J. Mol. Biol. (1994) **243**, 100–115

Protein Hydration Observed by X-ray Diffraction

Solvation Properties of Penicillopepsin and Neuraminidase Crystal Structures

Jian-Sheng Jiang and Axel T. Brünger

The Howard Hughes Medical Institute and
Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520 U.S.A.

Solvation in macromolecular crystal structures was studied by analyzing X-ray diffraction data of two proteins, penicillopepsin and neuraminidase. The quality of several solvent models was assessed by complete cross-validation in order to prevent overfitting the diffraction data. Radial solvent distribution functions were computed from electron density maps using phases obtained from multiple isomorphous replacement and from the protein's atomic model combined with the best solvent model. Distribution functions were computed around hydrophilic and hydrophobic groups on the protein's surface. Averaging of the distribution functions was performed in order to reduce the influence of noise. The first solvation shell is characterized by a peak in the average distribution functions. At 1·8 Å resolution, polar groups show a sharp peak while non-polar groups show a broad one. The distinction between hydrophobic and hydrophilic solvation sites is lost when using lower resolution (2·8 Å) diffraction data. Higher-order solvation shells are not observed in the average distribution functions. We hope that site-specific radial distribution functions obtained from high-quality diffraction data will produce a picture of macromolecular solvation consistent with available experimental data and computational results.

Keywords: X-ray crystallography; solvation; refinement; cross-validation; radial distribution function

1. Introduction

Water associated with the solvation of macromolecules plays an important role in biological processes such as enzymatic reactions, specific and non-specific macromolecular association and oligomerization, and ligand-binding. In transcriptional control, for example, water molecules can contribute to the specificity of the interactions between proteins and DNA. Water-mediated hydrogen bonds are found between bases and side-chains in a number of protein-DNA crystal structures (e.g. Otwinowski *et al.*, 1988; Hegde *et al.*, 1992). Comparative analysis of the water structure in the crystal structures of complexed and free DNA *trp* operator indicates several hydration sites in the free DNA that mediate specific protein-DNA interactions (Shakked *et al.*, 1994). Bound water is also affected by osmotic and hydrostatic pressure resulting in changes of specific protein-DNA binding (Robinson & Sligar, 1994). It is therefore not unreasonable to consider bound water as part of the macromolecule. This idea is supported by theoretical studies of lyzozyme solvation of Venable & Pastor (1988) who showed that simulated translational and rotational diffusion constants only agree with experimental data if bound water is included as part of the protein.

Solvent constitutes a large portion of the volume in macromolecular crystals (Matthews, 1968). Fully occupied hydration sites, "bound" or "ordered" water, only represent a small fraction of the solvent. The remaining solvent is disordered but not completely featureless (Lounnas *et al.*, 1992).

A significant body of experimental and theoretical work has been aimed at understanding the solvation of macromolecules (Squire & Himmel, 1979; Teeter, 1984, 1991; Saenger, 1987; Venable & Pastor, 1988; Levitt & Sharon, 1988; Thanki *et al.*, 1989; Clore *et al.*, 1990; Otting *et al.*, 1991; Kossiakoff *et al.*, 1992; Lounnas *et al.*, 1992, 1994; Steenivasan & Axelsen, 1992; Kuhn *et al.*, 1992; Parak *et al.*, 1992; Komeiji *et al.*, 1993; Steinbach & Brooks, 1993; Brunne *et al.*, 1993; Clore *et al.*, 1994). The predominant experimental techniques to study macromolecular solvation are X-ray crystallography and solution NMR spectroscopy. The former method can give structural information, while the latter can provide both structural and dynamical insight (Levitt & Park, 1993). An X-ray crystal structure represents a time and spatial average of the electron density distribution, so observed electron density represents a probability of an atom residing at a particular position. A fully occupied hydration site thus

Table 1
van der Waals radii used for solvent mask calculations

Atom type	Radius (Å)
Carbonyl and ring carbon	2·1
Other carbons	2·3
Nitrogen	1·6
Oxygen	1·6
Sulfur	1·9

represents a high probability for the presence of a water molecule, and not necessarily a continuous presence. Residence times of bound water molecules were observed by NMR spectroscopy and were generally found to be in the sub-nanosecond range (Otting *et al.*, 1991; Clore *et al.*, 1994).

To compute an image of the electron density in the crystal, observed intensity data must be augmented with phase information. Experimental phase information obtained by multiple isomorphous replacement (MIR†) is often prone to errors because of non-isomorphism of the derivative crystals. Thus, it is commonplace to provide phases through an appropriately chosen model, especially at the later stages of the structure determination process.

One approach to modelling solvent is atomistic, using individual scattering atoms for each solvent molecule. While this approach is reasonable for modelling fully occupied hydration sites is it less appropriate for the bulk solvent regions which are largely disordered. Furthermore, atomistic modelling of solvent introduces a large number of adjustable parameters, increasing the danger of overfitting the data (Brünger, 1992*a*). For example, Parak *et al.* (1992) refined a Monte Carlo derived configuration of water molecules in the unit cell of myoglobin crystals. The achieved improvement of the fit to the diffraction data must be viewed with caution since the refinement was carried out without cross-validation.

It is more appropriate to use continuous models of electron density to describe disordered solvent. The simplest possible model comes from the simplest possible assumptions, that scatterers in bulk solvent are positioned with equal likelihood everywhere, giving rise to constant, or flat electron density outside the macromolecule. Implementations of this "flat" model use Babinet's principle (Fraser *et al.*, 1978; Moews & Kretsinger, 1975) or the molecular surface to define a "solvent mask" (Phillips, 1980). In the vicinity of the solvated macromolecule one expects deviations from this flat model. A more detailed "radial shell" model divides the solvent volume into shells of constant (but possibly different) electron density extending outward from the macromolecular surface (Cheng & Schoenborn, 1990). It indicated the presence of two radial solvation shells in a neutron diffraction structure of myoglobin. The radial shell

† Abbreviations used: MIR, multiple isomorphous replacement; NMR, nuclear magnetic resonance; r.m.s., root-mean-square.

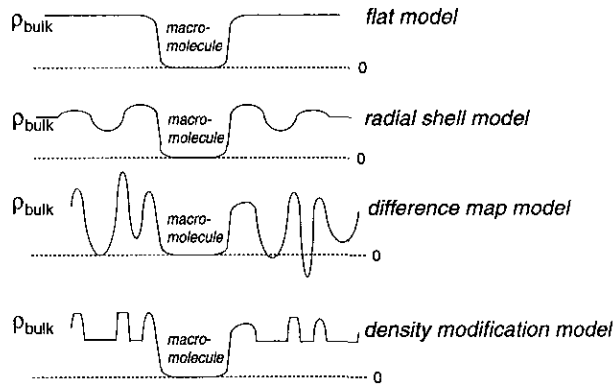

Figure 1. Schematic illustration for the 4 solvent models that were tested: flat model, radial shell model, difference map model and density modification model. The models are described in detail in the text.

model was devised to minimize the number of adjustable parameters by radial averaging. However, the surface of a macromolecular is normally anisotropic both in terms of shape and chemical composition. Badger & Casper (1991) attempted to model the resulting anisotropic solvent distribution through an iterative difference map procedure. This approach suggested non-random arrangements of water molecules extending several layers from the first solvation shell in a crystal structure in insulin. It must be remembered that all these observations are at relatively low electron density levels, close to the noise limit of the data, so overfitting is a definite possibility if precautions are not taken.

Here, two X-ray crystal structures are chosen as test cases in order to assess the quality of several solvent models and to analyse the solvent distribution around selected hydrophilic and hydrophobic groups. The crystal structure of penicillopepsin at 1·8 Å by James & Sielecki (1983) is selected because the observed MIR phases were of exceptionally high quality, which is useful for the validation of the solvent models. The crystal structure of neuraminidase from the N9 subtype of influenza virus at 2·2 Å resolution (Tulip *et al.*, 1991) is chosen because of the presence of large, 20 Å-wide, solvent pockets in the unit cell which allows us to investigate solvent distributions far away from the protein crystal lattice.

Solvent model quality is assessed by using the cross-validated of free R value and by evaluating the difference between model and MIR phases. In our original definition of R_{free}, 10% of the data are left out of the refinement process and later used to validate the model (Brünger, 1992*a*, 1993). At low resolution, however, statistical fluctuations become large due to a deminished number of reflections. We therefore use complete cross-validation (Brünger *et al.*, 1993), repeating the process ten times to measure the predictability of all reflections.

Radial distribution functions are a powerful tool for analysing the structural properties of liquids in the bulk phase and around solutes obtained from X-ray scattering experiments and computer simulations

Table 2
Least square minimization of ρ_s and B_s for the flat solvent model using penicillopepsin diffraction data

Cycle	ρ_s	B_s	k_{low}	k_{high}
0	1·000	0·00	0·631	0·894
1	0·791	7·61	0·871	0·960
7	0·378	36·37	0·971	1·004
8	0·375	36·25	0·969	1·004

(Allen & Tildesley, 1987). Site-specific solvation at the macromolecular surface, obtained by computer simulation, was analysed by using a proximity criterion for de-composing the radial distributions (Mehrotra & Beveridge, 1980). We apply this approach to the observed electron density maps. Since the electron density in a crystal structure obeys crystallographic symmetry we generalize Mehrotra & Beveridge's method by taking into account all symmetry mates of the macromolecule. Averaging over many sites on the protein's surface will reduce the amount of noise present in the distribution functions and allows one to determine if a particular feature is statistically significant.

