Based upon the protein coordinates and best solvent-shell scattering densities and liquidity factors from the shell analysis, a new neutron density map ($2|F_o| - |F_p|$, Φ_t) was calculated and analyzed on an interactive graphics system, *FRODO PS*300. At this stage, the 79 largest features with densities greater

than 0·56 fm Å^{-3} on the surface of the protein were fitted with water molecules, localizing the atoms to the neutron density while maintaining ideal molecular geometry and suitable hydrogen-bond donors and acceptors. These individual water molecules were then added to the protein coordinates and refined, applying the previously described water restraints. The coordinates were refined against the modified observed neutron structure factors (F_{om}). The occupancies of the oxygen atoms of water and hydrogen exchange rates of the H/D atoms were refined alternately with coordinates, and after 16 cycles an R factor of 16·3% was obtained. The process of the refinement is shown in Table 4.

After run No. 3, another solvent-shell analysis was performed and a new neutron density map was calculated. The water sites were carefully analyzed and two water sites with high densities (>0·70 fm Å^{-3}) were changed to ammonium ion sites. Inspection of the remaining Fourier features localized another ammonium ion and 50 more water sites, to a total of 127 water molecules. All three ammonium sites are adjacent to negative charges. Fig. 4 shows the Fourier representation of a COO^- group of one of the two propionic acids of the heme showing the effects of different phasing used to calculate Fourier maps. The contours depicted in Fig. 4(*b*) were interpreted as containing two D_2O and one ND_4^+. The inclusion of solvent in this analysis led to significant changes in the surface configuration of the protein, particularly in the placement of bound water and ions, without changing the interior features of the protein. An example of this is shown in Fig. 5 which depicts superimposed Fourier sections of a histidine residue C1 (36) that are based on phases calculated with different solvent phasing models corresponding to Figs. 4(*a*) to 4(*f*).

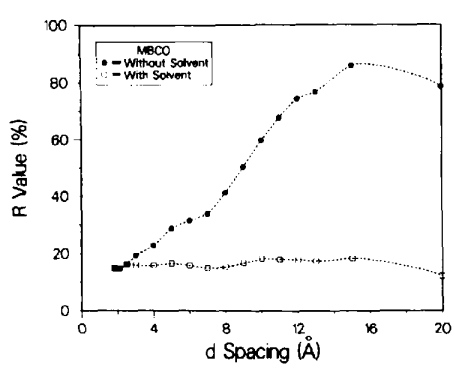


Fig. 6. A graph of the conventional R value as a function of d spacing ($n\lambda = 2d\sin\theta$). The top curve represents the R value without the solvent correction (the classical case), while the bottom curve shows the R value with the correction from the solvent-shell model.

Table 6. *Resolution breakdown*

An average of the value of $\sigma_F = \frac{1}{2}\langle F_o - F_c\rangle$ was used as weighting sigmas instead of individual counting statistics for the observed structure factors. Columns 3 to 5 are the refinement at the R = 12·7% stage, and columns 6 to 8 are the refinement at the present stage after eight cycles with relaxed weighting.

Resolution (Å)	No. of reflections	$\langle F_o - F_c\rangle$	Shell R (%)	Sphere R (%)	$\langle F_o - F_c\rangle$	Shell R (%)	Sphere R (%)
6·00–∞	323	43·5	9·7	9·7	40·4	9·0	9·0
5·00–6·00	221	37·6	12·0	10·5	35·4	11·3	9·4
4·00–5·00	485	35·6	10·1	10·3	33·1	9·4	9·1
3·00–4·00	1216	32·0	10·7	10·5	29·9	10·0	9·3
2·50–3·00	1414	29·0	13·6	11·4	27·5	12·9	10·2
2·10–2·50	1845	27·3	15·3	12·3	25·9	14·5	11·1
1·80–2·10	1016	25·9	16·5	12·7	23·7	15·1	11·5
1·80–∞	6520	30·1	–	12·7	28·2	—	11·5

Five more water molecules were added after run No. 4, while 45 water sites were rejected due to low occupancy (<0·3) and one Fourier feature close to the side chain of Lys H21 (145) was interpreted as a sulfate ion. The proof that this site is indeed an SO_4^{2-} ion must wait until the analysis of the equivalent non-deuterated (H_2O) crystal is complete. In the H_2O case, an SO_4^{2-} molecule would have the same density while a water molecule with the same occupancy would have a much lower density (−0·056 fm Å^{-3}, Table 1). Finally, one more water molecule and another ammonium ion joined the family of solvent molecules after run No. 5.

