A Global Model of the Protein-Solvent Interface

Valère Lounnas,* B. Montgomery Pettitt,* and George N. Phillips, Jr.[‡]

*W. M. Keck Center for Computational Biology and Department of Chemistry, University of Houston, Houston, Texas 77204-5641;
[‡]W. M. Keck Center for Computational Biology and Department of Biochemistry and Cell Biology, Rice University,
Houston, Texas 77251-1892 USA

ABSTRACT The solvent structure and dynamics around myoglobin is investigated at the microscopic level of detail by computer simulation. We analyze a molecular dynamics trajectory in terms of solvent mobility and probability distribution. Local events, occurring in the protein-solvent interfacial region, which are often masked by other approaches are thus revealed. Specifically, the local solvent mobility is greatly enhanced for certain locations at the protein surface and in its interior. In addition, a strong correlation between the solvent mobility and density emerges on both global and local scales. We propose a simple model where the solvent distribution measured perpendicularly to the protein surface is utilized to reconstruct the simulated network of hydration within 6 Å from the protein surface with a relative error of only 17%. The global precision of this solvation model matches results obtained with more complicated models usually used in refinement procedures in x-ray and neutron experiments but with far fewer parameters. The dramatically improved correspondence between observed and calculated x-ray intensities at low resolution relative to other methods both confirms the validity of the approach used in the MD (molecular dynamics) simulations and allows the results of this study to be implemented in solvent studies on real systems.

INTRODUCTION

The influence that proteins exert on the structure and dynamics of their aqueous surrounding for distances greater than 3–5 Å from their atomic surface is almost unknown. Classical crystallography experiments only probe the part of the protein-water interface where water is characterized by a high probability density and an implied reduced average mobility as compared with bulk. Thermodynamic experiments sensitive to changes in heat capacity (C_p) indicate that for a majority of hydrated globular proteins including lysozyme, ovalbumin, chymotrypsinogen, insulin and myoglobin, the properties of water do not differ significantly from bulk water for hydration degrees, h, greater than 0.38–0.40 (Yang and Rupley, 1979; Bull and Breese, 1968a, b; Suurkuusk, 1974; Hutchens et al., 1969). Such hydration degrees approximately correspond to the amount of water which is necessary to insure a complete coverage of the protein surface by a monolayer of tightly bound water molecules. For h values greater than 0.40 the specific heat capacity isotherm reaches a constant value which is not significantly altered as the hydration degree increases toward higher values.

Other types of thermodynamic experiments suggest, however, that more distant layers of water may be perturbed in the hydration of proteins. The dependence of the denaturation process on hydration level has been studied by differential scanning calorimetry for several proteins including β-lactoglobulin, lysozyme, chymotrypsinogen, and ovalbumin (Ruegg et al., 1975; Fujita and Noda, 1978, 1979, 1981a, b). Slight deviations from bulk water properties have been found in the denaturation temperature, T_d, and enthalpy, H_d, for hydration degrees between 0.35 and 0.75 h. Those results have been interpreted as reflecting a secondary hydration shell of water interfacing the bulk solvent with the ordered monolayer around the protein. This interpretation is in conflict with the interpretation of heat capacity isotherms which are invariant for hydration degree greater than 0.4 h.

Other experiments provide more convincing evidence of a longer-range perturbing effect of proteins on their aqueous environment and vice versa. Raleigh scattering of Mössbauer radiation experiments (RSMR) carried out on hydrated samples of human serum albumin and metmyoglobin Mb show that there is no additivity of spectral properties in the protein-water system in the entire range of studied hydration degrees; i.e., $0.05 < h < 0.75$ (Goldanskii and Krupyanskii, 1989; Krupyanskii et al., 1986; Kurinov et al., 1987a, b; Fullerton et al., 1986). This is interpreted as implying that the protein dynamics is continuously modified along the hydration coordinate. Adding water "loosens" the protein, increasing its mobility and decreasing the fraction of elastic RSMR scattering. Reciprocally, the water dynamics reaches the properties of free water for hydration degrees greater than 0.6 or 0.7 reflecting the mutual dynamic influence of both water and proteins on each other.

Measures of ^{1}H spin-lattice relaxation have been made during dehydration of lysozyme solutions approaching the dry state, during rehydration of lyophilized lysozyme powder by isopiestic equilibration and for high hydration degrees by titration with water (Fullerton et al., 1986). Breaks in the NMR response were successively found for h equal to 0.05, 0.22–0.27, and 1.22–1.62. ^{17}O and ^{2}H resonance experiments also performed on lysozyme powders similarly indicate a discontinuity in the NMR response for a higher hydration level corresponding to 1.7 or equivalently to 1400

Received for publication 23 August 1993 and in final form 17 December 1993.

Address reprint requests to B. Montgomery Pettitt at the Department of Chemistry, 4800 Calhoun, University of Houston, Houston, TX 77204-5641. E-mail: pettitt @uh.edu.

0006-3495/94/03/601/14 $2.00

waters per molecule (Lioutas et al., 1986). Analysis of the ^1H resonance data through comparison with the sorption isotherm by the D'Arcy-Watt method gave 19 mol of tightly bound water per mol of lysozyme, 148 mol of weakly bound water, and 2000 mol of so called "multilayer" water (Lioutas et al., 1987).

ESR spectra of lysozyme samples containing a noncovalently bound spin probe have exhibited strong dependence on hydration level (Rupley et al., 1980). In addition to the changes at 0.07 and 0.25 h equivalently found in the sorption and heat capacity isotherms a continuous decrease of the correlation time of the spin probe is observed until the solution value is reached for a very high level of hydration at 1.8–2.0 h.

The results cited above suggest that methods such as ESR and NMR that measure certain motional behavior can detect perturbations of (or by) water which some thermodynamic methods do not. Certain dynamic and thermodynamic properties show parallel changes for hydration levels below monolayer coverage (Yang et al., 1979; Goldanskii and Krupyanskii, 1989), and therefore it is reasonable to expect that the same should hold if there were changes above monolayer coverage. For instance, it is argued in the case of ESR measurement (Rupley et al., 1980) several layers of water may be needed for full solvation of the spin probe, even though a monolayer may be the only water perturbed by the protein. Another possibility is that the models used to interpret the resonance measurements are incomplete or that the collective motions of groups of water molecules too slightly perturbed to be detected are responsible.

X-ray crystallographic studies of the water structure around protein molecules have focused mainly on the tightly bound, immobilized water molecules, which are seen as a natural consequence of the determination of the protein structure. Some of these studies show elaborate clathrate-like and/or ring structures (Teeter, 1984). Approaches to studying the distribution of the more loosely bound, mobile water molecules by x-ray crystallography have included application of a bulk solvent "mask" of constant density everywhere outside the protein (Blake et al., 1983), traditional least squares refinement of occupancies and temperature factors on a regular grid outside the protein (Cheng and Schoenborn, 1990) and cyclical refinement of the electron densities at grid locations by Fourier inversion with phase constraints (Badger and Caspar, 1991; Badger, 1993). Each of these approaches has its strengths and weaknesses, but none of them provides a simple, accurate, and generally applicable water-protein distribution function.

Molecular dynamics simulations are capable, in principle, of revealing the details of the structure and dynamics of the solvent region at the protein water interface (Brooks and Karplus, 1986). Often the simulation analyses have focused on the solvent effect on the protein dynamics and structure rather than on the reciprocal influence of the protein on its aqueous environment (Chandrasekhar et al., 1992; Findsen, 1991). One reason is that simulations involving the presence of solvent in amounts matching hydration degrees $h > 2$ require several thousand explicit water molecules for globular proteins of sizes comparable to myoglobin. The solvent structure around small peptides in solution and in the crystal matrix of small or medium size proteins has been previously investigated by simulation methods (Hagler and Moult, 1978; Karplus and Rossky, 1980; van Gunsteren et al., 1983; Smith et al., 1991; Brooks and Karplus, 1989; Levitt and Sharon, 1988). Some recent efforts have been directed toward investigating the solvent properties in the vicinity of globular proteins (Brooks and Karplus, 1989; Levitt and Sharon, 1988).

The work we present in this paper presents a global model description of the solvent interface around globular proteins. The presence of water around myoglobin, from a previous simulation (Findsen et al., 1993), was used to investigate the solvent structure and dynamics in the range between 0 and 20 Å from the protein surface. We propose a simple parameterized model for the solvent structure at the interface with the protein.

METHOD

We will briefly review the simulation methodology. The details have been given elsewhere (Findsen et al., 1993). Initially, coordinates for the simulation were taken from the 2.0-Å resolution metmyoglobin structure by Takano (1977). The protein structure contained 1261 heavy atoms and 83 waters of hydration. It should be noted that essentially all of the crystallographic waters were exchanged during the simulation. The protein was hydrated creating a system suitable for simulation using the standard algorithm in Amber (Weiner et al., 1984) with a sample of bulk water to yield a simulation box of dimensions 56.32 × 56.32 × 44.45 Å. The total number of solvent molecules in a box included 3045 water molecules via the outlined procedure and 83 crystal waters found in the x-ray structure, making a total of 3128 solvent molecules. This gives a system about 12 mM the protein. The simple point charge (SPC) water model was used (Berendsen et al., 1981). The Leap Frog algorithm was used to propagate the equations of motion in the canonical ensemble (constant number, N, volume, V, and temperature, T) (Weiner et al., 1984). A timestep size of 2 fs was used in the propagation of the equations of motion. SHAKE was used for the hydrogenic bonds. A molecular dynamics trajectory of nearly 175 ps was then computed for metmyoglobin in water. We now consider the methodology of analysing the solvent near the protein.

A. Solvent mobility in the protein vicinity

The solvent mobility in the vicinity of a protein is conveniently measured by the diffusion coefficient D which is obtained from the slope of the mean-square displacement function,

$$\lim_{t\to\infty}\frac{d}{dt}\langle|\vec{r}(t)-\vec{r}(0)|^2\rangle = 6D, \tag{1}$$

where $\vec{r}(t)$ describes the position vector of a solvent molecule at time t. The brackets $\langle\ \rangle$ indicate that the quantity $|\vec{r}(t)-\vec{r}(0)|^2$ is averaged over both the solvent molecules and the time origins.