In the Materials and Methods section we summarize pertinent information about the diffraction data of penicillopepsin and neuraminidase. In the Theory section we describe the various solvent models, the method of complete cross-validation, and the algorithm used for computing the site-specific solvent distribution functions. In the Results and Discussion section we compare the quality of the solvent models and analyse the average distribution functions around selected hydrophobic and hydrophilic atoms on the protein's surface.

2. Materials and Methods

(a) *Penicillopepsin*

The first test case was the crystal structure of penicillopepsin from *Penicillium janthinellum* consisting of 323 amino acid residues and 320 ordered water molecules. It was solved by James & Sielecki (1983) with diffraction data collected to 1·8 Å resolution by Hsu *et al.* (1977). The space group is $C2$ with unit cell dimensions $a = 97{\cdot}37$ Å, $b = 46{\cdot}64$ Å, $c = 65{\cdot}47$ Å, $\beta = 115{\cdot}4°$. The crystals contained 38% (v/v) solvent and were grown from a 1·4 M $(NH_4)_2SO_4$ solution which corresponds to roughly 15 SO_4^{2-} and 30 NH_4^+ ions per asymmetric unit. Observed diffraction data were available up to 1·8 Å resolution, 95% complete between 2 and 23 Å resolution, and 99% complete between 6 and 23 Å resolution. All observed reflections with $|F_{obs}| > 2\sigma$ were used for our studies. Phase accuracy of atomic and solvent models was assessed by using the experimental phases to 2·8 Å obtained from MIR using eight heavy-atom derivatives (Hsu *et al.*, 1977) with an overall figure of merit of 0·9.

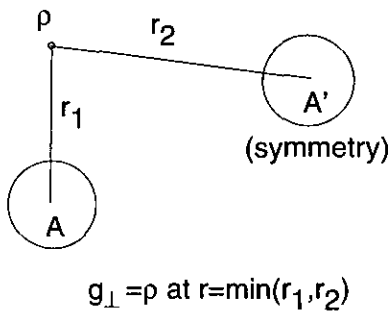

Figure 2. Radial distribution function $g_\perp$ in the presence of symmetry operations. The solvent density point ρ contributes to $g_\perp(r)$ at $r = \min(r_1, r_2)$, where r_1, r_2 are the distances between ρ and the centers of the two symmetry-related atoms A and A', respectively.

About half of the ordered water molecules in the deposited penicillopepsin crystal structure were only partially occupied (186 water molecules with occupancy $q<0{\cdot}75$ and 134 with $q > 0{\cdot}75$). We omitted all partially occupied water molecules with $q<0{\cdot}75$ from our calculations because they were not hydrogen-bonded to protein atoms or other water molecules and the corresponding electron density values for the oxygen atoms were quite weak.

(b) *Neuraminidase*

The second test case was the crystal structure of neuraminidase from the N9 subtype of influenza virus consisting of 388 amino acid residues, several asparagine *N*-linked monosaccharide and mannose ligands, and 94 fully-occupied ordered water molecules (Tulip *et al.*, 1991). The space group is cubic ($I432$) with unit cell dimension $a = 185{\cdot}1$ Å. The crystals contained 62·5% (v/v) solvent and were grown from 1·9 M phosphate solution which corresponds to about 95 ions per asymmetric unit. Observed diffraction data were 69·7% complete over the observed range between 2·21 and 76 Å, with nearly complete data at low resolution. All observed reflections with $|F_{obs}| > 2\sigma$ were used for our studies. MIR data were unavailable for this crystal structure since it was solved by difference Fourier techniques.

(c) *Computations*

All mask calculations, solvent refinements, electron density map calculations and solvent distribution analyses were carried out with a developmental version of X-PLOR (Brünger, 1992*b*) which will be made available in a future release through established procedures. Requests should be sent to A.T.B.

Graphical display was carried out using an interface between X-PLOR and the AVS graphics system (Advanced Visuals, Inc.) written by Warren L. DeLano (unpublished results).

3. Theory

(a) *Combined refinement of macromolecule and solvent*

The structure factor $F_{calc}(\mathbf{h})$ of a macromolecular structure is expressed as:

$$F_{calc}(\mathbf{h}) = F_{macro}(\mathbf{h}) + F_{bound}(\mathbf{h}) + F_{bulk}(\mathbf{h}), \quad (1)$$

where $F_{macro}(\mathbf{h})$ is obtained from the atomic model of the macromolecule, F_{bound} is computed from all bound

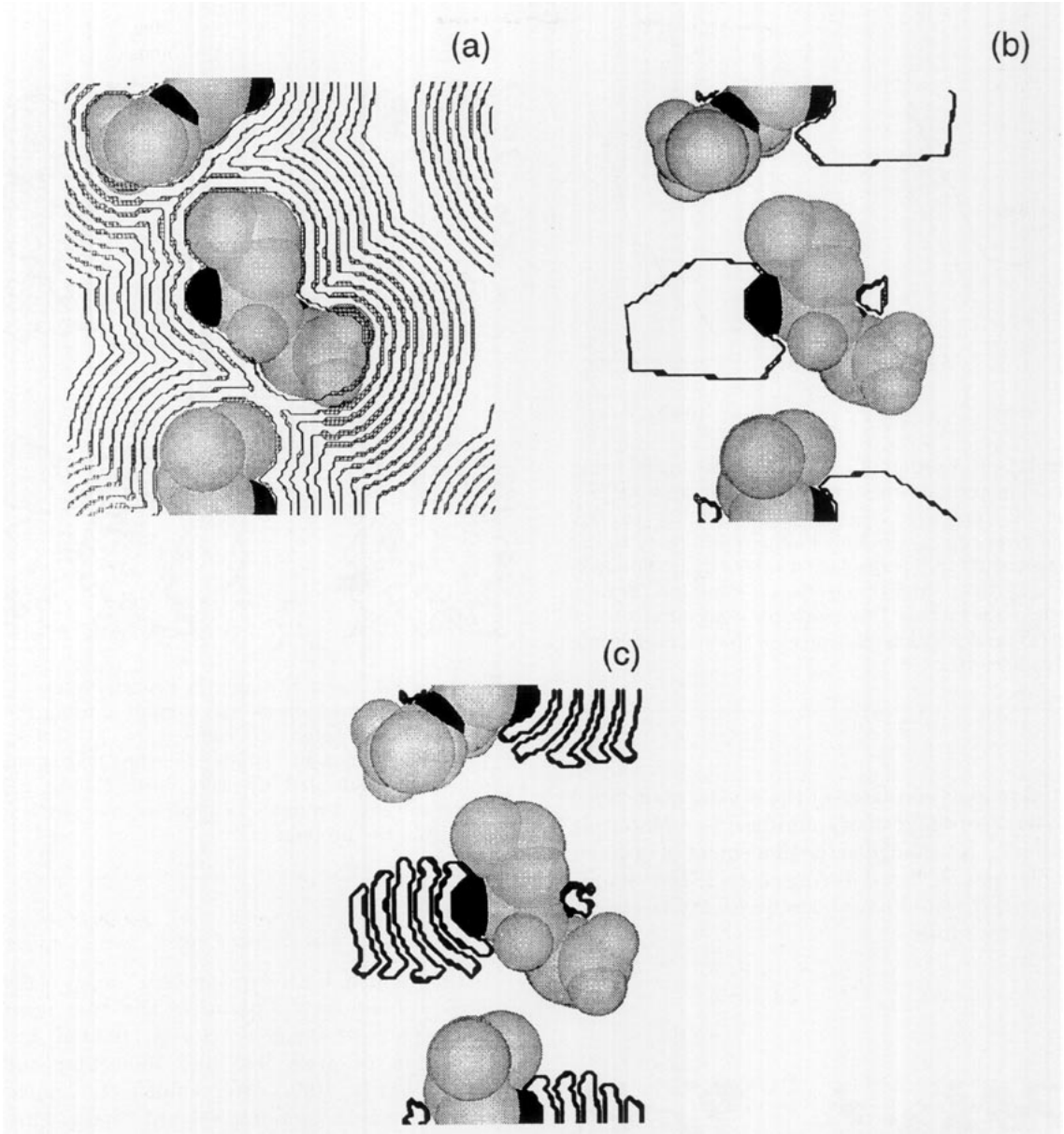

Figure 3. *Partitioning of the solvent volume by proximity to specific solute atoms. The solute consists of a leucine-serine* dipeptide. It is placed in a monoclinic unit cell and $P2_1$ symmetry is imposed. (a) Radical shells surrounding the solute and all its symmetry mates. (b) Partitioning of the available space into regions closest to the nitrogen atoms (in black) and the remaining space. One of the nitrogen atoms is fully solvent exposed while the other is almost completely buried. (c) Fractions of the solvent shells that are closest to the nitrogen atoms. Radial shells and proximity maps are computed up to 5 Å away from the surface of the solute.

water molecules and $F_{\text{bulk}}(\mathbf{h})$ is obtained from an appropriate model for disordered solvent.