Fig. 6 shows the effect of including the low-order data on the conventional R value of a structural refinement. The inclusion of the solvent model greatly improved the overall R value to 12·7% (Table 6) for all reflections. Atoms, including hydrogen and oxygen atoms of solvent molecules, in this model have an r.m.s. bond distance deviation from 'ideality' (Hendrickson, 1985) of 0·006 Å for non-hydrogen bond lengths (1–2 neighbor, notation as defined by Hendrickson), 0·022 Å for hydrogen-related bond distances, 0·039 Å for angle-related distances (1–3 neighbor), and 0·027 Å for intraplanar distances (1–4 neighbor). For repulsive water contacts (notation as defined in §5) this left an r.m.s. deviation from 'ideal' distances of 0·50 Å for water-to-protein O···O distances, 0·80 Å for water-to-water O···O distances, 0·50 Å for H···O distances and 0·44 Å for H···H′ distances. Eight more cycles with a relaxed weighting of the stereochemical conditions brought R to 11·5% (Tables 4 and 6). The weighting factor related to bonding distances, for example, is determined by σ_D (Hendrickson, 1985), the standard deviation for the distances of a particular type, which was relaxed from 0·01 to 0·03 Å for non-hydrogen 1–2 bonds, from 0·03 to 0·05 Å for hydrogen-related 1–2 bonds, from 0·04 to 0·06 Å for 1–3 and 1–4 neighbors, *etc.* Atoms shifted another 0·22 Å (r.m.s.), and no significant shifts were seen in the final two cycles.

Table 7. *Water-binding configuration: water-to-protein and water-to-water*

The hydrogen bonds (O or N···D—O) between protein-to-water and water-to-water are listed. The number after the first slash is the non-deuterium distance (O or N···O) (Å); the number after the second slash is the angle at the deuterium atom (°). The number in parentheses is the residue number of water or ion; the symbol # represents the neighboring symmetry-related molecule.

Residue number and type	Main chain N	Main chain C	Side chain
1 Val *NA1*	N—D···O(254)/2·80/165 O(254)···D—O(255)/2·63/151 O(255)—D···O(221)/3·21/141···(133 Lys)		
2 Leu *NA2*	D	O	
3 Ser *A1*	D	O	
4 Glu *A2*	N—D···O(253)/2·60/150 O(253)···O(#174 SO_4^{2-})/1·89···(#145 Lys)		$O_{\varepsilon 1}$···D—O(286)/2·47/160
5 Gly *A3*	D	O···O(225)/2·24 O(225)—D···O(226)/2·97/162	
6 Glu *A4*	D	O	$O_{\varepsilon 1}^{-}$···D—O(289)/3·03/124 $O_{\varepsilon 2}^{-}$···D—O(289)/3·59/144
7 Trp *A5*	D	O	
8 Gln *A6*	D	O	
9 Leu *A7*	D	O	
10 Val *A8*	D	O	
11 Leu *A9*	D	O	
12 His *A10*	D	O	
13 Val *A11*	D	O	
14 Trp *A12*	D	O	
15 Ala *A13*	D	O	
16 Lys *A14*	D	O···D—O(192)/2·72/170···(#59 Glu)	
17 Val *A15*	D	O	
18 Glu *A16*	D	O	$O_{\varepsilon 2}^{-}$···D—O(159)/2·54/164 O(159)···D—O(264)/2·57/156···(77 Lys)
19 Ala *AB1*	D	O	
20 Asp *B1*	D	O	
21 Val *B2*	D	O	
22 Ala *B3*	D	O	
23 Gly *B4*	D	O	
24 His *B5*	D	O	
25 Gly *B6*	D	O	
26 Gln *B7*	D	O	
27 Asp *B8*	D	O	$O_{\delta 2}^{-}$···D—O(287)/2·43/151···(118 Arg) O(287)—D···O(191)/3·16/144 $O_{\delta 2}^{-}$···D—O(#256)/2·97/159
28 Ile *B9*	D	O	
29 Leu *B10*	D	O	
30 Ile *B11*	D	O	
31 Arg *B12*	D	O	$N_{\eta 2}^{+}$···O(198)/3·29 $N_{\eta 2}^{+}$···O(232)/2·42 O(232)—D···O(229)/2·76/154
32 Leu *B13*	D	O	
33 Phe *B14*	D	O	
34 Lys *B15*	D	O	
35 Ser *B16*	D	O	
36 His *C1*	D	O	
37 Pro *C2*	—	O	
38 Glu *C3*	D	O···D—O(175)/3·12/143	
39 Thr *C4*	D	O	
40 Leu *C5*	D	O	
41 Glu *C6*	D	O···D—O(251)/2·98/177	$O_{\varepsilon 1}^{-}$···D—O(157)/2·84/157 O(157)···D—O(#264)/2·56/176···(#77 Lys)
42 Lys *C7*	D	O···D—O(250)/2·65/158	N^{+}—D···O(214)/3·50/140 O(214)—D···O(268)/2·93/177···(98 Lys)
43 Phe *CD1*	D	O	
44 Asp *CD2*	N—D···O(252)/3·22/174 D	O···D—O(184)/2·79/154···(48 His) O(184)—D···O(208)/2·89/170	$O_{\delta 2}^{-}$···D—O(215)/3·25/140 O(215)···D—O(216)/3·85/153
45 Arg *CD3*	N—D···O(205)/3·43/166···$O_{\beta 31}^{-}$ (154 Heme) D	O···O(188)/2·61/165···(60 Asp)	$N_{\eta 1}^{+}$—$D_{\eta 12}$···O(199)/3·38/158···(64 His) $N_{\eta 2}^{+}$—$D_{\eta 22}$···O(199)/3·61/143···(67 Thr) $N_{\eta 2}^{+}$···O(242)/3·67 O(242)···D—O(241)/2·67/163
46 Phe *CD4*	D	O	
47 Lys *CD5*	D	O	
48 His *CD6*	D	O	$N_{\delta 1}$—$D_{\delta 1}$···O(184)/2·95/132···(44 Asp) $N_{\varepsilon 2}$—$D_{\varepsilon 2}$···O(234)/3·06/150
49 Leu *CD7*	D	O	
50 Lys *CD8*	D	O	
51 Thr *D1*	D	O	
52 Glu *D2*	D	O	$O_{\varepsilon 2}^{-}$···D—O(193)/2·95/153
53 Ala *D3*	N—D···O(233)/3·01/151		
54 Glu *D4*	N—D···O(168)/2·86/168		$O_{\varepsilon 1}$···D—O(274)/2·93/133 O(274)···D—O(263)/3·02/147 O(263)···D—O(284)/2·80/159 O(284)—D···O(285)/3·14/152 O(263)···D—O(235)/3·07/152 O(235)···D—O(257)/2·69/137 O(257)···D—O(256)/2·84/154···(#27 Asp)