Unfortunately, there is a major drawback concerning the application of the above formula to characterize the solvent mobility around any solute molecule and more particularly around a large biological macromolecule. Indeed, the diffusion coefficient D obtained with Eq. 1 describes the behavior of the solvent mean square displacement regardless of the initial positions and the successive locations visited by each solvent molecule. In other words, the usual diffusion coefficient as defined in Eq. 1 has physical significance only for homogeneous and isotropic systems where averages over the ensemble of particles is a reflection of the properties of each single particle of the system averaged over an infinite period of time. In contrast, in a strongly anisotropic environment such as a protein-solvent interface, the different regions of the solvent present distinct diffusional characteristics. For instance, the water molecules attached to charged groups will often diffuse on a longer time scale than water molecules in the vicinity of uncharged or nonpolar hydrophobic groups. Consequently, an average property such as the diffusion coefficient is less useful and revealing than a corresponding local property.

It is possible, however, to use Eq. 1 to calculate the diffusional mobility in a restricted volume ΔV around a specific atomic group or region of the protein-solvent system (Ji et al., 1991; Lounnas and Pettitt, in press) by assuming that the diffusive regime will be reached in a time scale shorter than the actual residence time of water molecules in ΔV. The bulk value of D is usually found around 0.3 Å^2 ps^{-1} at 300 K when the SPC water model is used (Berendsen et al., 1981; Ji et al., 1991). Extrapolated diffusional motion of water molecules within 6-Å windows from nonpolar and polar groups have thus been determined from the slope of the mean square displacement calculated between 1 and 3 ps (Brooks and Karplus, 1989).

In the present study we have computed the diffusional mobility of water D_{uvw} at each particular point $\vec{r}_{uvw}$ of a 1-Å grid dividing the cartesian space defined by the 56.32 × 56.32 × 45.44 Å^3 periodic simulation box. The following finite difference expression,

$$6D_{uvw} = \frac{1}{(t_2-t_1)}(\langle|\vec{r}(t_2)-\vec{r}(0)|^2\rangle - \langle|\vec{r}(t_1)-\vec{r}(0)|^2\rangle), \tag{2}$$

was computed whenever $|\vec{r}(0)-\vec{r}_{uvw}| < 1$ Å. The values t_1 and t_2 were, respectively, fixed at 1 and 2 ps reasonably assuming that the diffusional regime would be reached after 1 ps and that water molecules within 1 Å from any grid point $\vec{r}_{uvw}$ at an initial instant $t = 0$ would not diffuse farther than a short distance (<2–3 Å) from its initial location $\vec{r}(0)$ in 1 ps.

B. Protein-solvent pair correlation functions

A problem similar to the one described in the previous subsection arises when dealing with the solvent structure around a non spherical solute. Specifically, the interpretation of the volume normalized, angle averaged, protein-solvent pair correlation functions, $g_i(r)$, which represent the relative probability of finding any solvent molecule j at a distance r from a specific solute atom i, is complicated when dealing with large polyatomic solute. These functions, also called radial distribution functions, are computed as follows,

$$g_{iw}(r) = \frac{1}{4\pi r^2 N_w \Delta r}\sum_{t=0}^{T}\sum_{j=1}^{N_w}\delta[|\vec{r}_i(t)-\vec{r}_j(t)| - r], \tag{3}$$

where T is the simulation length, N_w is the total number of water molecules present in the system, and $(1/4\pi r^2\Delta r)$ the normalization volume on a grid of spacing Δr. The normalization volume which is written in spherical coordinates as,

$$d\tau(r) = \left[\int_0^{\pi}\int_0^{2\pi} r^2\sin\theta\, d\theta\, d\phi\right] dr \tag{4}$$
$$=4\pi r^2\, dr$$

also includes the angles excluded by the multicenter van der Waals atomic core of the solute. This produces a profound change in the shape of $g(r)$ compared to super fluids due to the correlations with other protein atoms (not tagged) (Brooks and Karplus, 1986). To account for the protein solute excluded volume we have used the following expression,

$$g_{iw}(r) = \sum_{t=0}^{T}\sum_{j=1}^{N_w}\frac{\delta(|\vec{r}_i(t)-\vec{r}_j(t)| - r)}{N_w d\tau_\Omega(r, t)}, \tag{5}$$

which is similar to Eq. 3, except that the normalization volume d$\tau_\Omega(r, t)$, which accounts for the presence of the other protein atoms, becomes a time-dependent quantity because of the conformational fluctuations at the protein surface. The normalization volume at any instant t can then be computed as,

$$d\tau_\Omega(r, t) = \left[\int\int_{(r,\theta,\phi)\Omega} r^2\sin\theta\, d\theta\, d\phi\right] dr, \tag{6}$$

where

$$\Omega = \Omega_t(r, \theta, \phi) \tag{7}$$

represents the domain at the instant t where the points defined by the spherical angular coordinates (r, θ, ϕ) are accessible to solvent molecules. The determination of the conditional normalization volume $\Omega_t(r, \theta, \phi)$ for each increment of time t requires a larger computational effort than using Eq. 3 to evaluate the protein-solvent pair correlation $g_{iw}(r)$ for each myoglobin atom i. Notice also, that taking account of the excluded volume in Eq. 5 does not eliminate the local correlations of the protein since the function $g_{iw}(r)$ accounts also for the solvent probability distribution near other protein atomic sites. Indeed, a volume element $d\tau(\vec{r})$ at a given location $\vec{r}$ in the solvent region which is far from a protein site i may actually be close to another untagged protein atomic site i'.

C. Solvent distribution perpendicular to the protein surface

In order to explicitly quantify the influence of the protein surface on the protein-solvent pair correlation functions mentioned in the previous section we have considered another quantity we call $g_\perp(r)$. This quantity measures the probability distribution of the solvent molecules as a function of the distance r from the closest protein atom i.

$$g_\perp(r) = \sum_{t=0}^{T} \sum_{j=1}^{N_w} \frac{\delta(\mathrm{Inf}[|\vec{r}_i(t) - \vec{r}_j(t)|]_{i=1,N_w-r})}{\delta\tau(\vec{r}_j(t), k))} \qquad (8)$$

where $\mathrm{Inf}[|\vec{r}_i(t) - \vec{r}j(t)|]_{i=1,N_p}$ takes the minimum value of $|\vec{r}_i(t) - \vec{r}_j(t)|$ for each water molecule j at the instant t and for all protein atoms i. N_p is the total number of protein atoms, and k the atom for which,

$$\mathrm{Inf}[|\vec{r}_i(t) - \vec{r}_j(t)|]_{i=1,N_p} = |\vec{r}_k(t) - \vec{r}_j(t)| \qquad (9)$$

at some time t. The quantity $\delta\tau(\vec{r}_j(t), k)$ is the volume element around the location of the water molecule j at the instant t. It is defined by all vectors $\vec{r}$ of the solvent region such that,

$$\vec{r} \in \delta\,\tau(\vec{r}_j(t), k) \qquad \text{when} \quad |\vec{r}_k - \vec{r}| \leq |\vec{r}_i - \vec{r}|; \quad \forall i\,[1, \ldots, N_p]. \qquad (10)$$

The meaning of $g_\perp(r)$ is that of a conditional pair correlation function and describes the solvent solvent structure at a given distance r perpendicularly to the protein surface which is defined by the atoms directly in contact with the solvent. Notice that the pair correlation functions defined according to Eqs. 3 and 5 locally describe the solvent distribution around surface atoms of the solute, whereas the distribution function defined in Eq. 8 is a characteristic of the whole protein.

Similar to classical radial distribution functions, this method allows the study of the solvent structures in the vicinity of specific atomic sites with the condition that they are distributed on the protein surface. For instance, when evaluating $g_\perp(r)$, it is possible to select only certain types of atomic sites by restricting the increment i to a specific species such as atom type. This allows the quantitative characterization of different aspects of the protein-solvent interface for instance as a function of either the hydrophilic or hydrophobic nature of the selected species. Distinction can also be made between atomic groups in function of their belonging to the backbone or side chains.

The exact evaluation of protein-solvent perpendicular correlation functions $g_\perp(r)$ defined by Eq. 8 represents a computational challenge because the determination of $\delta\tau(\vec{r}_j(t))$ is a nontrivial problem that has to be solved at each instant t for each solvent molecule j. Therefore we made one further reduction and define an averaged perpendicular distribution functions $g_\perp^\mathrm{A}(r)$ where the protein and solvent atomic positional fluctuation are preaveraged in time separately,

$$g_\perp^\mathrm{A}(r) = \int_{\vec{r}} \delta(\mathrm{Inf}\,|\vec{r}_i - \vec{r}\,'| - r)\, d\rho^1(\vec{r}\,'), \qquad (11)$$

where

$$\vec{r}_i = \frac{1}{N_t} \sum_{t=0}^{N_t} \vec{r}_i(t), \qquad (12)$$

and, with i defined as in Eqs. 9 and 10,

$$\rho_w^1(\vec{r}) = \sum_{j=1}^{N_w} \sum_{t=0}^{N_t} \frac{\delta(\vec{r}_j(t) - \vec{r})}{\delta\tau(\vec{r})}. \qquad (13)$$

The later quantity, $\rho_w^1(\vec{r})$ which is the three-dimensional water singlet density distribution, is easy to compute on a grid from a simulation trajectory (Lounnas and Pettitt, in press) and Eq. 11 can thus be rewritten in discrete form as

$$g_\perp^\mathrm{A}(r) = \sum_u \sum_v \sum_w \delta(\mathrm{Inf}\,|\vec{r}_i - \vec{r}\,'_{uvw}| - r)\rho_w^1(\vec{r}\,'_{uvw}) \qquad (14)$$

where u, v, and w are the grid points indices along the X, Y, and Z directions, respectively.

Although intrinsically differing from $g_\perp(r)$, the function $g_\perp^\mathrm{A}(r)$ can also be viewed as a pair distribution function between the protein surface (a fixed condition) and the solvent molecules. Instead of being the average of instantaneous correlation in space, $g_\perp^\mathrm{A}(r)$ appears rather as the correlation resulting in the average positions of both protein and solvent atoms measured with respect to a reference frame attached to the protein. The function $g_\perp^\mathrm{A}(r)$ is thus not only easier to compute than $g_\perp(r)$ but also relevant to x-ray crystallographic measurements (Lounnas and Pettitt, in press). Indeed, $g_\perp^\mathrm{A}(r)$ is determined in part from the knowledge of the three-dimensional density distribution $\rho^1(\vec{r})$ which is obtained by Fourier synthesis and phase refinement of the x-ray structure factor $F(\vec{K})$.