We refined parameters of the macromolecular model, of the bound water molecules, and of the disordered solvent in three stages:

Stage 1: The solvent model for disordered solvent was generated in the presence of the macromolecular model and bound water molecules, and its parameters were refined.

Stage 2: The placement of ordered water molecules was checked by using a conservative criterion: the difference density map had to exhibit a peak with several standard deviations above the mean, the peak had to be within hydrogen bonding distance from a polar or charged group on the macromolecular surface or from another water molecule, and the refined thermal factor of the placed water oxygen atom had to be less than 50 Å^2. Water molecules were added or removed if necessary.

Stage 3: Atomic positions and thermal factors for the macromolecular model and the bound waters were refined against the whole resolution range of the observed diffraction data. The bound water molecules stayed in the vicinity of the initial positions during the simulated annealing stage, so it was not necessary to restrain their positions.

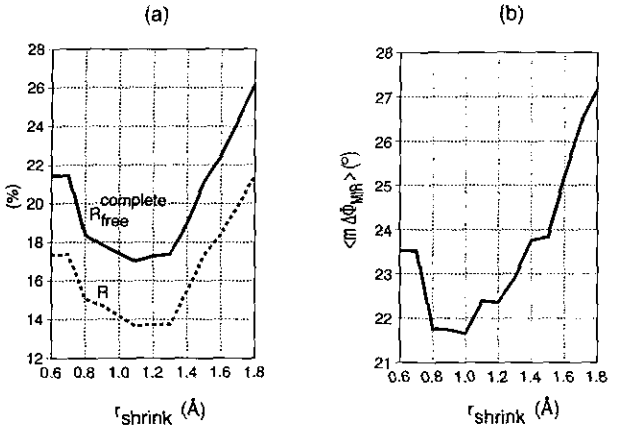

Figure 4. Optimization of r_{shrink} for the flat model using the penicillopepsin diffraction data. (a) R and $R_{free}^{complete}$ obtained by complete cross-validation at 23 to 6 Å resolution *versus* r_{shrink}. (b) Model's phase difference to the MIR phases at 23 to 6 Å resolution *versus* r_{shrink}. For each value of r_{shrink} the ρ_s and B_s parameters equation (2) were refined by alternating least-squares optimization of equation (11) and position refinement of the macromolecule and its bound water.

The initial models consisted of the crystal structures of penicillopepsin and neuraminidase (see Materials and Methods). No change in the placement of ordered water molecules occurred during stage 2. If a change had occurred it would have been necessary to repeat the above procedure.

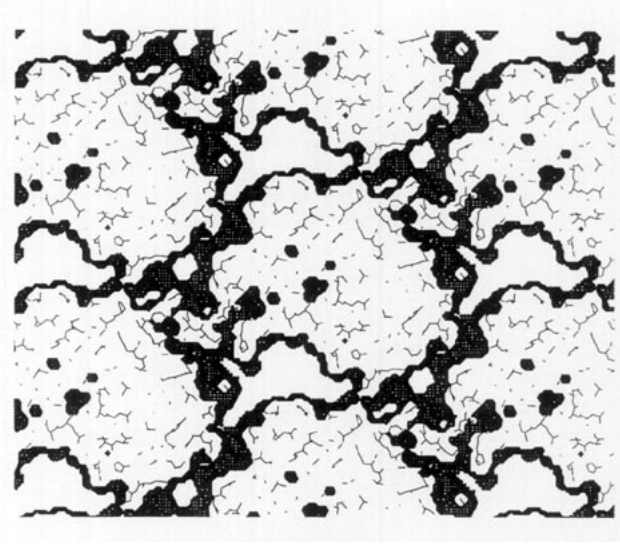

Figure 5. A 1 Å slice through the penicillopepsin crystal structure. The solvent mask (thick lines) was computed using the van der Waals radii listed in Table 1, r_{probe} = 1·0 Å and r_{shrink} = 1·1 Å. Covalent bonds of the penicillopepsin molecules are indicated by thin lines. Empty regions represent solvent. A number of small solvent cavities were found within the protein.

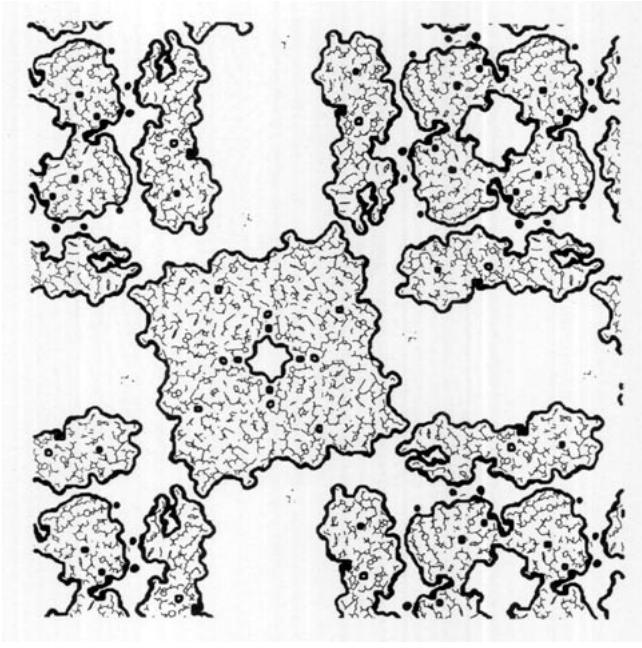

Figure 6. A 1 Å slice through the neuraminidase crystal structure. The solvent mask was computed using the van der Waals radii listed in Table 1, r_{probe} = 1·0 Å and r_{shrink} = 1·1 Å. Covalent bonds of the neuraminidase molecules are indicated by thin lines. Empty regions represent solvent. A number of small solvent cavities were found within the protein.

(b) *Solvent mask*

To distinguish between "protein" and "solvent" regions, a boundary separating the two must be defined. This problem is closely related to the computation of accessible and molecular surface areas (Richards, 1985). We defined the molecular surface and a corresponding solvent "mask" through the following procedure:

Setup: A map M is defined on a grid that covers an asymmetric unit of the crystal. The map values are restricted to 0 and 1. The grid size is chosen to be small enough to avoid Fourier series truncation errors. By trial and error, we found that 1/4 of the high resolution limit is sufficient (0·45 Å in the case of penicillopepsin). Smaller grid sizes did not change the results. All grid points of M are initially set to 1.

Accessible surface: All grid points of M within a distance of r_i around atom i of the atomic model and its symmetry mates are set to 0. The atomic model includes the macromolecule and bound water molecules. r_i is defined as the sum of the van der Waals radius r_{vdw} of atom i and the probe radius r_{probe}. The van der Waals radius is defined as half the distance at which the Lennard-Jones potential energy function reaches its minimum (Table 1).

Contact and re-entrant surface: All grid points of M marked 0 are tested to see if they fall within a distance r_{shrink} from a grid point set to 1. If this is the case, the tested grid point is set to 1. This procedure shrinks the

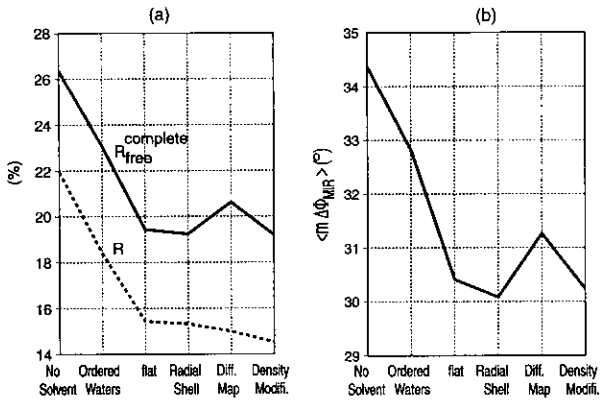

Figure 7. Assessment of the quality of solvent models for the penicillopepsin crystal structure. (a) R and $R_{free}^{complete}$ at 23 to 1·8 Å resolution. (b) $\langle m\Delta\phi_{MIR}\rangle$ (model's phase difference to the MIR phases) at 23 to 2·8 Å resolution. The following cases are shown: no solvent model; inclusion of ordered water molecules; additional modelling of disordered solvent by the flat model, radial shell model, difference map model, and density modification model. The solvent mask parameters were set to $r_{probe} = 1{\cdot}0$ Å and $r_{shrink} = 1{\cdot}1$ Å. Solvent and atomic model parameters were refined as out-lined in the Theory section. For the radial model, N_{shells} was set to 8.

accessible surface area. The resulting boundary between solvent and macromolecule is a combination of contact and re-entrant surface areas (Richards, 1985). The grid points of M marked 1 comprise the solvent regions whereas those marked 0 are associated with the atomic model and its symmetry mates. The resulting map M is referred to as the solvent mask.

The definition of the solvent mask M includes two adjustable parameters, r_{probe} and r_{shrink}. Optimization of these parameters is discussed in the Results and Discussion section.