Table 7 (*cont.*)

Residue number and type	Main chain N	Main chain C	Side chain
55 Met *D5*		D O	
56 Lys *D6*		D O···D—O(207)/3·29/157 O(207)···D—O(270)/3·82/164	N^+—D···O(260)/3·27/157
57 Ala *D7*		D O	
58 Ser *E1*		D O	
59 Glu *E2*		D O···D—N(209 ND_4^+)/2·90/169	$O_{\epsilon 1}^-$···N'(209 ND_4^+)/2·89
	N—D···O(240)/3·02/159		$O_{\epsilon 2}^-$···D—O(#231)/2·45/178···(119 His)
			$O_{\epsilon 1}^-$···D—O(#192)/2·69/157···(16 Lys)
60 Asp *E3*	N—D···O(239)/2·74/154		$O_{\delta 1}^-$···D—O(188)/2·71/158···(45 Arg)
61 Leu *E4*		D O	
62 Lys *E5*		D O	
63 Lys *E6*		D O···D—O(210)/2·68/160	
64 His *E7*		D O	$N_{\delta 1}$—$D_{\delta 1}$···O(199)/3·64/178···(45 Arg)
65 Gly *E8*		D O	
66 Val *E9*		D O	
67 Thr *E10*		D O	$O_{\gamma 1}$···D—O(199)/3·31/172···(45 Arg)
			$O_{\gamma 1}$—$D_{\gamma 1}$···O(211)/2·57/136
68 Val *E11*		D O	
69 Leu *E12*		D O	
70 Thr *E13*		D O	
71 Ala *E14*		D O	
72 Leu *E15*		D O	
73 Gly *E16*		D O	
74 Ala *E17*		D O	
75 Ile *E18*		D O	
76 Leu *E19*		D O	
77 Lys *E20*		D O	N'—D···O(264)/3·17/127
78 Lys *EF1*		D O	N^+—D···O(161)/3·10/162
79 Lys *EF2*		D O	N^+—D···O(190)/3·09/158
80 Gly *EF3*	N—D···O(181)/3·04/161···(82 His)		
		D O···D—N^+(218 ND_4^+)/2·80/145···(141 Asp)	
81 His *EF4*		D O	$N_{\epsilon 2}$—$D_{\epsilon 2}$···O(288)/3·35/118
82 His *EF5*		D O	$N_{\delta 1}$—$D_{\delta 1}$···O(181)/2·70/178···(80 Gly)
83 Glu *EF6*		D O	
84 Ala *EF7*		D O	
85 Glu *EF8*	N—D···O(176)/3·24/161		$O_{\epsilon 1}^-$···D—O(176)/2·95/150
86 Leu *F1*		D O	
87 Lys *F2*		D O	N^+—D···O(238)/2·69/170
88 Pro *F3*		— O	
89 Leu *F4*		D O	
90 Ala *F5*		D O	
91 Gln *F6*		D O	
92 Ser *F7*		D O	
93 His *F8*		D O	
94 Ala *F9*		D O	
95 Thr *FG1*		D O	
96 Lys *FG2*		D O	
97 His *FG3*		D O	$N_{\delta 1}$—$D_{\delta 1}$···O(203)/3·01/157
98 Lys *FG4*		D O	N^+—D···O(213)/3·29/147
			N^+···O(268)/2·78
			O(268)···D—O(#171)/3·53/173
			O(268)—D···O(#226)/3·20/165
99 Ile *FG5*		D O	
100 Pro *G1*		— O	
101 Ile *G2*		D O	
102 Lys *G3*	N—D···O(162)/3·08/147		
		O(162)—D···O(290)/3·30/174	
103 Tyr *G4*		D O	
104 Leu *G5*		D O	
105 Glu *G6*		D O	$O_{\epsilon 2}^-$···D—O(280)/2·58/130
			O(280)···D—O(282)/2·76/170···(147 Lys)
			O(280)—D···O(177)/3·36/175
106 Phe *G7*		D O···D—O(262)/2·95/136	
107 Ile *G8*		D O	
108 Ser *G9*		D O	
109 Glu *G10*		D O	
110 Ala *G11*		D O	
111 Ile *G12*		D O	
112 Ile *G13*		D O	
113 His *G14*		D O	$N_{\delta 1}$—$D_{\delta 1}$···O(261)/2·78/158
114 Val *G15*		D O	
115 Leu *G16*		D O	
116 His *G17*		D O	
117 Ser *G18*		D O	
118 Arg *G19*		D O	N_ϵ—D_ϵ···O(287)/2·83/159···(27 Asp)
			$N_{\eta 2}$—$D_{\eta 21}$···O(287)/3·26/2·50/133
119 His *GH1*		D O	$N_{\delta 1}$—$D_{\delta 1}$···O(231)/2·38/149···(#59 Glu)
			O(231)—D···$O_{\delta 2}^-$(122 Asp)/3·11/131
120 Pro *GH2*		— O	
121 Gly *GH3*		D O	
122 Asp *GH4*		D O	$O_{\delta 2}$···D—O(231)/3·11/131···(119 His)
123 Phe *GH5*		D O	
124 Gly *GH6*	N—D···N^+(178 ND_4^+)/3·05/144···(126 Asp)		
125 Ala *H1*		D O	