In the present study, we have decomposed $g_\perp^\mathrm{A}(r)$ into three distinct functions $g_{\perp\mathrm{C}}^\mathrm{A}(r)$, $g_{\perp\mathrm{N}}^\mathrm{A}(r)$, and $g_{\perp\mathrm{O}}^\mathrm{A}(r)$ in such a way that

$$g_\perp^\mathrm{A}(r) = \frac{n_\mathrm{C}(r) \cdot {}^{\mathrm{A}}_{\perp\mathrm{C}}(r) + n_\mathrm{N}(r) \cdot g_{\perp\mathrm{N}}^\mathrm{A}(r) + n_\mathrm{O}(r) \cdot g_{\perp\mathrm{O}}^\mathrm{A}(r)}{n_\mathrm{C}(r) + n_\mathrm{N}(r) + n_\mathrm{O}(r)} \qquad (15)$$

where n_C, n_N, and n_O are, respectively, the number of carbon, nitrogen, oxygen, and protein sites which are closest to solvent for a given distance r. The determination of $g_{\perp\mathrm{C}}^\mathrm{A}(r)$, $g_{\perp\mathrm{N}}^\mathrm{A}(r)$, and $g_{\perp\mathrm{O}}^\mathrm{A}(r)$ is a byproduct of the evaluation of $g_\perp^\mathrm{A}(r)$. Whenever

$$\mathrm{Inf}\,|\vec{r}_i - \vec{r}_{uvw}|_{i=1,\ldots,N_p} = |\vec{r}_k - \vec{r}_{uvw}| \qquad (16)$$

for a given distance r the histogram for either $g_{\perp\mathrm{C}}^\mathrm{A}(r)$, $g_{\perp\mathrm{N}}^\mathrm{A}$ (r), or $g_{\perp\mathrm{O}}^\mathrm{A}(r)$ is incremented by the value $\rho^1(\vec{r}_{uvw})$

according to the the nature of the kth protein atom,

$$g^{A}_{\perp S}(r) = \frac{g^{1}_{\perp S}(r) + \rho^{1}_{w}(\vec{r}_{uvw})}{n_S(r) + 1} \qquad (17)$$

where S stands for either the C, N, or O sites. Notice that hydrogens have been ignored when evaluating $g^{A}_{\perp}(r)$.

D. Modeled reconstruction of the protein-solvent interface

In this section we introduce a model we have used to reconstruct the solvent density distribution for the protein-solvent interface. Such a model may find an application in addressing the general problem of x-ray crystallographic refinement of globular proteins where the lack of information on the solvent distribution in the crystal lattice results in some problems concerning the protein structure itself and confusion concerning the role of hydration water in biological systems.

We base our model on the analysis of the solvent distribution at the interface with the protein utilizing the method we describe in Part C of Method. Specifically, we make use of the perpendicular distribution function $g^{A}_{\perp}(r)$ for three type of atomic sites which are respectively the "nonpolar" sites centered on extended carbon atoms (C, CA, CB, CG, CD, . . .), the "polar" or charged sites centered on nitrogen atoms (N, NE, NZ, . . .), and the sites centered on oxygen atoms (O, OE, OD, OH). The modeled solvent density, $\rho^{1}_{m}(\vec{r}_{uvw})$, in the protein-solvent interfacial region is rebuilt from functional fitting of $g^{A}_{\perp C}(r)$, $g^{A}_{\perp N}(r)$, and $g^{A}_{\perp O}(r)$ according to

$$\rho^{1}_{m}(\vec{r}_{uvw}) = g^{A}_{\perp S}(r'), \qquad (18)$$

where

$$r' = \mathrm{Inf}(|\vec{r}_i - \vec{r}_{uvw}|)_{i=1,\ldots,N_p}, \qquad (19)$$

and S = C, N, or O according to the nature of the atom k for which $r' = |\vec{r}_k - \vec{r}_{uvw}|$. This procedure can be viewed just as the inverse of the procedure described in the previous subsection. The resulting precision of the constructed model $\rho^{1}_{m}(\vec{r}_{uvw})$ versus the actual singlet density $\rho^{1}(\vec{r}_{uvw})$ maybe measured by the relative error, R, between the two distributions,

$$R = \sum_{u}\sum_{v}\sum_{w}\frac{|\rho^{1}_{m}(\vec{r}_{uvw}) - \rho^{1}_{w}(\vec{r}_{uvw})|}{\rho^{1}_{w}(\vec{r}_{uvw})}. \qquad (20)$$

The purpose of this part of our study is to propose a mathematical model, $g^{m}_{\perp}(r)$, which mimics the behavior of the perpendicular distribution functions $g^{A}_{\perp C}(r)$, $g^{A}_{\perp N}(r)$, and $g^{A}_{\perp O}(r)$ for the complete range of distances from 0 to 20 Å. For each type of species S our model function is expressed as the following product of functions,

$$g^{m}_{\perp S}(r) = p_s(r) \times q_s(r) \times \rho_0, \qquad (21)$$

where the functions $p_s(r)$ and $q_s(r)$ are defined as

$$q_s(r) = \left\{ 1 - \exp\left[-\left(\frac{r}{\sigma(r)}\right)^{\gamma}\right] j_0\left[\frac{\pi}{2}\left(\frac{r}{\sigma(r)}\right)^{\beta}\right]\right\}^{\delta}, \qquad (22)$$

and

$$p_s(r) = 1 - A\left\{ 1 - \exp\left[-\left(\frac{a\sigma_0}{r}\right)^{b}\right] + \exp\left[-\left(\frac{cr}{\sigma_0}\right)^{d}\right]\right\}. \qquad (23)$$

The parameter σ_0 is expected to be consistent with the equilibrium distance of closest approach between the protein atomic species and the solvent molecules, j_0 is the spherical bessel function of order 0, and ρ_0 is the bulk limit of the solvent density for large τ. The function $\sigma_s(r)$ is defined as

$$\sigma(r) = \frac{1}{2}\left[\left(\frac{r}{\sigma_0}\right)^{\alpha} + 1\right]\sigma'_0. \qquad (24)$$

The general form of $g^{m}_{\perp}(r)$ and the purpose of the parameters are discussed in details in the Appendix. The resulting modeled distribution functions denoted $g\perp Cm(r)$, $g\perp Nm(r)$, and $g^{m}_{\perp O}(r)$ provide a close fit of $g^{A}_{\perp C}(r)$, $g^{A}_{\perp N}(r)$, and $g_{\perp O}(r)$. Because of its great flexibility the modeled distribution function $g^{m}_{\perp}(r)$ is intended to provide a basis for further implementation of the method in x-ray crystallographic refinement procedures.

Clearly a more precise fit can be achieved by using a finer categorization scheme than the simple atom types discussed above. Considering the differences to be found in the solvation of various atoms in different bonding (electronic and geometrical) situations (see Results) a clear improvement can be made by subdividing the atom types as is common in, for instance, molecular mechanics calculations. For this demonstration we have opted for the simpler (less quantitative) model to show the qualitative behavior of the method in general.

E. X-ray crystallographic determination of the radial solvent function

X-ray diffraction intensities were calculated from a Fourier transform of a uniform grid of values representing the solvent density. The density was determined by first finding the distance of each grid point to the nearest protein atom, and noting the distance and the atom type. The solvent function described above was then evaluated to produce the solvent density at each grid point, and the resulting function transformed to give the solvent contribution to the diffraction pattern. At this point, the solvent portion of the diffraction was entered along with the experimental data (Phillips et al., 1990) into the program XPLOR (Brünger, 1991), and the best scale factor, temperature factor, and gamma value for the solvent was determined by direct search. For the "step function" calculation, both the XPLOR "solmask" option with default parameters, which uses a hard edge approach, and a step function in our procedure were tested, also by finding the optimum bulk density and temperature factors. The difference between these latter two is that XPLOR uses a range of van der Waals radii for different classes of atoms consistent with a molecular mechanics approach, whereas our implementation considers one radius per element. The function $\rho(r)$ was set to 1.0 for these comparisons with the x-ray data.

RESULTS

A. Mobility of the solvent in the vicinity of the protein

The solvent mobility has been studied in various simulations of biopolymers performed in periodic boxes sufficiently large to obtain solvation shells four to five water molecules thick. It turns out that the diffusion coefficient, D, averaged over all water molecules present in the periodic box may differ by a factor of 2 to 3 from the value obtained with the same water model in absence of solute (Wong and McCammon, 1987; McCammon et al., 1987). Furthermore, the solvent mobility seems to be strongly dependent on the distance from the protein. This effect was observed in a simulation of trypsin in an environment of 4785 SPC water molecules where the diffusion coefficient increased continuously from 0.08 to 0.6 Å^2/ps for distances between 2.7 and 15 Å from the protein surface defined by the cartesian location of the atoms in van der Waals contact with solvent molecules (Wong and McCammon, 1987; McCammon et al., 1987).

We have computed the diffusion coefficient as local property of space on a three-dimensional grid using the method proposed in the Method section. Fig. 1 displays the average diffusion coefficient, D, as a function of the distance τ from the protein closest atom. The diffusion coefficient quickly increases from a value of about 0.1 Å^2/ps near the protein surface to a maximum value of 1.1 Å^2/ps in the (5–8)-Å interval. Then, the mobility decreases regularly and smoothly reaches the approximate value of 0.6 Å^2/ps for distances near 20 Å. Thus, the values of the diffusion coefficient found in the simulation of myoglobin at short and large distances from the protein surface are consistent with those of the trypsin simulation. However, the two results are qualitatively different since for this simulation of myoglobin the diffusion reaches a peak for intermediate distances whereas it continuously increases in the case of trypsin simulation.