(c) *Solvent models*

We tested the four solvent models shown schematically in Figure 1.

(i) *Flat model*

The flat model assumes that solvent regions outside the molecular surface show relatively little variation in density as compared with the macromolecule (Phillips, 1980). The structure factor of the solvent (F_{bulk}) is computed by Fourier transformation of the solvent mask M:

$$F_{bulk}(\mathbf{h}) = \rho_s \exp(-B_s \sin^2\theta/\lambda^2)\mathrm{FT}(M), \quad (2)$$

where FT denotes the Fourier transformation. In order to blurr the sharp boundary between macromolecule and solvent as imposed by the mask, resolution-dependent scaling in reciprocal space is applied (represented by the factor in front of the Fourier transformation). The two adjustable parameters ρ_s and B_s are refined as described below.

(ii) *Radial shell model*

The model by Schoenborn (1988) and Cheng & Schoenborn (1990) defines shells of constant electron density at certain distances perpendicular to the molecular surface. The first shell is in contact with the molecular surface of the atomic model and of its symmetry mates. The last shell comprises the space not covered by any other shell or the macromolecule. Each shell requires two solvent parameters, a scattering density ρ_{sn} and a "liquidity" factor B_{sn}. The solvent structure factor is given by:

$$F_{bulk}(\mathbf{h}) = \sum_{n=1,n_{shell}} F_{sn}(\mathbf{h}) \quad (3)$$

$$F_{sn}(\mathbf{h}) = \rho_{sn}\exp(-B_{sn}\sin^2\theta/\lambda^2)\mathrm{FT}(M_n), \quad (4)$$

where M_n is a mask that is set to unity inside shell n and to zero outside. The radial shell model reduces to the flat model if $n_{shell} = 1$.

Adjustable parameters are r_{probes}, r_{shrink}, ρ_{sn}, B_{sn}, the shell thickness d and the number of shells n_{shell}. d must be wider than or equal to the grid size that is used to sample the Fourier transformation in order to obtain meaningful results. We used $d = 0{\cdot}6$ Å and n_{shell} was set to 8 and 16 for penicillopepsin and neuraminidase, respectively. Refinement of ρ_s and B_s is described below.

(iii) *Difference map model*

Badger & Casper (1991) and Badger (1993) proposed a model where an electron density map of the solvent is obtained from a procedure analogous to solvent flattening (Wang, 1985). This model is obtained by iteration of the following steps: (1) calculation of a conventional difference Fourier map with $|F_{obs}| - |F_{calc}|$ coefficients and F_{calc} phases, (2) addition of the difference Fourier map to the phasing model at grid-points within the solvent region, and (3) calculation of a new set of solvent structure factors (F_{bulk}) from this solvent-modified map followed by update of F_{calc} (equation (1)). Step (2) can be carried out in reciprocal space after inverse Fourier transform (FT^{-1}) of the masked difference Fourier map:

$$F_{bulk}^{j+1}(\mathbf{h}) = \mathrm{FT}^{-1}[M(\mathrm{FT}[\Delta F^j(\mathbf{h})]) + \overline{\rho_{macro}}(1-M)], \quad (5)$$

where M is the solvent mask, $\overline{\rho_{macro}}$ is the mean density in the difference map averaged over the macromolecular region and:

$$\Delta F^j(\mathbf{h}) = (|F_{obs}(\mathbf{h})| - |F_{calc}^j(\mathbf{h})|)\exp(i\Phi_{calc}^j(\mathbf{h})). \quad (6)$$

$|F_{calc}^j|$ and Φ_{calc}^j are the calculated amplitudes and phases for the jth cycle, respectively:

$$F_{calc}^j(\mathbf{h}) = F_{macro}(\mathbf{h}) + F_{bound}(\mathbf{h}) + F_{bulk}^j(\mathbf{h}). \quad (7)$$

The iteration is started by setting:

$$F_{calc}^0(\mathbf{h}) = F_{macro}(\mathbf{h}) + F_{bound}(\mathbf{h}). \quad (8)$$

The difference map model has no adjustable parameters except for those of the solvent mask (r_{probe} and r_{shrink}).

(iv) *Density modification model*

Here we propose a modification of the difference map model which uses density modification. In the difference map model, large positive and negative peaks emerge which are related to noise and Fourier series termination errors. As a consequence, the difference map model overfits the diffraction data (see Results and Discussion). Application of both high and low density truncation would hopefully dampen these spurious peaks. This idea is supported by the work of Shiono & Woolfson (1992) who showed that low-density elimination can improve protein phases. As in the case of the flat model, we have also blurred the sharp boundary between the atomic model and the solvent region by resolution-dependent scaling using a thermal factor B_s and a scale factor ρ_s.

$$F^{j+1}_{bulk}(\mathbf{h}) = \rho_s \exp(-B_s \sin^2\theta/\lambda^2) \times \mathrm{FT}^{-1}[M'(\mathrm{FT}[\Delta F^j(\mathbf{h})])], \quad (9)$$

where $\Delta F^j(\mathbf{h})$ has been defined as in equation (6) and M' is a density modification function:

$$M'(r) = \begin{cases} \min(t, \max(b, (\rho(r) - \langle\rho\rangle)/\epsilon)) & \text{grid point } r \text{ outside protein} \\ 0 & \text{grid point } r \text{ inside protein} \end{cases} \quad (10)$$

$\langle\rho\rangle$ is the average density in the solvent region and ϵ is the base level of the solvent density. The success of the density modification model was fairly independent of the precise choice of t, b, ϵ. By trial and error, we achieved good results with ϵ set to the expected average solvent density of pure water ($0{\cdot}33e^-/A^3$), t set to 3 and b set to 1. In the presence of ions the expected solvent density will be higher than $0{\cdot}33e^-/A^3$). This effect is compensated by application of resolution-dependent scaling (ρ_s, B_s which are determined by least-squares optimization against equation (14)). The only other adjustable parameters are those of the solvent mask (r_{probe} and r_{shrink}).

(d) *Refinement of solvent model parameters*

Adjustable parameters $\mathbf{p}$ of the various solvent models were refined by least-squares optimization against the target function:

$$G(\mathbf{p},k) = \sum_{\mathbf{h}} \| F_{obs}(\mathbf{h})| - k|F_{calc}(\mathbf{h},\mathbf{p})\|^2 / \sum_{\mathbf{h}} |F_{obs}(\mathbf{h})|^2, \quad (11)$$

where the overall scale factor k was obtained by the requirement that the first derivative with respect to k of $G(\mathbf{p},k)$ has to vanish resulting in:

$$k = \sum_{\mathbf{h}} |F_{obs}(\mathbf{h})| |F_{calc}(\mathbf{h},\mathbf{p})| / \sum_{\mathbf{h}} |F_{calc}(\mathbf{h},\mathbf{p})|^2. \quad (12)$$

Simultaneous refinement of G against solvent parameters turned out to be ill-behaved even in the simple case that:

$$F_{calc}(\mathbf{h}) = F_{macro}(\mathbf{h}) + F_{bound}(\mathbf{h}) + \rho_s \exp(-B_s \sin^2\theta/\lambda^2)\mathrm{FT}(M), \quad (13)$$

where M is the solvent mask. Straightforward application of least-squares optimization using all observed diffraction data produced numerical instabilities, resulting in a poor fit against the low resolution data (W. I. Weis & A. T. Brünger, unpublished results).

One way to avoid this problem is to make k resolution dependent. We initially computed separate scale factors k_{low}, k_{high} for diffraction data less than and greater than a certain resolution cutoff (5 Å), respectively. Refinement was carried out against:

$$G'(\mathbf{p},k) = G_{low}(\mathbf{p}, k_{low}) + G_{high}(\mathbf{p}, k_{high}), \quad (14)$$

where G_{low} and G_{high} are the target, G, restricted to reflections at resolution lower than and above the cutoff, respectively. The scale factors k_{low}, k_{high} were obtained by equation (12). The solvent parameters $\mathbf{p}$ were determined by least-squares optimization of equation (14). The process was repeated until convergence was achieved at which point k_{low} and k_{high} were approximately equal (illustrated in Table 2 for penicillopepsin using the flat solvent model). Finally, the two scale constants were replaced by a single overall one for the calculation of R values and density maps. This procedure was used for the flat model and the density modification model. For the radial shell model we extended this procedure to refinement of an overall liquidity factor $B_s = B_{sn}$ and refinement of individual ρ_{sn} values. We refined an overall liquidity value as opposed to individual ones because the narrow spacing of the individual shells allows adequate modelling of the electron density distribution by the individual ρ_{sn}.