Table 7 (*cont.*)

Residue number and type	Main chain N	Main chain C	Side chain
126 Asp *H*2	D	O···D—O(183)/3·12/169···(133 Lys)	$O_{\delta1}^-$···D—O(172)/3·17/140 $O_{\delta1}$···D—O(173)/2·31/152 $O_{\delta2}^-$···D—O(172)/3·07/146 $O_{\delta2}^-$···D—N^+(178 ND_4^+)/2·34/155···(124 Gly) O(172)···D—O(186)/2·78/150 O(173)—D···O(171)/2·73/159
127 Ala *H*3	D	O	
128 Gln *H*4	D	O	$O_{\varepsilon1}$···D—O(189)/3·05/143 O(189)—D···O(223)/3·06/153···$O_{\zeta21}^-$(≠154 Heme) O(189)···D—O(261)/2·70/158···(113 His) O(223)···D—O(196)/2·91/153 O(196)···D—O(261)/2·59/145
129 Gly *H*5	D	O	
130 Ala *H*6	D	O	
131 Met *H*7	D	O	
132 Asn *H*8	D	O	$N_{\delta2}$—$D_{\delta22}$···O(182)/3·27/124···(136 Glu)
133 Lys *H*9	D	O···D—O(221)/3·05/131	N^+···O(183)/3·31/156···(126 Asp)
134 Ala *H*10	D	O	
135 Leu *H*11	D	O	
136 Glu *H*12	D	O	$O_{\varepsilon2}^-$···D—O(182)/2·61/155···(132 Asn) $O_{\varepsilon1}$···D—O(221)/4·20/145
137 Leu *H*13	D	O	
138 Phe *H*14	D	O	
139 Arg *H*15	D	O	
140 Lys *H*16	D	O	
141 Asp *H*17	D	O	$O_{\delta1}$···D—O(227)/3·27/156 $O_{\delta2}^-$···D—N^+(218 ND_4^+)/2·91/146···(80 Gly) N^+(218 ND_4^+)—D···O(276)/2·63/155 O(276)—D···O(247)/3·01/139
142 Ile *H*18	D	O	
143 Ala *H*19	D	O	
144 Ala *H*20	D	O	
145 Lys *H*21	D	O	N^+—D···O(237)/3·39/138 N^+···O(174 SO_4^{2-})/3·18
146 Tyr *H*22	D	O	
147 Lys *H*23	D	O	N^+—D···O(282)/2·60/160
148 Glu *H*24	D	O	$O_{\varepsilon1}^-$···D—O(164)/2·74/175···(≠152 Gln) O(164)···D—O(≠162)/3·47/174···(≠102 Lys) $O_{\varepsilon1}^-$···D—O(167)/2·97/124 $O_{\varepsilon2}^-$···D—O(167)/3·40/151
149 Leu *HC*1	D	O	
150 Gly *HC*2	D	O	
151 Tyr *HC*3	D	O	
152 Gln *HC*4	D	O	$N_{\varepsilon2}$···D—O(220)/3·41/166 $O_{\varepsilon1}$···D—O(≠164)/3·24/128···(≠148 Glu)
153 Gly *HC*5	D	O O_t^-···D—O(163)/3·65/147	O(163)···D—O(236)/3·38/141
154 Heme			$O_{\delta32}^-$···D—O(271)/2·40/173 $O_{\delta32}^-$···D—N^+(283 ND_4^+)/2·97/152 $O_{\delta31}^-$···D—O(195)/3·14/175 $O_{\delta31}^-$···D—O(205)/3·21/155···(45 Arg) N^+(283)—D···O(197)/2·60/145 O(197)···D—O(195)/3·50/144 O(195)—D···O(252)/3·11/158···(44 Asp) O(195)···D—O(203)/2·61/149···(97 His) O(252)—D···O(250)/3·21/170···(42 Lys) O(205)—D···O(252)/3·05/140 $O_{\zeta21}^-$···D—O(≠223)/3·38/156