In Fig. 2, a slice is cut through the three-dimensional grid to allow a detailed examination of the local fluctuation of the solvent mobility in the protein vicinity. Different contour levels are drawn from lower (light grey) to higher (dark grey) solvent mobility. Beside the global dependence of the solvent mobility on the distance from the protein, one can observe well localized spots of higher mobility in the interior region of the protein-solvent interface otherwise characterized by an overall reduced solvent mobility. Specifically, the diffusion is two to three times higher than for bulk SPC at some locations in the interior of the protein and near hydrophobic surface sites. Such locations are observed near the residues Ile^{99} and His^{97} lining the entrance of the active site (on the lower right hand side), and near Leu^{29}, Phe^{33}, and His^{64} in the heme cavity. They also appear in the hydrophobic cavities I, and J, and in the channel between the cavities K and

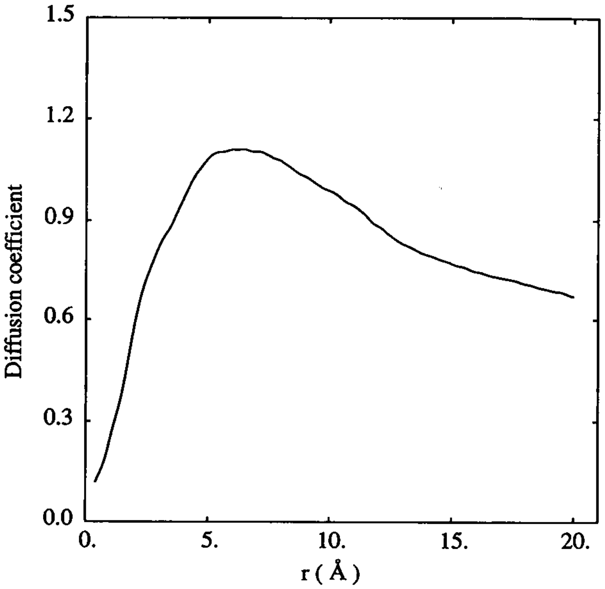

FIGURE 1 The diffusion coefficient $D(r)$ is plotted as a function of the distance r from the average cartesian location of the closest nonhydrogenic protein atoms. $D(r)$ is obtained from the local diffusion coefficient, $D(r_{ijk})$ (see Methods), computed on a grid of 1. Å meshing dividing the simulation box into 146,205 points.

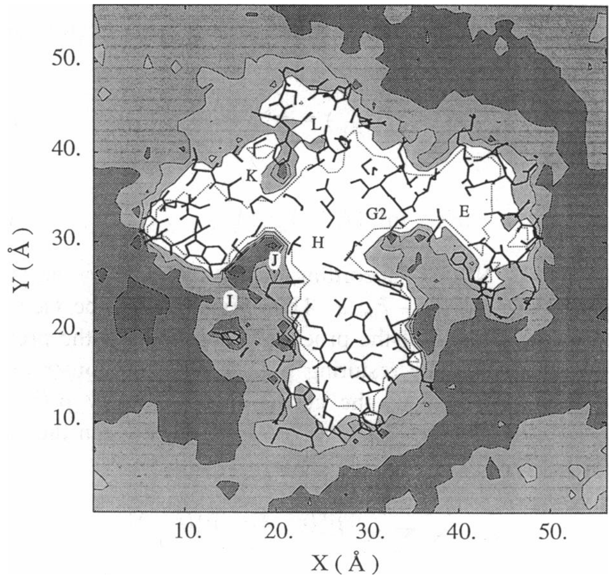

FIGURE 2 The local diffusion coefficient, $D(r_{ijk})$, obtained from the slope of the average mean-square difference $\langle |\vec{r}(0) - \vec{r}(t)|^2 \rangle$ between t_1 = 1 ps and t_2 = 2 ps is average in a 5-Å thick layer cut in the middle (X, Y) plane of the simulation box. The shading from light grey to dark grey successively indicate diffusion coefficient in the following range of values; (*a*) 0.02–0.33, (*b*) 0.33–0.66, (*c*) 0.66–1., and (*d*) 1.–1.3 Å^2/ps. The capital letters indicates the positions of the different cavities in the myoglobin. The stick figure indicate the protein atoms contained within the volume defined of the displayed slice.

L at the exact average position of the phenyl ring of Phe[122]. The later observation suggests that the phenyl ring may act as a swinging door which controls the passage of solvent molecules from the cavity K to L.

B. Protein-water pair distribution functions

As a first step, the structure of the solvent in the vicinity of the protein is studied through the examination of the classical and solute excluded solute-solvent pair correlation functions or solute-solvent radial distribution functions (Figs. 3–5). The carbon, nitrogen, and oxygen sites of the backbone and sidechains of the protein are successively considered as the solute. The effect of the volume excluded by the protein core in the determination of the protein-solvent pair distribution function is addressed for this definition by comparison with the insets. While convenient for our analysis this method does less to correct for the protein volume than previous techniques (Brooks and Karplus, 1986). Water coordination numbers and structural characteristics for the various protein-solvent pair distribution functions are gathered in Table 1. While most of the N and O sites have low coordination numbers the carbons have somewhat larger values due to the broader, more diffuse, nature of carbons radial distribution functions.

The pair correlation functions between the protein oxygens and water is shown in Fig. 3. The distribution of water around side chain oxygens appears very strongly structured with a sharp peak at 2.9 Å and a split secondary peak centered at 5.2 Å. For distances greater than 7 Å the distribution

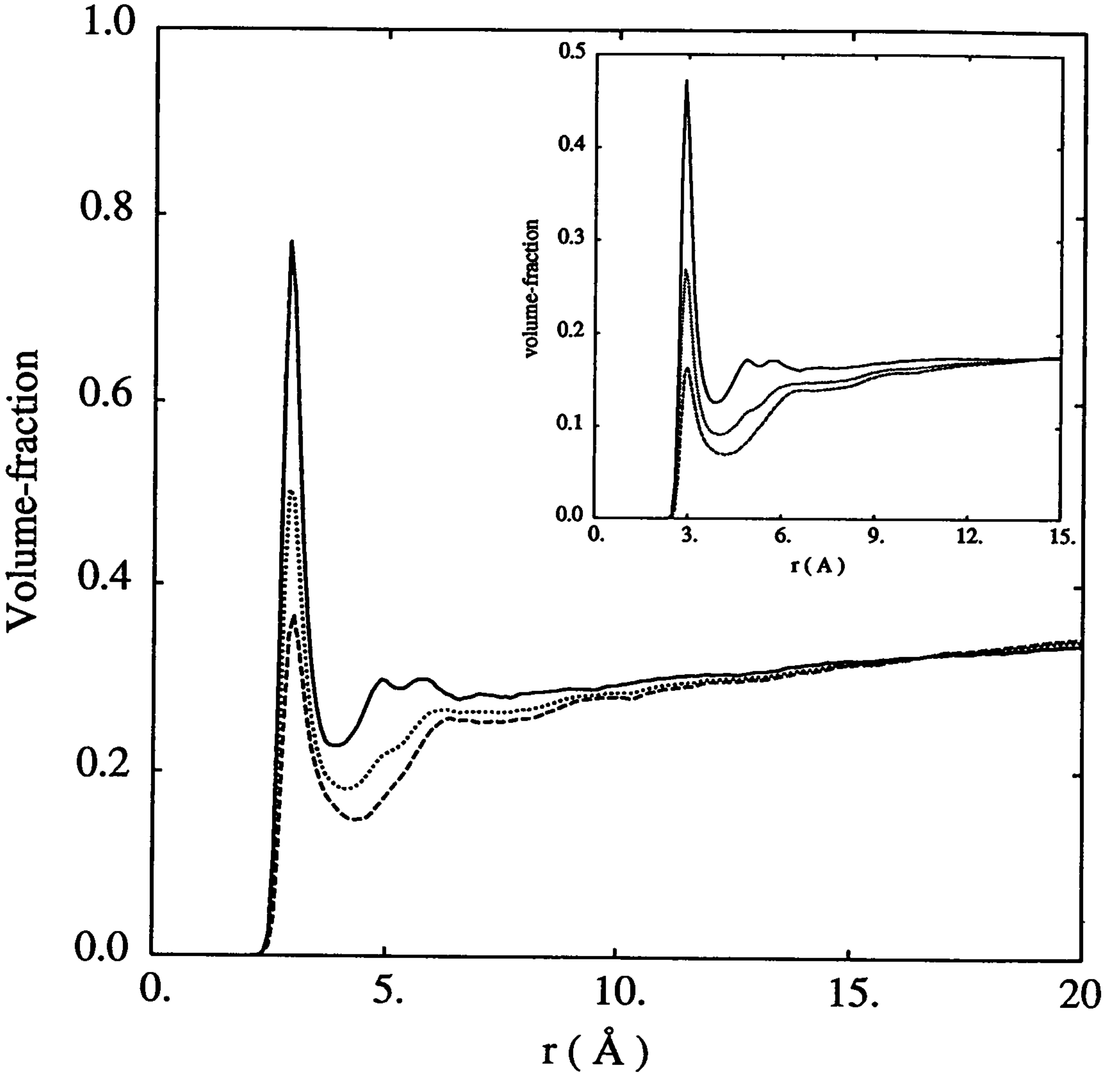

FIGURE 3 The protein-solvent pair correlation function are computed for the protein oxygen sites. The spherical volume element $\Delta\tau_\omega(r)$ accessible to the solvent at each instant t of the trajectory is computed for each protein oxygen site and for each distance increment r. The pair correlation is then computed for each site as a time average of the ratio of the number of water molecules found at distance r from the considered protein site over the volume element $\Delta\tau_\omega(r)$. The resulting pair distribution function is then normalized to volume-fraction occupancy units by multiplying it with the volume of a 2.8-Å diameter sphere to accounts for the van der Waals core of water molecules. The solid and dashed lines represent, respectively, the average water distribution for side chains and backbone oxygen sites for which the pair distribution was non zero for $r = 2.8$ Å. The dotted line is the weighted average of the two previous distributions. The inset represent similar distribution computed with a spherical volume element $\Delta\tau$ which does not account for the protein excluded volume.

TABLE 1 Characteristics of the protein-solvent pair correlation functions for solvent exposed C, N, and O sites

g (r)	N_{site}*	peaks	minima	N_{coord}‡
		(Å)	(Å)	
O_{backbone}	64	3.0	4.2	2.0
		6.7	7.0	14.0
$O_{\text{sidechain}}$	168	2.9	3.9	3.5
		4.9	5.2	8.4
		5.7	6.6	17.6
O_{total}	232	2.9	4.1	2.6
		4.9	5.2	5.6
		5.7	6.6	13.5
N_{backbone}	54	3.3	3.9	1.4
		5.1	5.4	5.7
		6.5	6.7	13.3
$N_{\text{sidechain}}$	60	3.0	4.0	3.6
		5.2	5.4	9.3
		5.7	6.7	18.0
N_{total}	114	3.0	4.0	2.6
		5.2	5.4	7.6
		6.1	6.7	15.8
C_{backbone}	116	3.4	5.4	7.8
		6.7	7.2	19.3
$C_{\text{sidechain}}$	236	4.0	5.0	4.5
		8.1	8.4	29.5
C_{total}	352	3.5	5.4	7.1

* Number of sites (of each type) in contact with the solvent which contribute to the average pair distribution $g(r)$. A particular site i is counted when $g_i(r) \neq 0$ at $r = 2.8$ Å.