(e) *Complete cross validation*

The R value is a measure of the fit between the model and the observed diffraction data:

$$R = \frac{\Sigma_{\mathbf{h}} \| F_{obs}(\mathbf{h})| - k|F_{calc}(\mathbf{h})\|}{\Sigma_{\mathbf{h}} |F_{obs}(\mathbf{h})|}, \quad (15)$$

where the F_{obs} denotes the set of observed structure factor amplitudes, and the constant k is a scale factor. The R value is a poor criterion for accuracy as it can be made arbitrarily small by introducing an increasing number of parameters without improving the information content of the model (Brünger, 1992*a*). This problem is avoided by cross-validation; R_{free} shows a higher correlation with the phase accuracy of the model than R (Brünger, 1992*a*, 1993). Cross-validation consists of omitting a certain subset or test set of the observed data, refining the model

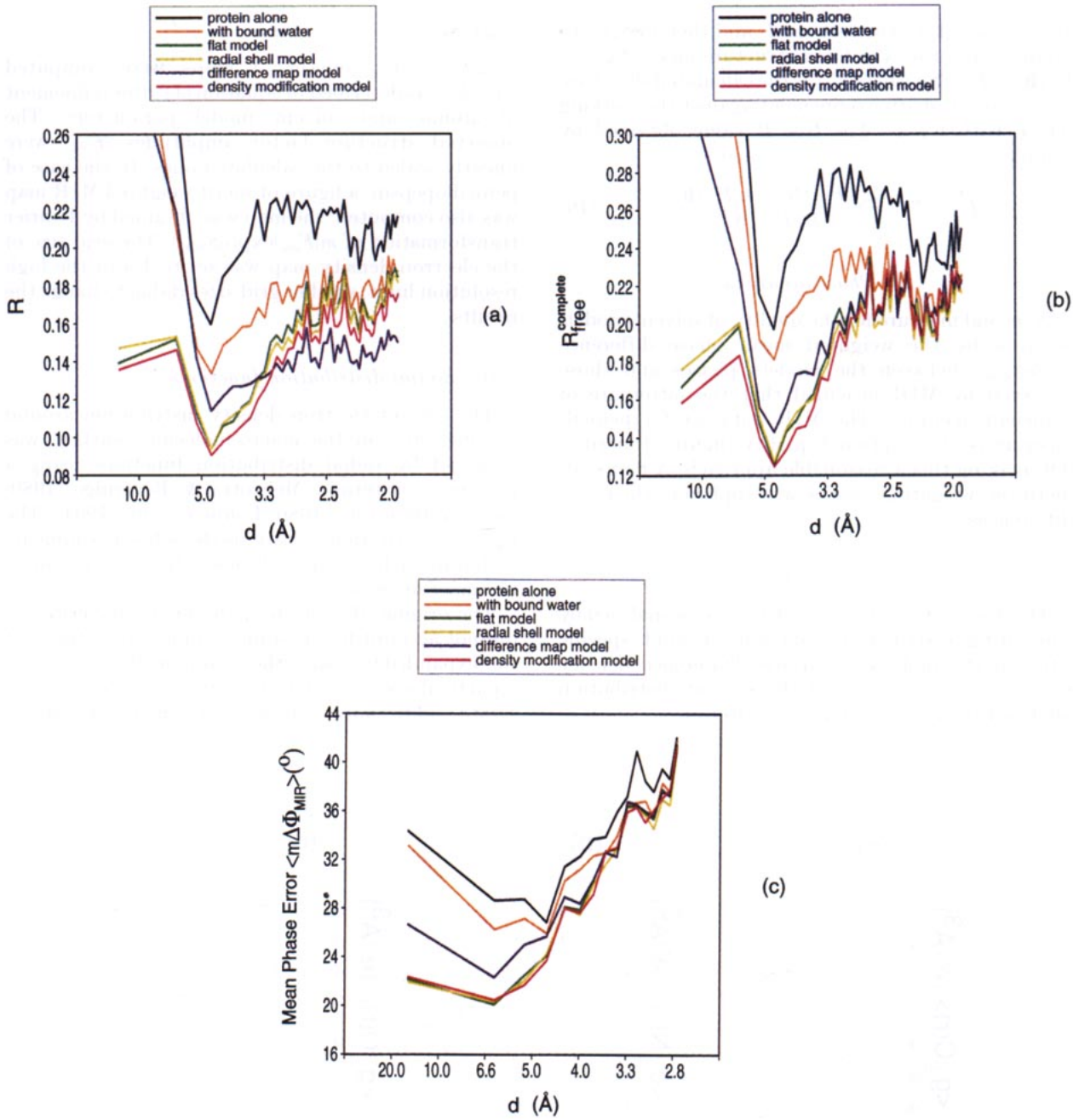

Figure 8. (a) R, (b) $R_{free}^{complete}$ and (c) $\langle m\Delta\phi_{MIR}\rangle$ *versus* resolution, d, for the penicillopepsin crystal structure (note the reciprocal scale for d). R and $R_{free}^{complete}$ were computed at 23 to 1·8 Å resolution while $\langle m\Delta\phi_{MIR}\rangle$ was computed at 23 to 2·8 Å resolution. Black, no solvent model; orange, ordered water molecules included; green, ordered water and flat model; yellow, ordered water and radial shell model; blue, ordered water and difference map model; red, ordered water and density modification model. Parameters are identical to those described in Figure 7.

against the remaining data, and evaluating the R value from the test set.

R_{free} shows little dependence on the choice of the test set as long as the selection is purely random and the test set contains a sufficient number of data points (Brünger, 1993). This is usually not a problem if one is interested in an overall value of R_{free}. However, if the diffraction data are broken down by resolution, R_{free} can show considerable variation at low resolution, i.e. in the region where solvent shows a predominant effect. In the light of this problem, using only a single test set to assess the quality of a solvent model (Badger, 1993) must be viewed with caution.

To increase the reliability of R_{free} at low resolution we used complete cross-validation (compare Brünger *et al.*, 1993). The observed diffraction data set is partitioned into ten test sets ($T_1, \ldots, T_{10}$) where each set contains a different 10% of the data. For each test set T_1, a corresponding working set A_i is defined of all data excluding T_i. Solvent parameter optimizations and protein refinements are carried out ten times, once for each of the working sets A_i. Structure factors

are calculated for the test set T_i and then merged to produce the cross-validated structure factor F_{cv}, i.e. $F_{cv}(\mathbf{h}) = F_{calc}(\mathbf{h})$, where F_{calc} is the calculated structure factor obtained after refinement against the working set A_i with $h \notin A_i$. The free R value obtained by complete cross-validation is defined as:

$$R_{free}^{complete} = \frac{\Sigma_h \| F_{obs}(\mathbf{h}) | - k|F_{cv}(\mathbf{h})\|}{\Sigma_h |F_{obs}(\mathbf{h})|}. \qquad (16)$$

(f) *Phase accuracy*

A second measure for the quality of solvent models is given by the weighted mean phase difference $\langle m\Delta\phi_{MIR} \rangle$ between the model's phases and those observed by MIR provided that the latter are of sufficient accuracy. The MIR data set for penicillopepsin is of exceptional quality (figure of merit is 0·9) making this a reasonable approach. A figure-of-merit (m) weighting scheme was applied to the phase differences.

(g) *Analysis of solvent density*

The distribution of solvent density was analysed by computing distribution functions around specific sites on the molecular surface. Difference density maps were computed and the solvent distribution analysed by using a proximity criterion.

(i) *Electron density maps*

$(2|F_{obs}| - |F_{calc}|)\exp(i\phi_{calc})$ maps were computed with F_{calc} as described in equation (1) after refinement of atomic and solvent model parameters. The observed structure factor amplitudes $|F_{obs}|$ were linearly scaled to the calculated ones. In the case of penicillopepsin, a figure-of-merit weighted MIR map was also computed; the map was obtained by Fourier transformation of $m|F_{obs}|\exp(i\phi_{MIR})$. The grid size of the electron density map was set to 1/4 of the high resolution limit; smaller grid sizes did not change the results.

(ii) *Radial distribution functions*

The solvent electron density distribution around specific sites on the macro-molecule's surface was analysed by radial distribution functions using a proximity criterion (Mehrotra & Beveridge, 1980; Mezei & Beveridge, 1986; Lounnas *et al.*, 1994). The proximity criterion partitions the solvent volume by assigning each solvent molecule to the closest atom of the macromolecule.

The original definition of the proximity criterion did not account for crystallographic symmetry, so it was expanded by taking the minimum distance r from a particular solvent density point to the centers of a solute and to those of its symmetry mates (Figure 2).