The 87 water sites and 5 ion sites that were determined (Fig. 7) are described in Table 7.* In all, the 87 water and 5 ion molecules correspond to about 30% of the solvent in the crystal. Of the 87 water molecules, 37 are bound to side-chain atoms, 16 are bound to main-chain atoms, and 12 are in bridges between protein atoms (ten intra- and two intermolecular). Twenty-two water molecules are bound only to other water molecules. All water molecules bound to protein are hydrogen bonded to polar or charged groups. Twelve main-chain amide deuteriums form hydrogen bonds with water or ion molecules. Fifteen peptide carbonyl groups are hydrogen bonded to water or ion molecules. Only hydrophilic side chains form hydrogen bonds to water or ion molecules. Analysis of the polar nature of the amino-acid side chains by residue type (Table 8) suggests that the affinity for water occurs in the order acidic > basic > polar > nonpolar. Six out of the seven (86%) Asp residues are hydrogen bonded, and 11 out of 14 (79%) Glu residues are hydrogen

* Atomic coordinates and structure factors have been deposited with the Protein Data Bank, Brookhaven National Laboratory (Reference: 2MB5, R2MB5SF), and are available in machine-readable form from the Protein Data Bank at Brookhaven or one of the affiliated centres at Melbourne or Osaka. The data have also been deposited with the British Library Document Supply Centre as Supplementary Publication No. SUP 37032 (as microfiche). Free copies may be obtained through The Technical Editor, International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, England.

Table 8. *Water affinity of amino-acid side chains and main chains*

Amino acid	Side chain			Relative hydrophilicity		Peptide backbone	
	Number[a]	Number[b] observed	This study (%)[c]	Sweet & Eisenberg (OMH scale)[d]	Wolfenden *et al.* (kcal mol^{-1})[e]	ND %(no.)[f]	CO %(no.)[f]
Asp	7	6	86	1·31	13·31	29 (2)	29 (2)
Glu	14	11	79	1·22	12·58	29 (4)	21 (3)
Arg	4	3	75	0·59	22·31	25 (1)	25 (1)
His	12	7	58	0·64	12·62	...	—
Lys	19	10	53	0·67	11·91	5 (1)	26 (5)
Asn	1	1	100	0·92	12·07	—	—
Gln	5	2	40	0·91	11·77	—	—
Thr	5	1	20	0·28	7·27	—	—
Tyr	3	—	—	−1·67	8·50	—	—
Ser	6	—	—	0·55	7·45	—	—
Gly	11	—	—	0·67	0	18 (2)	18 (2)
Phe	6	—	—	−1·92	3·15	17 (1)	—
Ala	18	—	—	0·40	0·45	6 (1)	—

Notes: (*a*) Number of residues existing in CO myoglobin. (*b*) Number of residues that form hydrogen bonds to water. (*c*) Number in column three divided by number in column two gives the percentage in this column. (*d*) The hydrophilicity (OMH) scale was derived by Sweet & Eisenberg (1983) to optimize the match between similar sequences in three-dimensional protein structures. The scale has been normalized with a mean of 0 and a standard deviation of 1·0. (*e*) Measured by Wolfenden *et al.* (1979, 1981) from the distribution coefficient between H_2O and vapor for a small molecule representing only the side chain of the amino acid at pH 7; the value is expressed relative to that of Gly. (1 kcal = 4·187 kJ.) (*f*) Values given are the percentage of peptides which form hydrogen bonds to water, and the values in parentheses are the number of peptides.