‡ Coordination number computed as the volume normalized summation of $g(r)$ between 0 and each minimum.

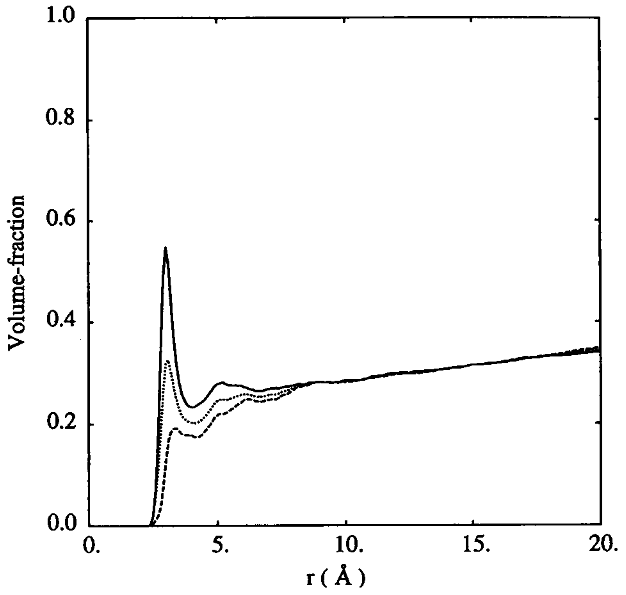

FIGURE 4 Pair distribution functions similar to those described in Fig. 3 are displayed for the protein nitrogens. The solid and dashed line represent, respectively, side chain and backbone nitrogens, whereas the dot line is their weighted average.

smoothly approaches the bulk value of the volume occupancy at about 0.33 water molecule per water molecule volume. In contrast, the pair distribution for the backbone oxygen exhibits a different quantitative behavior with a less strongly pronounced first peak at 2.9 Å.

Coordination numbers (Table 1) indicate an average of 3.5 and 17.6 water molecules in the first and second solvation shells of any side chain oxygen whereas only 2.0 and 14.0 water molecules are counted for backbone oxygens which are less exposed to the solvent. When the volume excluded by the protein is neglected (see inset Fig. 3), the general aspects of the pair correlation are essentially identical. However, the observed magnitudes are about half the volume fraction occupancies obtained when including the protein excluded volume.

Aspects of the pair distribution functions between the nitrogen atoms of the protein and the water oxygens, depicted in Fig. 4, are similar to those computed for the protein oxygens. One difference is found between the pair distributions for the backbone oxygens and nitrogen. A backbone nitrogen atom is on average less accessible to the solvent than a backbone oxygen and this is reflected in the lower correlations.

The pair distribution functions computed for carbons or "nonpolar" sites shown in Fig. 5 contrast with those obtained for polar and charged groups. Specifically, the strongly structured first peak observed near nitrogen and oxygen sites disappears in favor of a broad structure characteristic of hydrophobic solute with an equilibrium population below bulk density. More differences can be observed between the pair distributions from nonpolar sites than polar sites concerning the effect of neglecting the volume excluded by the protein (see inset Fig. 5).

The results depicted in the present section have to be compared with molecular simulation and integral equation results obtained on small peptides in polar solvent (Pettitt and Karplus, 1987; Brady, 1989). Those studies have shown that the water distribution near hydrocarbon sites has broad low intensity features for distances up to 10 to 12 Å. On the other hand, the water distribution is strongly structured with well defined peaks near polar sites and approaches bulk for distances greater than 10 Å in smaller systems. One important similarity with our results concerns the less intense structure for the water distribution near nitrogen sites involved in a peptide bond versus that of the carbonyl oxygens.

C. Distribution perpendicular to the protein surface

The perpendicular distribution functions $g\perp_{\text{C}}(r)$, $g\perp_{\text{N}}(r)$, and $g\perp_{\text{O}}(r)$ resulting from the method described in Part C of Methods are displayed in Fig. 6. They present a much smoother and simpler shape than the classical pair distribution functions described in the previous section. Principally, the first peak is less strongly resolved and has a larger width when compared to classical pair distribution functions. This is an effect of the averaging procedure resulting from the fact that the perpendicular distribution functions are actually related to the singlet density distribution of water $\rho^1_{\text{w}}(\vec{r})$ about the average locations of the protein atoms. In this case the solvent distribution is related to the time average instead of instantaneous location of the protein sites.

Another noticeable difference concerns the depletion of probability density observed in the intermediate region be-

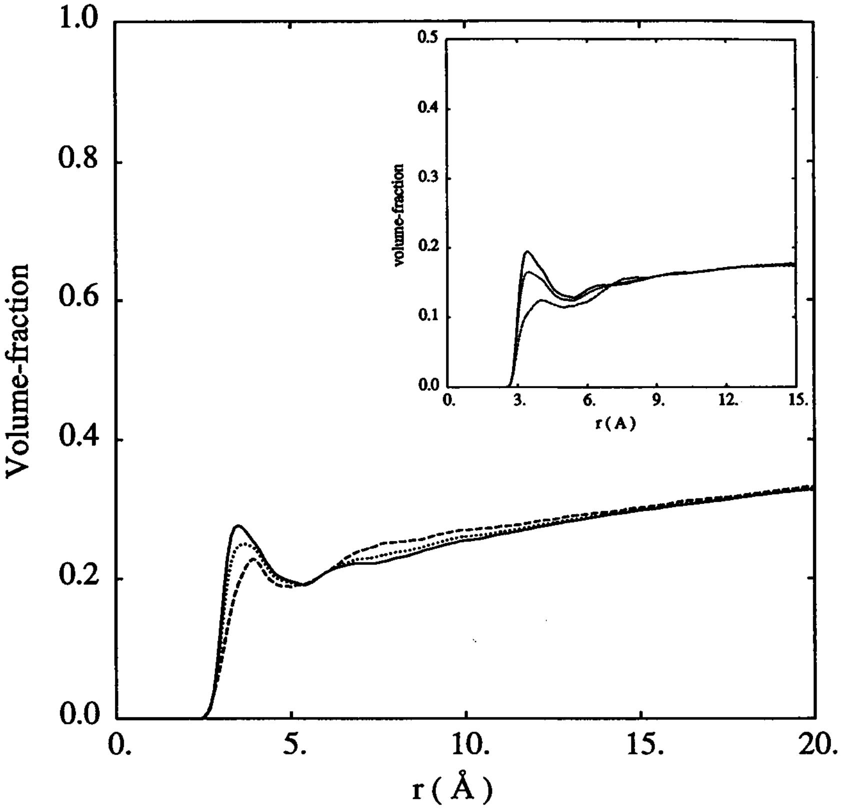

FIGURE 5 Pair distribution functions computed for the hydrocarbon sites of the protein (cf. Fig. 3). The solid and dashed line represent, respectively, side chain and backbone hydrocarbon sites whereas the dot line is their weighted average.

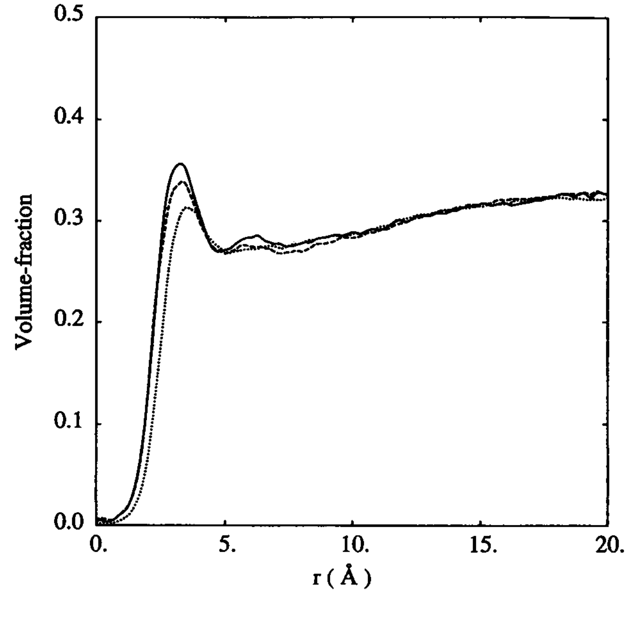

FIGURE 6 The perpendicular distribution functions between the protein O, N, C sites and the solvent are respectively depicted by the solid, dashed, and dotted lines. They represent in terms of volume fraction occupancy the probability of finding a water molecule at a given distance from the average position of the protein closest atom which is either a C, N, or O site.

tween the first layer of hydration and the bulk region extending beyond 12 Å. The origin of this observed effect is addressed later in the Discussion. However, notice that this is not simply related to the classically computed pair distribution function where the protein excluded volume is neglected (cf. inset Fig. 3) and is only slightly reflected in the regular upward drift observed in the pair distribution function when the protein excluded volume is taken into account (cf. Figs. 3, 4, and 5).