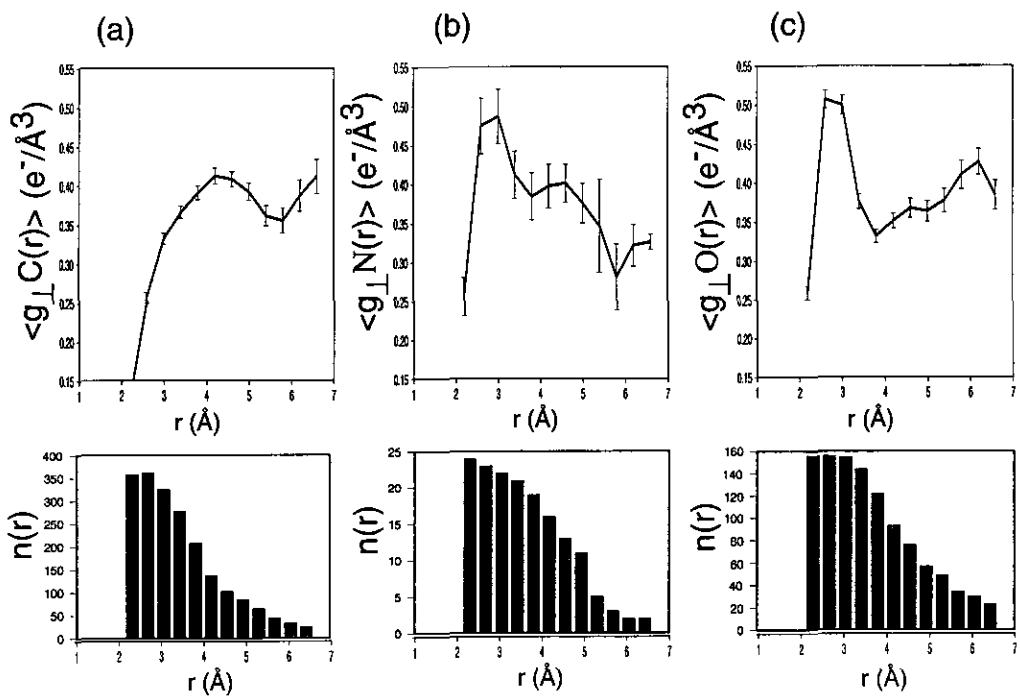

Figure 9. Average solvent distribution functions around specific atoms on the surface of penicillopepsin. (a) $\langle g_\perp C(r)\rangle$, (b) $\langle g_\perp N(r)\rangle$, (c) $\langle g_\perp O(r)\rangle$. Only those surface atoms were considered that had thermal factors less than 24 Å^2. Distribution functions were computed from $(2|F_{obs}| - |F_{calc}|)\exp(i\phi_{calc})$ maps. The computation of F_{calc} included ordered water molecules and the density modification model (eqn (1)). Parameters for the refinement were identical to those described in Figure 7. Error bars indicate the estimated error of the mean for the average distribution functions (eqn (17)). The sample size $n(r)$ for the computation of the average of $g(r)$ at a particular distance r is shown in the corresponding bar plots. Averaging of the distribution functions was carried out in 0·4 Å-wide bins.

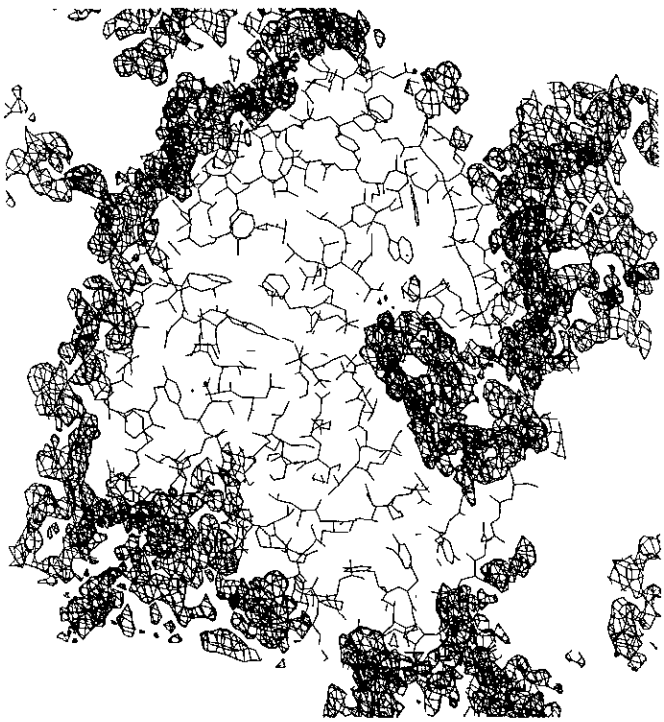

Figure 10. First solvation shell around penicillopepsin. Shown is the electron density within 5 Å around the protein obtained from a 5 Å slice of a $(2|F_{obs}| - |F_{calc}|)\exp(i\phi_{calc})$ map. F_{bulk} was obtained through the density modification model. The contour level is 0·4 $e^-/Å^3$ which corresponds to roughly 0·5 standard deviation above the mean.

The definition of the radial distribution function $g_\perp$ can also be applied to a specific subset of atoms.

The algorithm to compute radial distribution functions is illustrated in Figure 3 for two nitrogen atoms located on the surface of a dipeptide. First, radial shells are computed for the whole dipeptide (Figure 3(a)). Next, a proximity map is defined that indicated for each grid point the atom whose centre is closest to the grid point. The boundary between grid points belonging top the nitrogen atoms and all other points is shown in Figure 3(b). Overlay of the radial shells and the proximity map is used to identify arcs that are used for the computation of $g_\perp N(r)$ (thick lines in Figure 3(c)). The average density in each arc is stored in $g_\perp N(r)$ after division by the volume of the arc. Normally, radial distribution functions are computed in dimensionless units. We elected to compute them in electron density per volume units, instead, in order to indicate the absolute electron density in the crystal.

Weighted averages were computed for the radial distribution functions around selected hydrophobic and hydrophilic sets of atoms. The weight is given by the number of grid points in a particular arc at distance r around each selected atom. The weighted average will be indicated by angular brackets, e.g. $\langle g_\perp N(r)\rangle$ is the weighted average radial distribution function around all nitrogen atoms. The error of the average was estimated by:

$$\epsilon(r) = \sqrt{\langle (g_\perp N(r) - \langle g_\perp N(r)\rangle)^2\rangle / n(r)}, \qquad (17)$$

where $n(r)$ is the sample size that is used to compute $g_\perp N(r)$. $n(r)$ is equal to the number of selected atoms of the macromolecule (in this case nitrogen atoms) that contribute to $g_\perp N(r)$ at distance r. The error of the average $\epsilon(r)$ is related to the root-mean-square (r.m.s.) fluctuation by:

$$\epsilon(r) = \frac{\mathrm{rms}(g_\perp N(r))}{\sqrt{n(r)}}. \qquad (18)$$

(iii) *Estimation of the average density*

In dilute solution, radial distribution functions approach the average density of the solvent as r goes to infinity. One would expect the same to be true for large solvent-filled cavities in macromolecular crystals. Distribution functions that are computed from electron density maps are only known up to an additive constant. This constant is the mean of the density map ($F_{000}/V_{unitcell}$) where $V_{unitcell}$ is the volume of the unit cell. Since the F_{000} term cannot be measured it has to be estimated from a model that fits the diffraction data.

We have estimated the F_{000} term by:

$$F_{000} = (n_e + \rho_s V_{solvent}). \qquad (19)$$

where n_e is the total number of X-ray scattering electrons in the macro molecule and bound water molecules (including those belonging to hydrogens). $V_{solvent}$ is the volume of the solvent, and ρ_s is the solvent electron density which was obtained by optimization of equations (2) and (11).

4. Results and Discussion

(a) *Optimization of solvent model parameters*

Solvent models depend on the definition of the boundary that separates the bulk solvent regions from the rest of the crystal volume. As discussed in the Theory section this boundary is defined through the molecular surface of the macromolecule. The molecule surface is in turn a function of the atomic radii used for the macromolecule and the solvent molecules. Furthermore, the definition of cavities depends on the choice of atomic radii (Kleywegt & Jones, 1994). The radii are associated with the non-bonded potential energy function that describes the van der Waals interaction between pairs of atoms. Possible choices for atomic radii are half the value at which the non-bonded potential energy function assumes a minimum (van der Waals radius) or half the value at which the potential is zero (σ value). Richards (1985) concluded that both sets are physically reasonable and that there may be no set of correct values.

Our solvent mask calculations used the van der Waals radii described in the CHARMM19 force field of Brooks *et al.* (1983). A probe radius corresponding

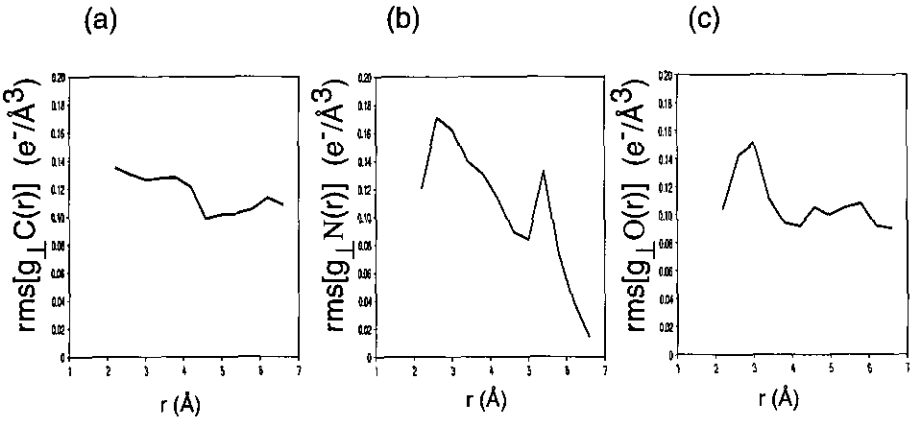

Figure 11. r.m.s. fluctuations (rms($g(r)$) = $\sqrt{(\langle g(r) - \langle g(r)\rangle)^2}$) of the solvent distribution functions around specific atoms on the surface of penicillopepsin. (a) rms($g_\perp C(r)$), (b), rms($g_\perp N(r)$), (c) rms($g_\perp O(r)$).

to the van der Waals radius of water oxygen atoms (1·6 Å) caused certain solvent cavities that contained ordered water molecules to be excluded from the solvent mask. This can be explained by the fact that atoms can get closer to each other than their combined van der Waals distance. By trial and error the probe radius was reduced until all internal water molecules were accounted for in the solvent mask, but small cavities due to packing defects of the protein were not. This resulted in $r_{\text{probe}} = 1$ Å.