bonded to water, *etc.* There are no waters bound to the side chains of Tyr and Ser residues, but they are involved in intraprotein hydrogen bonding. The hydrophilicity of polar side chains is expressed relative to that of nonpolar, say Gly, chains and given in Table 8. (The hydrophobicity of an amino-acid side chain is a measure of its lack of affinity for water while the hydrophilicity is the converse.) Comparisons with the refined optimal matching hydrophilicity scale (Sweet & Eisenberg, 1983) and the measurement of the hydrophilicity of the amino acid by using analogous small molecules (Wolfenden, Cullis & Southgate, 1979; Wolfenden, Andersson, Cullis & Southgate, 1981) shows a certain degree of agreement (correlation), and the values are also given in Table 8 (columns 5 and 6). Not surprisingly, the hydrophilicities of the peptides (Table 8) are related to the hydrophilicities of their side chains. Table 8 shows that there is a quantitative relationship for polar residues: the average hydrophilicity of a peptide is one third of that of its side chain. Asp *CD*2 (44) and Arg *CD*3 (45) have both amide deuterium and oxygen atoms of their peptides bound to water moelcules. The negatively charged side chain and the carbonyl oxygen of Glu *E*2 (59) are bound to an ammonium ion, and its amide is bound to a water molecule. Gly *EF*3 (80) and Gly *GH*6 (124) are in contact with ions, through which they are bridged to the side chains of Asp *H*17 (141) and Asp *H*2 (126), respectively. The amide deuterium of Gly *EF*3 (80) is bound to a water molecule. The neighboring residues of Gly *A*3 (5), Ala *D*3 (53) and Phe *G*7 (106) are Glu residues; the former three peptides form bonds to water. It should be noted again that MBCO crystals were grown from a saturated ammonium sulfate solution at pH 5·7.

Table 7 describes the water hydrogen-bonding geometry: the distance between donor and acceptor, and the angle between donor, hydrogen and acceptor. The only criterion used to define hydrogen bonds was that the donor-to-acceptor distance had to be less than 3·5 Å. Distributions of the distances and angles for these hydrogen bonds that satisfy this condition are given in Fig. 8 and Table 9. Hydrogen-bonded structures are divided into three groups according to the target atom of the bond: bonds between water and main-chain atoms, water and side-chain atoms, and water and water. The O···O distances between water and main chains vary between ~2·60 and 3·40 Å, and hydrogen-bond angles vary between ~130 and 180°. The average values of distances and angles in this water and main-chain group are 2·96 Å and 158°. The O···O distances between water and water vary between

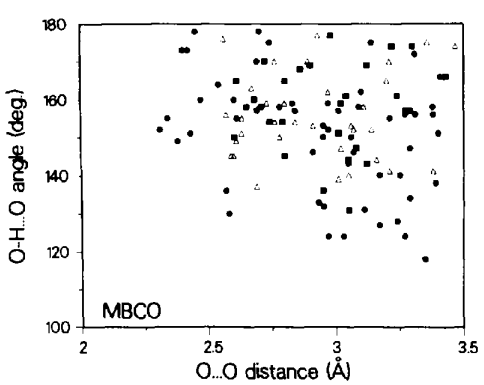


Fig. 8. Plot of O—H···O angle *versus* O···O distance: ■, hydrogen bonds between water and main-chain atoms; ●, hydrogen bonds between water and side-chain atoms; △, hydrogen bonds between water and water.

Table 9. *Observed hydrogen-bond summary*

The summary is based on a cutoff of a maximum donor-to-acceptor distance of 3·50 Å.

	O···O distance (Å)			O—D···O angle (°)		
Group	Min.	Max.	Ave.	Min.	Max.	Ave.
Water-to-main	2·60	3·43	2·96 ± 0·22	131	177	158 ± 12
Water-to-side	2·31	3·41	2·94 ± 0·39	118	178	151 ± 14
Water-to-water	2·56	3·47	2·92 ± 0·34	137	177	156 ± 11

Table 10. *Distance geometry of the ions*

The number in parentheses is the residue number; water and ion molecules are treated like a residue in the refinement.