D. Reconstruction of the water network around the protein surface

In order to address the general problem of the relationships of bound water positions to protein surface topography and residue type we have attempted a reconstruction of the water three-dimensional density distribution $\rho_w^1(\vec{r})$ from the knowledge of the protein average structure and the perpendicular distribution functions $g\perp_C(r)$, $g\perp_N(r)$, and $g\perp_O(r)$. The procedure we have utilized is described in Method Section C. The resulting modeled solvent singlet density $\rho_w^1(\vec{r})$ is displayed in Fig. 7 for a two-dimensional slice cut in a (X, Y) plane of the simulation box and must be compared with the actual solvent singlet density $\rho_w^1(\vec{r})$ obtained from the simulation trajectory which is similarly depicted in Fig. 8. Both global and local details of the hydration network are fairly well reproduced by the constructed solvent model. The features of the primary layer of hydration as well as the solvent penetration into the protein interior are reproduced with an equivalent degree of accuracy. Notice also that the model allows the presence of solvent (at reasonable probability levels) in a few internal cavities inside the protein which were not accessed by the solvent during the simulation. Indeed, these cavities are large enough to accommodate water mol-

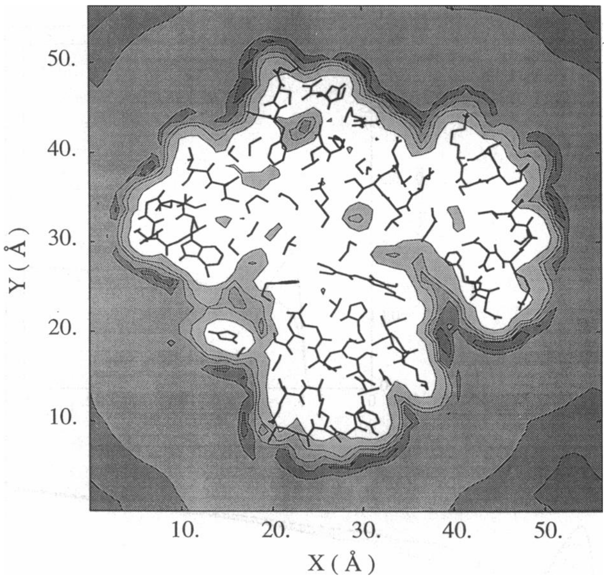

FIGURE 7 The modeled singlet density distribution $\rho^1_{\mathrm{mod}}(\vec{r})$ computed on a 1. Å meshing grid is displayed for a 5-Å thick slice cut perpendicularly to the Z axis of the simulation box. Contours of isodensity level are drawn and regions with increasing density values are shaded from light grey to dark grey. In terms of volume occupancy the five density contours drawn are respectively 10, 20, 25, 30, and 35%.

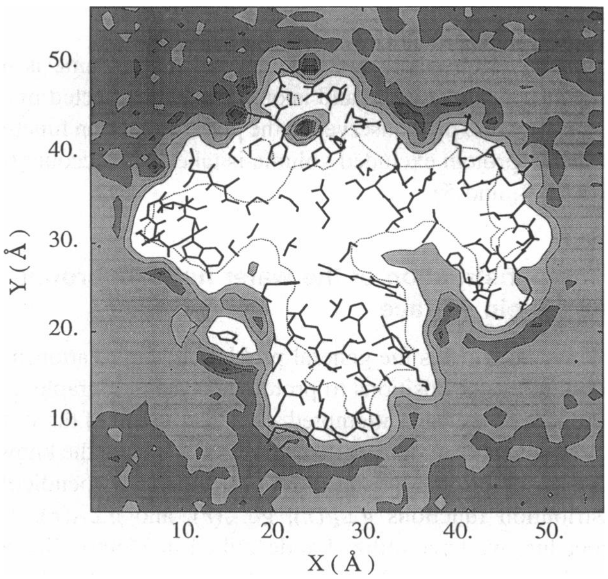

FIGURE 8 The actual singlet density distribution $\rho^1_{\mathrm{mds}}(\vec{r})$ obtained from the simulation trajectory is represented with settings similar to those of Fig. 7. The drawn contours are at 0.1, 10, 20, 30, 35, and 40% of volume occupancy. The region with the lowest density between 0.1 and 10% volume occupancy represent the intimate part of the protein-solvent interface which is dynamically accessed by the solvent molecules only through conformational changes of the protein surface.

ecules and were characterized in earlier studies on protein surface and solvent accessibility (Lee and Richards, 1971).

The relative error between the actual density distribution and the modeled density distribution is 0.17 for the hydration region within 6 Å from the protein surface and 0.11 for the complete simulation box. These values are comparable to those obtained in our previous work (Lounnas and Pettit (in press)) Ref. 31 where the solvent network was reconstructed from the knowledge of the hydration site locations, occupancy weights, and temperature factors.

E. A model for the solvent density distribution

The reconstruction described in the previous section by its simplicity suggests an interesting possibility of application in data refinement from diffraction experiments such as x-ray and neutron scattering. In those experiments the solvent structure description is often taken as a mixture of a crude model and/or a detailed one. In the first component a step function (or sigmoid) model with a fixed density value is used. In the other component each hydration water molecule is described by its cartesian coordinates, occupancy weight, and Debye-Waller temperature factor (tensor) resulting in a large number of free parameters to optimize. This number could add up to several thousand when all the possible water molecules present in the crystal are included in the refinement which leads to a nontractable problem.

We describe in Method Section D and the Appendix at the end of this paper an analytical function which is designed to mimic the protein-solvent distribution functions $g^{\mathrm{A}}_{\perp}(r)$ depicted in the previous section. This function is reasonably flexible which allows most of its features to be adjusted by a proper choice of few parameters.

As depicted in Fig. 9 the function can be adjusted to reproduce the first and second peak as well as aspects of the medium and long range behavior for the three types of perpendicular probability distributions $g^{\mathrm{A}}_{\perp\mathrm{O}}(r)$, $g^{\mathrm{A}}_{\perp\mathrm{N}}(r)$, and $g^{\mathrm{A}}_{\perp\mathrm{C}}(r)$. Table 2 gives the parameter values of the analytical form introduced in Eqs. 21, 22, and 23 of Part D of Methods used to obtain the distributions (a), (b), and (c) shown in Fig. 9. Indeed the function can be adjusted to eliminate the depletion zone between 5 and 15Å if that is an artifact related to our choice of thermodynamic state.

F. Refinement of interface

We now consider the usefulness of the approach in interpreting the low angle or low resolution x-ray data. The radial distribution function derived from the simulation can be used to calculate x-ray diffraction intensities from protein crystals of myoglobin. These calculated intensities can then be compared with experimentally determined intensities to test the validity of the functions in a real, physical system. The effect of mobile solvent on the diffraction pattern is limited to the low resolution region of the x-ray diffraction pattern (<5 Å). In fact, because of inadequate modeling of the solvent regions, these data are customarily ignored in protein crystal structure analysis. Hydration sites of high probability (Lounnas and Pettitt, in press) contribute more directly to the higher resolution data.

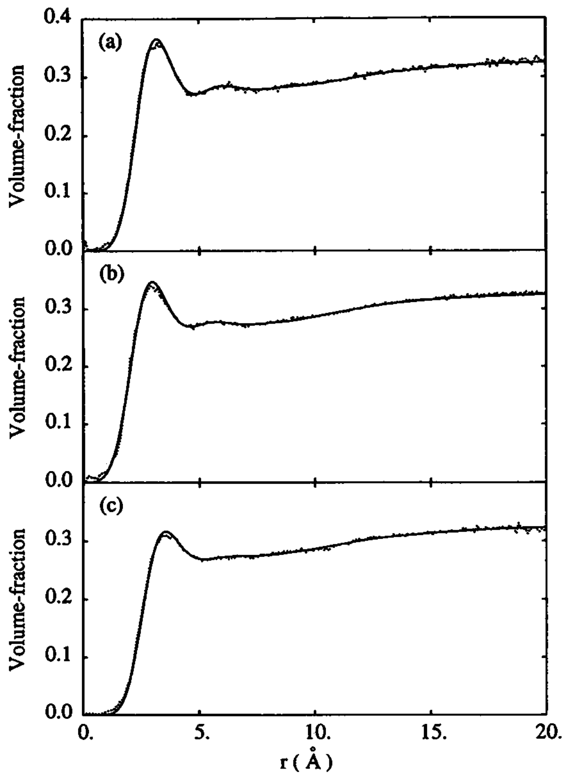

FIGURE 9 The modeled distribution functions $g_\perp^m(r)$ are displayed respectively for the oxygen (*a*), nitrogen (*b*), and hydrocarbon sites (*c*) of the protein. In each case the dotted line represent the equivalent perpendicular distribution computed from the MD trajectory.

TABLE 2 Fitting function parameters for the C, N, and O sites

Parameters	Site type C	 N	 O
σ_0 (Å)*	3.75 (3.60)	3.15 (3.05)	3.35 (3.25)
$q_s(r)$			
α	0.75	0.75	0.75
β	2.45	2.45	2.45
γ	3.00	3.40	4.40
ϵ	1.25	1.60	1.60
δ	2.00	1.10	1.10
$p_s(r)$			
A	0.17	0.17	0.15
a	2.95	3.50	3.50
b	3.60	4.00	4.00
c	∞	0.93	0.93
d	3.00	3.00	3.00

* Values in parenthesis indicate first peak positions of $g_\perp s(r)$ profiles (S = C, N, O). The given values for the parameters of σ_0 are adjusted in order to best reproduce the computed $g_\perp s(r)$ profiles.

The R factor, defined as

$$R = \frac{\sum |F_{\text{observed}} - F_{\text{calculated}}|}{\sum F_{\text{observed}}}, \qquad (25)$$

is a measure of how closely diffraction amplitudes that are calculated from a model structure match the experimental measurements. This R can be plotted as function of resolution to compare various solvent models (Fig. 10). Simply ignoring the effects of solvent clearly does not satisfactorily account for the data at low resolution, resulting in the inability to use the calculated amplitudes (and phases) in the crystallographic analysis. Using a "smoothed step function" improves the situation somewhat, but still does not give an agreement with the data that is consistent with the high resolution data or the low level of noise in the data. By using the function described in Eq. 25, considerably better fits can be achieved. In fact the best fits to experiment are obtained with a rather more "peaky" solvent distribution function obtained by setting $p(r) = 1$ (see Appendix) rather than the atom averaged functions shown in Fig. 6. With our solvent function the fit at low angles is about as good as at high angles and represents a breakthrough in the examination of solvent structure around proteins.

DISCUSSION AND CONCLUSION

The bulk limit that the diffusion coefficient seems to reach for distance greater than 15 Å from the protein surface is about twice as large as the regular value at 300 K usually obtained with pure water simulated with the SPC model. This problem also observed in other simulations of solvated proteins may be indicative of procedures currently used such as cutoff, temperature or pressure baths, and even initialization conditions. In particular our peak in the radial diffusion coefficient occurs at the solute-solvent interaction cutoff distance used in the simulation. This suggests that the results concerning our dynamic physical model systems are convoluted with our theoretical methods; a common situation in

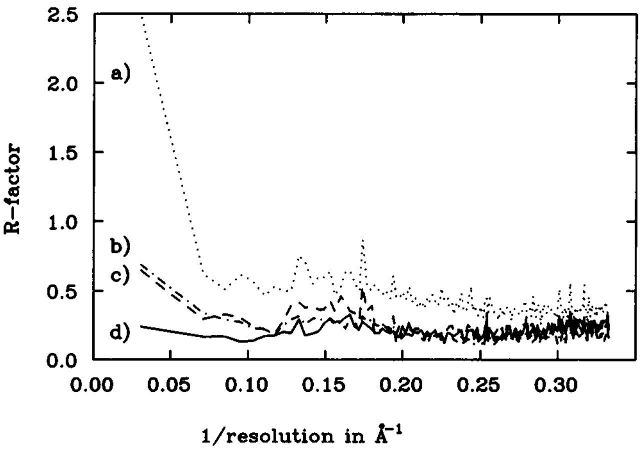

FIGURE 10 Comparison of observed and calculated diffraction amplitudes for myoglobin crystals. The crystallographic R factor is plotted as a function of one over the resolution. (*a*) No solvent correction of any kind. The curve is shifted up at higher resolution due to poor overall scaling caused by the large discrepancies at very low resolution. (*b*) Curve from XPLOR with default values, (*c*) curve from our step function with only three atom types, (*d*) curve from the radial solvent function with parameters from Table 2 but with γ values multiplied by 3.3. The lower R factor using the radial distribution solvent model supports the validity of the function in real systems.

experimental results. Thus, we will attempt to concentrate less on individual details and more on the average, global, properties found.