The probe radius and the atomic radii define the accessible surface. The construction of the accessible surface described in the Theory section is equivalent to rolling a probe over the macromolecule and defining the accessible surface through the collection of all center positions of the probe. The solvent volume must include the space between the probe center and its surface. We accounted for this volume by reducing the accessible surface to the molecular surface, a combination of contact and re-entrant surface areas (Richards, 1985). We determined the amount of this reduction (r_{shrink}) by minimizing $R_{\text{free}}^{\text{complete}}$ (Figure 4(a)). This yielded $r_{\text{shrink}} = 1{\cdot}1$ Å, which is very close to the value which optimizes the mean phase difference to the MIR phases and the conventional R value (Figure 4(a) and (b)).

The optimum choice of r_{probe} and r_{shrink} produced solvent volumes of 38% and 58% for penicillopepsin and neuraminidase, respectively. The resulting solvent masks are shown in Figures 5 and 6.

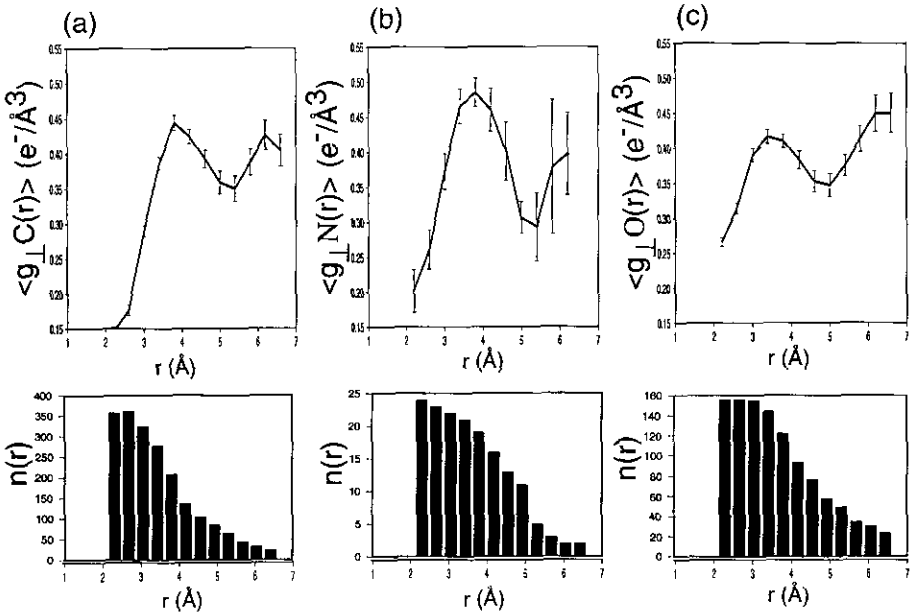

Figure 12. Average solvent distribution functions around specific atoms on the surface of penicillopepsin using figure-of-merit weighted MIR maps ($m|F_{\text{obs}}|\exp(i\phi_{\text{MIR}})$). (a) $\langle g_\perp C(r)\rangle$, (b) $\langle g_\perp N(r)\rangle$, (c) $\langle g_\perp O(r)\rangle$. Error bars indicate the estimated error of the mean for the average distribution functions (equation (17)). The sample size $n(r)$ is shown in the corresponding bar plots.

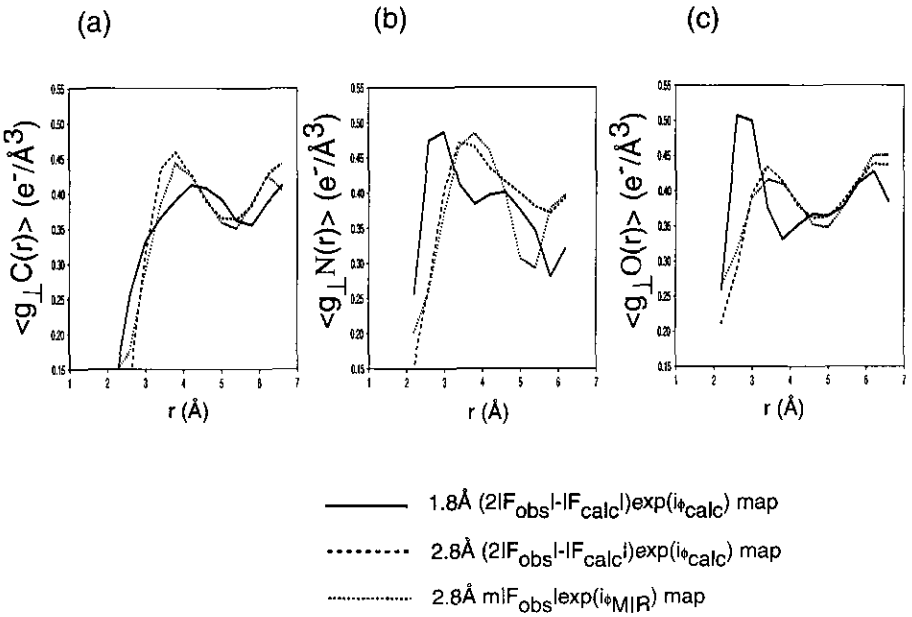

Figure 13. Effect of resolution on average solvent distribution functions around specific atoms on the surface of penicillopepsin. (a) $\langle g_\perp C(r)\rangle$, (b) $\langle g_\perp N(r)\rangle$, (c) $\langle g_\perp O(r)\rangle$. Continuous line, $(2|F_{obs}| - |F_{calc}|)\exp(i\phi_{calc})$ map at 1·8 Å resolution: $(2|F_{obs}| - |F_{calc}|)\exp(i\phi_{calc})$ map at 2·8 Å resolution; dotted line, figure-of-merit weighted MIR maps $(mF_{obs}|\exp(i\phi_{MIR}))$ at 2·8 Å resolution. The grid size set to 0·45 Å in all cases. Averaging of the distribution functions was carried out in 0·4 Å-wide bins.

(b) *Comparison of solvent models in penicillopepsin*

The quality of the various solvent models was assessed by complete cross-validation and by the mean phase difference from the MIR phases. Without any solvent model $R_{free}^{complete} = 26{\cdot}2\%$ and $\langle m\Delta\phi_{MIR}\rangle = 34{\cdot}4°$. Upon inclusion of 134 ordered water molecules, $R_{free}^{complete}$ drops by 1·% accompanied by a 1·5° decrease of $\langle m\Delta\phi_{MIR}\rangle$. If the flat solvent model is added both quantities drop further (1·7% and 2·3°, respectively). The radial shell model and the density modification model produce a slight further improvement (around 0·5% and 0·2°, respectively). The difference map model, however, overfits the diffraction data at high resolution, indicated by a lower R value but a higher free R value compared to the radial shell model. Figure 7 indicates that the flat model produces the most significant improvement of the model's phases and $R_{free}^{complete}$, and that the more sophisticated models produce only a slight further improvement.

The same conclusion can be drawn from the resolution-dependent break-down of $R_{free}^{complete}$ and $\langle m\Delta\phi_{MIR}\rangle$ as shown in Figure 8. The unsatisfactory performance of the difference map model is caused by poorly fitting the low resolution data while overfitting the high resolution data (compare Figure 8(a) and (b)). R is lower than any of the other models at high resolution while $r_{free}^{complete}$ is similar to the other models at high resolution.

Complete cross-validation was essential to produce the $R_{free}^{complete}$ distributions. A single test set produced fairly large fluctuations over several percent of the free R value at low resolution.

(c) *Solvent distribution in penicillopepsin*

Average solvent distributions around carbon, nitrogen, and oxygen atoms on the surface of penicillopepsin are shown in Figure 9 using the density modification model which had the best free R value. The other two models (flat and radial shell) with comparable free R values produced very similar results (not shown). The distribution functions exhibit a pronounced first peak at hydrogen bonding distance (2·8 Å) between water and nitrogen or

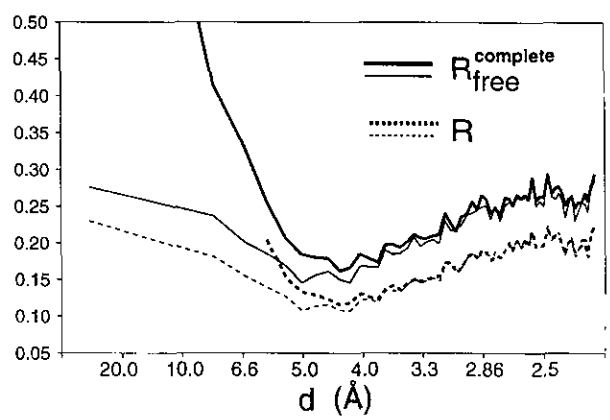

Figure 14. R (broken lines) and $R_{free}^{complete}$ (continuous lines) distributions for neuraminidase *versus d* (reciprocal scale). Thin lines, refined protein, bound water and density modification model. Thick lines, refined protein and bound water without bulk solvent model; in this case, $R_{free}^{complete}$ was computed only between 6 and 2 Å resolution because this was the range that was used in the refinement of the neuraminidase crystal structure.