Atom pair	Distance (Å)
N(283 ND_4^+)···$O_{\bar{B}32}$(154 heme)	2·97
N(283) ···O(197 water)	2·60
N(283) ···O(271 water)	2·35
N(218 ND_4^+)···$O_{\delta 2}^-$(141 Asp)	2·91
N(218) ···O(80 main)	2·80
N(218) ···O(276 water)	2·63
N(178 ND_4^+)···$O_{\delta 2}^-$(126 Asp)	2·34
N(178) ···D(124 main)	2·30
N(178) ···D(172 water)	2·91
N(209 ND_4^+)···$O_{\epsilon 1}^-$(59 Glu)	2·89
N(209) ···O(59 main)	2·90
S(174 SO_4^{2-}) ···N^+(145 Lys)	3·55
O(174) ···D(145 ND_3^+)	2·18
O(174) ···O(253 water)	1·89

~2·56 and 3·47 Å, and hydrogen-bond angles vary between ~140 and 180°. The average values of distances and angles in this group are 2·92 Å and 156°. We found a wider range of distances and angles between water and side chains; O···O distances vary between ~2·30 and 3·40 Å, and angles vary between ~120 and 180°. The average values of distances and angles in this group are 2·94 Å and 151°. The three groups have almost identical averages. The average hydrogen-bond distance is longer than that found in ice (Pauling, 1967) (2·76 Å), and the average angle deviates from those for linear hydrogen bonds. The environment of water is varied and some water molecules form networks containing up to seven molecules. D_2O (199), for example, is surrounded by side chains of three residues: its oxygen atom is directed towards the side chain of Arg *CD*3 (45) and the $N_{\delta 1}$—$D_{\delta 1}$ group on imidazole His *E*7 (64); one of its deuterium atoms is directed towards the side chain of Thr *E*10 (67). This water molecule site is close to the location for a sulfate ion postulated in an early X-ray study by Takano (1977*a*,*b*) on metmyoglobin; the neutron density in this site is much higher than that for a sulfate ion can be (0·257 fm $Å^{-3}$, Table 1). The water network between Glu *D*4 (54) and neighboring symmetry-related residue Asp *B*8 (27) contains seven molecules. Another seven-membered water network connects one of the two propionic acids of the heme to side chains of Lys *C*7 (42), Asp *CD*2 (44), Arg *CD*3 (45) and His *FG*3 (97). A ring of four water molecule sits above the plane of His *A*10 (12) with approach distances from ~3·2 to ~3·7 Å.

The stereochemistry of four ammonium ions and one sulfate ion is listed in Table 10. Three ammonium ions are in bridges between protein atoms. Ammonium ion (178) is in contact with the side chain of Asp H2 (126) and the main-chain amide group of residue Gly *GH*6 (124). Ammonium ion (209) contacts directly to the side chain of Glu *E*2

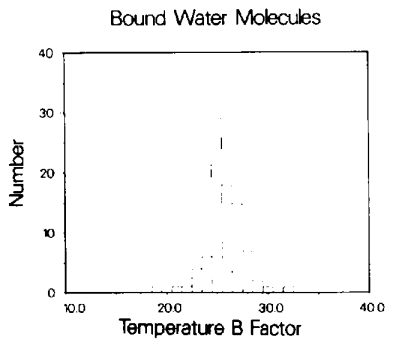


Fig. 9. Histogram of temperature factors of observed discrete water molecules.

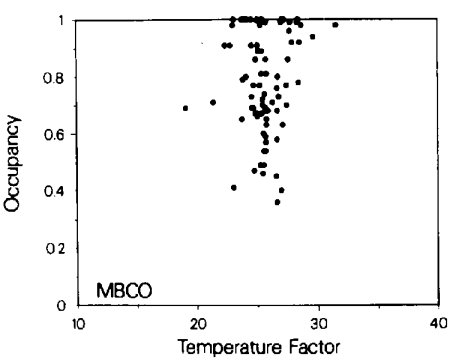


Fig. 10. Plot of the observed solvent-occupancy values against their respective temperature factors.

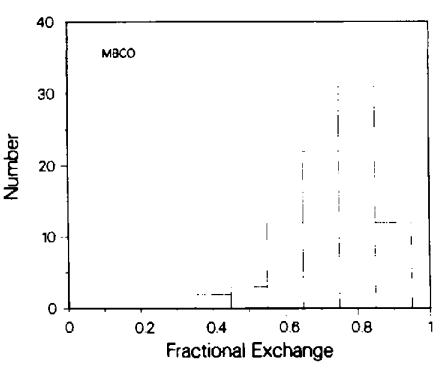


Fig. 11. Distribution of the fractional hydrogen-exchange rate of H/D atoms of observed bound water and ion molecules.

(59), and one ammonium D binds to the main-chain oxygen of same residue. Ammonium ion (218) is in contact with the side chain of Asp *H*17 (141), with one ammonium D bound to the main-chain oxygen of residue Gly *EF*3 (80) and another bound to a water molecule. One of the two propionic acids of the heme is bound to one ammonium ion (283), with two ammonium D atoms bound to two water molecules. One sulfate ion is in a channel between symmetry-related protein molecules. Intermolecular contacts are made by the side chain of Lys *H*21 (145) in contact with the sulfate ion, and by contacts *via* a water molecule and main-chain amide group of residue Glu *A*2 (4).