First, the persistence of the water structure beyond the primary layer of hydration is reflected unambiguously in the density profiles we report. This is qualitatively in agreement with the observed experimental features which relate non-bulk water properties for hydration degrees $h > 0.4$. Second, the three-dimensional network distribution of water is strongly correlated to the protein surface topography on both local and global scales. The relationship between local surface shape and bound water sites has been recently investigated else for high resolution crystal structures of few dozens of proteins (Kuhn et al., 1992). "Surfractal" surface groove accessibility (Hausdorff, 1919; Mandelbrot, 1983; Peitgen et al., 1992) shows that most x-ray water molecules bind along re-entrant curves at the surface of the protein. However, those results are biased concerning the extent of waters existing in the x-ray structure, whereas, our approach provides (within the limitations of MD methods) a direct simulation of the protein surface topography versus water distribution relationship.

One important implication of our results is the possibility of implementing a better solvent model in protein x-ray crystallography refinement procedures. This model based on our geometry versus probability distribution analysis of the protein-solvent interface (Lounnas and Pettitt, in press) allows the water density to be reconstructed in the entire crystal cell with a degree of accuracy approaching the crystallographic data. This method provides two major advantages over the classical methods. The first one is the complete elimination of the a priori need for local cartesian positioning of hydration water molecules. The second advantage is a considerable reduction in the number of degrees of freedom or parameters required to optimize the solvent description in the unit cell over other methods. The model does not preclude, however, the treatment of well-ordered water molecules as discrete atoms refined in the classical sense. The determination of a dozen parameters allows our model to account for not only the primary layer of hydration but also for additional external layers and solvent density structural features extending up to bulk levels many angstroms away. A least squares refinement of all of the water function parameters to best fit the crystallographic data is currently in progress (Phillips et al., work in progress). In addition we intend to expand the number of atom types used to obtain a better fit and test to what extent the classical well-ordered water molecules affect the fit and final result.

The function describing the radial distribution of solvent is a clear breakthrough in the analysis of the solvent structure around protein molecules. The dramatically improved correspondence between observed and calculated x-ray intensities at low resolution relative to other methods both confirms the validity of the approach used in the MD simulations and allows the results of this study to be implemented in solvent studies on real systems. For example, it would be useful to determine what the best parameters in the function are for myoglobin, and to see if different crystals with different solvent systems have different radial functions. Once this is known, crystallographers be able to make use of the low resolution experimental data in crystallographic analyses, instead of the currently common practice of ignoring the data dominated by solvent contributions.

APPENDIX

The function $g^m_\perp(r)$ we introduced in order to mimic the time averaged perpendicular distribution described under Method and Results is designed to fulfill several criteria. First, the function must obey the essential criteria characterizing any pair distribution function. Primarily, the conditions at the limits must be fulfilled,

$$g^m_\perp(0) = 0 \tag{A1}$$

and

$$\lim_{r\to\infty} g^m_\perp(r) = \rho_0 \tag{A2}$$

where ρ_0 is the bulk limit of the solvent density. Second, the short range behavior is dominated by van der Waals repulsion forces between atoms. Third, the long range tail reaches the average density smoothly.

Because of the damped oscillating structure of the perpendicular distribution obtained from the MD simulation, we have investigated the possibility of a function having the general form

$$g^m_\perp(r) = p(r) \times [1 - f(r) \times j_0\,(h(r))]^\gamma \times \rho_0 \tag{A3}$$

where j_0 is the spherical Bessel function of order 0, $p(r)$, $f(r)$, and $h(r)$ are functions to be determined in order to reproduced the features of the observed distribution functions. Implicitly, $p(r)$, $f(r)$, and $h(r)$ must obey certain requirements so that $g^m_\perp(r)$ behaves as stated above. We first assume $p(r)$ is bounded, that is that the structure is smooth for all r $[0, \infty]$, and that it approaches to 1 when r goes to large values. The purpose of $p(r)$ is discussed later at the end of this Appendix. As a consequence $f(r)$ and $h(r)$ must obey

$$f(0) = 1, \tag{A4}$$

$$h(0) = 0, \tag{A5}$$

and

$$\lim_{r\to\infty} \frac{f(r)}{h(r)} = \infty. \tag{A6}$$

The purpose of $h(r)$ and $f(r)$ is to control the oscillatory structure of $g^m_\perp(r)$ by respectively determining the peak positions and the damping regime. We propose the following forms for h and f,

$$h(r) = \frac{\pi}{2}\left(\frac{r}{\sigma(r)}\right)^\beta \tag{A7}$$

and

$$f(r) = \exp\left[-\epsilon\left(\frac{r}{\sigma_0}\right)^\delta\right], \tag{A8}$$

where

$$\sigma(r) = \frac{1}{2}\left[\left(\frac{r}{\sigma_0}\right)^\alpha + 1\right]\sigma_0'. \tag{A9}$$

The parameter σ_0 represents the average radius of exclusion between water molecules and protein atoms. Notice that σ_0 is function of the nature of the protein atomic site from which the perpendicular distribution function is considered. Theoretically, between two different species 1 and 2, the resulting hard sphere exclusion radius σ_{12} is defined as

$$\sigma_{12} = \tfrac{1}{2}(\sigma_1 + \sigma_2), \tag{A10}$$

where σ_1 and σ_2 characterize the species 1 and 2, respectively. In the present study the parameters σ_0 for each type of atom C, N, or O is preliminary interpreted as the distance corresponding to the position of the first peak in the perpendicular density distribution profile.

The function $\sigma(r)$ and the power parameters α and β are intended to control the periodicity and shape of the modeled function and are set so that the position and width of the first and second peak in $g^{A}_{\perp N}(r)$ and $g^{A}_{\perp O}(r)$ are correctly reproduced by $g^{m}_{\perp N}(r)$ and $g^{m}_{\perp O}(r)^{O}$, respectively. The value of σ'_0 has to be determined as a function of β in order to insure that the position of the first peak of $g^{m}_{\perp}(r)$ remains at its fixed predetermined value σ_0. Therefore, the constraint on $g^{m}_{\perp}(r)$ at r = σ_0 can be written as,

$$\left.\frac{dg^{m}_{\perp}(r)}{dr}\right|_{r=\sigma_0} = 0 \tag{A11}$$

which leads to the following expression when $g^{m}_{\perp}(r)$ is at critical points

$$f(r)\frac{dj_0(h(r))}{dr} + f'j_0(r)(h(r)) = 0. \tag{A12}$$

which gives

$$h(r)\cos(h(r)) + (F(r) - 1)\sin(h(r)) = 0, \tag{A13}$$

where

$$F(r) = \frac{f'(r)h(r)}{h'(r)f(r)}. \tag{A14}$$

It turns out that if the quantity $F(r)$ remains quite small (less than 1) for the complete range of distances r [0, ∞], in which the case with the functions $f(r)$ and $h(r)$ we propose, then Eq. 13 can be approximately reduced to

$$h(r)\cos(h(r)) - \sin(h(r)) = 0, \tag{A15}$$

which has a unique solution h_k in each interval $[(2k-1)(\pi/2), (2k+1)\pi/2]$ *where* k takes all values 0, 1, . . . , N. For $k = 0$, the solution $h_0 = 0$ in the interval $[-(\pi/2), \pi/2]$ and corresponds to the overall minimum of $g^{m}_{\perp}(r)$ for the value $r = 0$. For $k = 1$, the value h_1 satisfying equation 15 determines the position r_1 of the first peak which must be $r_1 = \sigma_0$ and consequently leads to the following condition

$$\frac{\pi}{2}\left(\frac{\sigma_0}{\sigma(r_1)}\right)^{\beta} = h_1, \tag{A16}$$

which after manipulation gives

$$\sigma'_0 = \sigma_0\left(\frac{\pi}{2h_1}\right)^{1/\beta}. \tag{A17}$$

The following values of $k = 2, 3, \ldots, N$ will determine the positions r_k of the successive minima and maxima of $g^{m}_{\perp}(r)$ according to k being odd or even. The positions r_k are then related to the parameter α and β through the following relation

$$\left(\frac{h_k}{h_1}\right)^{1/\beta}\left[1 + \left(\frac{r_k}{\sigma_0}\right)^{\alpha}\right] = 2r_k. \tag{A18a}$$

In the case where $\alpha = 1$ then equation 18a reduces to

$$r_k = \frac{(h_k/h_1)^{1/\beta}\sigma_0}{[2 - (h_k/h_1)^{1/\beta}]}. \tag{A18b}$$

Notice that in that case there is a limited number of possible positive solutions of r_k depending on the value of β. Thus, if $\beta = 1$ then only $k = 1$ and $k = 2$ lead to positive r_k values which will be reflected by a profile with only one peak. This effect comes from the fact that $h(r)$ goes to $2\sigma_0$ or 0 when, respectively, $\alpha = 1$ or $\alpha > 1$ (see Eq. 7 and 9). Conversely, when $\alpha < 1$ then $h(r)$ goes to ∞ which leads to an infinite number of solutions r_k.

When the value of β increases the position r_2 of the secondary peaks is displaced closer to the first peak and conversely when β is decreased the secondary peak is displaced toward larger distances. Notice that r_1 the position of the primary peak is independent of both α and β. One can control the peak magnitudes through a proper adjustment of the parameter γ.