A histogram of temperature factors of discrete water molecules is shown in Fig. 9. There is a clear peak at $B = 25$ Å^2, corresponding to the root-mean-square vibration of 0·56 Å. Occupancy parameters were refined and are plotted against the temperature-factor values (Fig. 10). All solvent sites with occupancy factors Q of less than 0·3 were eliminated. The distribution of the observed hydrogen exchange of bound water molecules is shown in Fig. 11. There is a clear peak at $f = 0{\cdot}80$, measured as the fractional deuterium present. This value agrees well with the initial overall solvent-scattering density D_2O/H_2O that was used in crystallization of the MBCO crystals.

The detailed structure of MBCO and the comparisons of MBCO with the X-ray structures of metmyoglobin (Takano, 1977*a*), deoxymyoglobin (Takano, 1977*b*), oxymyoglobin (Phillips, 1980) and carbonmonoxymyoglobin (Kuriyan, Wilz, Karplus & Petsko, 1986) will be described elsewhere. The original MBCO data was collected in collaboration with A. C. Nunes and J. C. Norvell. The original myoglobin coordinates were provided by J. C. Kendrew and H. C. Watson.

We are grateful to Dr A. S. Brill for helpful discussions and critically reading the manuscript. We thank Drs A. D. Woodhead and S. W. White for critically reading the manuscript. This research was done under the auspices of the Office of Health and Environmental Research and calculations were performed under the supercomputing program of the US Department of Energy.

References

COOKE, R. & KUNTZ, I. D. (1974). *Annu. Rev. Biophys. Bioeng.* **3**, 95–126.

EDSALL, J. T. & MCKENZIE, A. A. (1983). *Adv. Biophys.* **16**, 53–183.

FINNEY, J. L. (1979). *Water: A Comprehensive Treatise*, Vol. 6, edited by F. FRANKS, p. 47. New York: Plenum.

GRIFFIN, J. F. (1986). Editor. *Am. Crystallogr. Assoc. Trans.* **22**.

HANSON, J. C. & SCHOENBORN, B. P. (1981). *J. Mol. Biol.* **153**, 117–146.

HENDRICKSON, W. A. (1985). *Methods Enzymol.* **115**, 252–270.

HENDRICKSON, W. A. & KONNERT, J. H. (1981). *Biomolecular Structure, Function, Conformation and Evolution*, Vol. 1, edited by R. SRINIVASAN, pp. 43–57. Oxford: Pergamon.

IBEL, K. & STUHRMANN, H. B. (1975). *J. Mol. Biol.* **93**, 255–265.

KELLENBERGER, E. (1978). *Trends Biochem. Sci.* **3**, N135–137.

KOSSIAKOFF, A. A. (1985). *Annu. Rev. Biochem.* **54**, 1195–1227.

KUNTZ, I. D. & KAUZMANN, W. (1974). *Adv. Protein Chem.* **28**, 239–345.

KURIYAN, I., WILZ, S., KARPLUS, M. & PETSKO, G. A. (1986). *J. Mol. Biol.* **192**, 133–154.

LEHMANN, M. S., MASON, S. A. & MCINTYRE, G. J. (1985). *Biochemistry*, **24**, 5862–5869.

LEHMANN, M. S. & STANSFIELD, R. F. D. (1989). *Biochemistry*, **28**, 7028–7033.

NORVELL, J. C., NUNES, A. C. & SCHOENBORN, B. P. (1975). *Science*, **190**, 568–570.

PAULING, L. (1967). *The Chemical Bond.* Ithaca: Cornell Univ. Press.

PHILLIPS, S. E. V. (1980). *J. Mol. Biol.* **142**, 531–554.

PHILLIPS, S. E. V. (1984). *Neutrons in Biology, Basic Life Sciences*, Vol. 27, edited by B. P. SCHOENBORN, pp. 305–322. New York: Plenum.

SAVAGE, H. (1986). *Water Science Reviews*, Vol. 2, edited by F. FRANKS, pp. 1–82. Cambridge Univ. Press.

SAVAGE, H. F. J. & FINNEY, J. L. (1986). *Nature (London)*, **322**, 717–720.

SCHOENBORN, B. P. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 569–575.

SCHOENBORN, B. P. (1988). *J. Mol. Biol.* **201**, 741–749.

SCHOENBORN, B. P. (1989). Unpublished results.

SWEET, R. M. & EISENBERG, D. (1983). *J. Mol. Biol.* **171**, 479–488.

TAKANO, T. (1977*a*). *J. Mol. Biol.* **110**, 537–568.

TAKANO, T. (1977*b*). *J. Mol. Biol.* **110**, 569–584.

WOLFENDEN, R., ANDERSSON, L., CULLIS, P. M. & SOUTHGATE, C. C. B. (1981). *Biochemistry*, **20**, 849–855.

WOLFENDEN, R. V., CULLIS, P. M. & SOUTHGATE, C. C. B. (1979). *Science*, **206**, 575–577.