Finally, the function $p(r)$ is introduced to provide a fine tuning of the overall aspect of $g^{m}_{\perp}(r)$. Specifically, we wanted an exact control on the depletion which is observed in the 5–15-Å range of the perpendicular density profiles $g^{A}_{\perp C}(r)$, $g^{A}_{\perp N}(r)$, and $g^{A}_{\perp O}(r)$. This was achieved by choosing the following functional form

$$p(r) = 1 - A\left\{1 - \exp\left[-\left(\frac{a\sigma_0}{r}\right)^{b}\right] + \exp\left[-\left(\frac{cr}{\sigma_0}\right)^{d}\right]\right\} \tag{A19}$$

where A, a, b, c, and d are parameters to be determined to best fit the observed perpendicular density distribution profiles. The parameter A controls the depth of the depletion whereas a, b, c, and d regulate its extent and steepness.

Notice that the form of Eq. 3 is very general and allows an infinite number of variations in the definitions of $p(r)$, $h(r)$ and $f(r)$.

B. M. P. and V. L. thank the National Institutes of Health (NIH), and the Robert A. Welch Foundation for partial support. The National Science Foundation is acknowledged for local computational support. G. N. P. thanks the W.M. Keck Foundation, NIH Grant AR40252, and Robert A. Welch Foundation Grant C-1142. G. N. P. also thanks Kenneth Forman for helpful discussions.

REFERENCES

Badger, J. 1993. Multiple hydration layers in cubic insulin crystals. *Biophy. J.* 65:1656–1659.

Badger, J., and D. L. D. Caspar. 1991. Water structure in cubic insulin crystals. *Proc. Natl. Acad. Sci. USA.* 88:622–626.

Berendsen, H. J. C., J. P. M. Postma, W. F. van Gunsteren, and J. Hermans. 1981. *In* Intermolecular Forces. B. Pullman, Editor. Reidel, Dordrecht.

Blake, C. C. F., W. C. A. Pulford, and P. J. Artymiuk. 1983. X-ray studies of water in crystals of lysozyme. *J. Mol. Biol.* 167:693–723.

Brady, J. 1989. Molecular dynamics simulation of α-D-glucose in aqueous solution. *J. Am. Chem. Soc.* 111:5155–5165.

Brooks, C. L., III, and M. Karplus. 1986. Theoretical approaches to solvation of biopolymers. *Methods Enzymol.* 127:369–400.

Brooks, C. L., III, and M. Karplus. 1989. Solvent effects on protein motion and protein effects on solvent motion. *J. Mol. Biol.* 208:159–181.

Brünger, A. T. 1991. XPLOR Version 3.1. Yale University Press, New Haven, CT.

Bull, H. B., and K. Breese. 1968a. Protein hydration. I. Bonding sites. *Arch. Biochem. Biophys.* 128:497–502.

Bull, H. B., and K. Breese. 1968a. Protein hydration. II. Specific heat of egg albumin. *Arch. Biochem. Biophys.* 128:488–496.

Chandrasekhar, I., G. M. Clore, A. Szabo, A. M. Gronenborn, and B. R. Brooks. 1992. A 500 ps molecular dynamics study of interleukin-1β in water. *J. Mol. Biol.* 226:239–250.

Cheng, X., and B. P. Schoenborn. 1990. Hydration in protein crystals. A neutron diffraction analysis of carbonmonoxymyoglobin. *Acta Cryst.* B46:195–208.

Findsen, L. A., S. Subramanian, V. Lounnas, and B. M. Pettitt. 1993. *In* Principle of Molecular Recognition. A. D. Buckingham Editor. Chapman Hall, London.

Fujita, Y., and Y. Noda. 1978. Effect of hydration on the thermal stability of protein as measured by differential scanning calorimetry. *Bull. Chem. Soc. Jpn.* 51:1567–1568.

Fujita, Y., and Y. Noda. 1979. Effect of hydration on the thermal stability of protein as measured by differential scanning calorimetry: lysozyme-water-d2 system. *Bull. Chem. Soc. Jpn.* 52:2349–2352.

Fujita, Y., and Y. Noda. 1981a. Effect of hydration on the thermal stability of protein as measured by differential scanning calorimetry: chymotrypsinogen. *Int. J. Pept. Protein Res.* 18:12–17.

Fujita, Y., and Y. Noda. 1981b. Effect of hydration on the heat stability of proteins. *Bull. Chem. Soc. Jpn.* 54:3233–3234.

Fullerton, G. D., V. A. Ord, and I. L. Cameron I. L. 1986. An evaluation of the hydration of lysozyme by an NMR titration method. *Biochim. Biophys. Acta.* 869:230–246.

Goldanskii, V. I., and Y. F. Krupyanskii. 1989. Protein and protein-bound water dynamics studied by Raleigh scattering of Mössbauer radiation (RSMR). *Quart. Rev. Biophys.* 22l:39–92.

Hagler, A. T., and J. Moult. 1978. Computer simulation of the solvent structure around biological macromolecules. *Nature (Lond.).* 272: 223–226.

Hausdorff, F. 1919. Dimension und äusseres Mass. *Mathematische Annalen.* 79:157–179.

Hutchens, J. O., A. G. Cole, and J. W. Stout. 1969. Heat capacities from 11 to 305° K, and entropies of hydrated and anhydrous bovine zinc insulin and bovine chymotrypsinogen A. Entropy change for formation of peptide bonds. *J. Biol. Chem.* 244:26–32.

Ji, J., T. Çagin, and B. M. Pettitt. 1991. Dynamic simulation of water at constant chemical potential. *J. Chem. Phys.* 96:1333–1342.

Karplus, M., and P. Rossky. 1980. Solvation: a molecular dynamics study of a dipeptide in water. *Biopolymers.* 23–42.

Krupyanskii, Y. F., I. V. Sharkevitch, Y. I. Khurgin, I. P. Suzdalev, and V. I. Goldanski. 1986. Investigation of trypsin hydration by RSMR. *Mol. Biol.* 20:1356–1363.

Kuhn, L. A., M. A. Siani, M. E. Pique, C. L. Fisher, E. D. Getzoff, and J. A. Tainer. 1992. The interdependence of protein surface topography and bound water molecules revealed by surface accessibility and fractal density measures. *J. Mol. Biol.* 228:13–22.

Kurinov, I. V., Y. F. Krupyanskii, I. P. Suzdalev, and V. I. Goldanskii. 1987a. The study of the influence of hydration of dynamics of some globular proteins by Raleigh scattering of Móssbauer radiation. *Biofizika.* 32:210–214.

Kurinov, I. V., Y. F. Krupyanskii, I. P. Suzdalev, and V. I. Goldanskii. 1987b. RSMR study of the hydration effects on the dynamics of some globular proteins. *Hyperfine Interact.* 33:223–232.

Lee, B., and F. M. Richards. 1971. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Bio.* 55:379–400.

Levitt, M., and R. Sharon. 1988. Accurate simulation of protein dynamics in solution. *Proc. Natl. Acad. Sci. USA.* 85:7557–7561.

Lioutas, T. S., I. C. Baianu, and M. P. Steinberg. 1986. Oxygen-17 and deuterium nuclear magnetic resonance studies of lysozyme hydration. *Arch. Biochem. Biophys.* 247:68–75.

Lioutas, T. S., I. C. Baianu, and M. P. Steinberg. 1987. Sorption equilibrium and hydration studies of lysozyme: water activity and 360-MHz proton NMR measurements. *J. Agric. Food Chem.* 35:133–137.

Lounnas, V., and B. M. Pettitt. Distribution implied dynamics versus residence times and correlations: solvation shell of myoglobin. *Proteins: Struc. Funct. Genet.* In press.

Mandelbrot, B. B. 1983. The fractal geometry of nature. W.H. Freeman Editor, New York. pp 109–115.

McCammon, J. A., O. A. Karim, T. P. Lybrand, and C. F. Wong. 1987. Ionic association in water: from atoms to enzymes. *Ann. N.Y. Acad. Sci.* 210–217.

Peitgen, H. O., H. Jürgens, and D. Saupe. 1992. Fractals for the classroom. Springer-Verlag Editor, New York. 218–239.

Pettitt, B. M., and M. Karplus. 1987. The structure of water surrounding a peptide: a theoretical approach. *Chem. Phys. Lett.* 136:383–386.

Phillips, G. N., Jr., R. M. Arduini, B. A. Springer, and S. G. Sligar. 1990. *Proteins Struct. Funct. Genet.* 7:358–365.

Ruegg, M., U. Moor, and B. Blanc. 1975. Hydration and thermal denaturation of β-lactoglobulin calorimetric study. *Biochim. Biophys. Acta.* 400: 334–342.

Rupley, J. A., P. H. Yang, and G. Tollin. 1980. Thermodynamic and related studies of water interacting with proteins. *ACS Symp. Ser.* 127:111–132.

Smith, P., L. Dang, and B. M. Pettitt. 1991. Simulation of the structure and dynamics of the bis(penicillamine) enkephalin zwitterion. *J. Am. Chem. Soc.* 113:67–73.

Smith, P. and B. M. Pettitt. 1991. Effects of salts on the structure and dynamics of the bis(penicillamine) enkephalin zwitterion: a simulation study. *J. Am. Chem. Soc.* 113:6029–6037.

Suurkuusk, J. 1974. Specific heat measurements on lysozyme chymotrypsinogen, and ovalbumin in aqueous solution and in solid state. *Acta Chem. Scand. Ser. B.* 28:409–417.

Takano, T. 1977. Crystal structure of myoglobin. *J. Mol. Biol.* 110:537–568.

Teeter, M. M. 1984. Water structure of a hydrophobic protein at atomic resolution: pentagon rings of water molecules in crystals of crambin. *Proc. Natl. Acad. Sci. USA.* 81:6014–6018.

van Gunsteren, W. F., H. J. C. Berendsen, J. Hermans, W. G. J. Hol, and J. P. M. Postma. 1983. Computer simulation of the dynamics of hydrated protein crystals and its comparison with X-ray data. *Proc. Natl. Acad. Sci. USA.* 80:4315–4319.

Weiner, S. J., P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner. 1984. *J. Am. Chem. Soc.* 106:825–833.

Wong, C. F., and J. A. McCammon. 1987. Computer simulation and the design of new biological molecules. *Isr. J. Chem.* 27:211–215.

Yang, P. H., and J. A. Rupley. 1979. Protein-water interactions. Heat capacity of the lysozyme-water system. *Biochemistry.* 18:2654–2661.