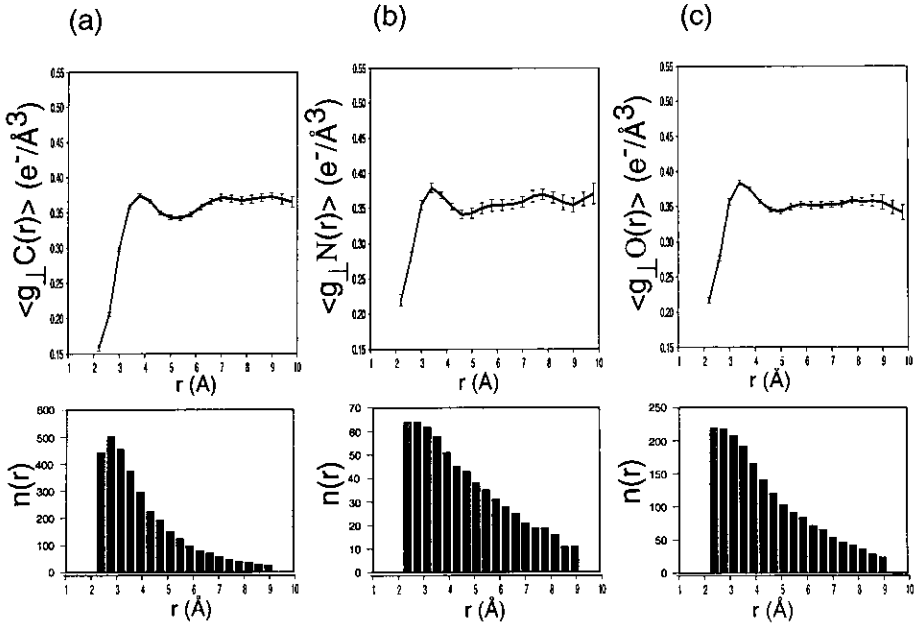

Figure 15. Average solvent distribution functions around specific atoms on the surface of neuraminidase. (a) $\langle g_{\perp}C(r)\rangle$, (b) $\langle g_{\perp}N(r)\rangle$, (c) $\langle g_{\perp}O(r)\rangle$. Only those surface atoms were considered that had thermal factors less than 24 Å^2. Distribution functions were computed from $(2|F_{obs}| - |F_{calc}|)\exp(i\phi_{calc})$ maps. The computation of F_{calc} included ordered water molecules and the density modification model (equation (1)). Parameters for the refinement were identical to those described in Figure 7 except that all calculations were carried out at 76 to 2·3 Å resolution. Error bars indicate the estimated error of the mean for the average distribution functions (equation (17)). The sample size $n(r)$ is shown in the corresponding bar plots. Averaging of the distribution functions was carried out in 0·4 Å-wide bins.

oxygen atoms, and at roughly van der Waals distance (4 Å) between water and carbon atoms. The first solvation layer consists of a highly connected network of solvent density (Figure 10).

One can gain an estimate of the reliability of these distribution functions by computing the estimated error of the mean (indicated by the error bars in Figure 9) and the sample size $n(r)$ of the average. The first peak of the distribution functions is statistically significant while higher order features are probably insignificant because of the small sample size at longer distances.

Some fluctuation of the distribution functions should be expected, as the distribution functions are clearly context, or environment, dependent. The r.m.s. fluctuation of the distribution functions are fairly large (Figure 11).

Whether these fluctuations are due to differences between individual distributions or due to noise cannot be decided based on the available diffraction data. It is interesting to note that the solvent fluctuations around carbon atoms are nearly distance-independent and that the magnitude of the fluctuations is similar for all selected sets of atoms. This observation suggests a strong noise component in the individual distribution functions. Averaging over many sites reduces the influence of the noise (Figure 9).

Another estimate of the reliability of the average radial distribution functions is provided by using a model-free MIR map for the analysis (Figure 12). The peaks of the nitrogen and oxygen atoms are shifted to slightly longer distances while the peak for carbon atoms has become significantly more pronounced. This effect can be explained by the lower resolution (2·8 Å) of the MIR data. Indeed, when computing the radial distribution functions from a $(2|F_{obs}| - |F_{calc}|)\exp(i\phi_{calc})$ map with data restricted to 2·8 Å resolution, close agreement is achieved with those computed from the MIR map (Figure 13). Apparently, lower resolution blurs the distinction between individual hydration sites and solvent density belonging to nitrogen or oxygen atoms "bleeds" into that belonging to carbon atoms.

The broad peak for the solvent distribution function around carbon atoms (Figure 9(a)) is remarkable since bound water molecules in contact with exposed hydrophobic surfaces are rarely seen in protein crystal structures, the water molecules are only well-ordered (high occupancy and low thermal factor) if the protein donates anchor points in suitable hydrogen-bonding positions (Jeffrey & Saenger, 1991). In fact, there are only 179 out of 492 surface carbon atoms close to ordered water molecules and all of these are involved in hydrogen bonds with neighbouring polar or charged groups. On average there appears to be a significant amount of electron density in the first solvation shell around exposed carbon atoms despite the absence of well-defined positions of high occupancy.

Figure 16. First solvation shell around neuraminidase. Shown is the electron density within 5 Å around the protein obtained from a 5 Å slice of a $(2|F_{obs}| - |F_{calc}|)\exp(i\phi_{calc})$ map. F_{bulk} was obtained through the density modification model. The contour level is 0·4 $e^-/Å^3$ which corresponds to roughly 0·5 standard deviation around the mean.

(d) *Solvent distribution in neuraminidase*

The R and $R_{free}^{complete}$ distribution for the neuraminidase crystal structure (Figure 14) indicates improvement of the fit to the diffraction data when using the density modification model, especially at low to medium resolution (76 to 4 Å). The solvent distribution functions for neuraminidase show a similarly pronounced first solvation shell as in penicillopepsin (Figures 15 and 16). There is a significant contribution from disordered water to the first solvation shell because of the relatively small number of observed bound water molecules. A slight shift of the first peaks of the nitrogen and oxygen distribution functions to longer distances is observed. This shift is caused by the lower resolution of the neuraminidase crystal structure compared to that of penicillopepsin (compare Figure 13).

The large solvent pockets in the neuraminidase crystal structure allowed us to compute solvent distribution functions with a distance of more than 12 Å from the protein. No statistically significant features emerged at distances greater than 6 Å (Figure 15(a), (b), (c)). The radial distribution functions that are obtained from difference maps using phases obtained fromthe other solvent nodels (flat, radial shell and difference map model) fall within the error bounds to those shown in Figure 15. Thus, our conclusions are solvent-model independent.

5. Conclusions

The simplest model that assumes nearly flat electron density in the bulk solvent region works suprisingly well. Some of the more elaborate models, such as the radial shell model or the density modification model only slightly improve the fit to the diffraction data. The difference map model overfits the diffraction data. The most pronounced effect on the fit to the diffraction data is seen at low to medium resolution as expected by the disordered character of bulk solvent.

Although the flat solvent model fits the diffraction data quite well the electron density in the solvent region is not completely homogeneous. Fluctuations of solvent density are observed, giving rise to characteristic solvent distributions around polar and non-polar atoms.

Detailed analysis of the solvent features in electron density must be viewed with caution because of the low level of solvent electron density that is observed. Our approach was to use radial averaging over many sites on the protein's surface which will reduce the amount of noise present in the distribution functions. We only considered features that were statistically significant in the radial distribution functions (Figures 9 and 15). Solvent distribution functions showed a well-defined first solvation shell. No statistically significant higher order solvation shells emerged. The observation of a flat average electron density distribution in the bulk solvent region justifies the use of solvent flattening procedures which are commonplace in macromolecular crystallography.

The first solvation shell around hydrophobic groups is fairly pronounced despite the absence of ordered water molecules near most hydrophobic groups. Solution NMR experiments indicate the presence of bound but moving water molecules near methyl groups (Clore *et al.*, 1994). The motion of these water molecules should define a diffuse or broad solvation shell but relatively few well-defined peaks of electron density, in agreement with our observations.

Our analysis tools, to compute site-specific radial distribution functions, can be used to characterize the solvation properties in active sites or binding pockets. This might produce insight into the role of solvent for binding specificities and catalytic mechanisms.

Complete cross-validation allowed the assessment of the quality of solvent models. This method was necessary because the free R value tends to show large fluctuations at low resolution when only a single test set is used. Complete cross-validation could have applications in low-resolution diffraction or imaging techniques where the data to parameter ratio is also critical.

The best solvent models result in R and free R values around 15% and 19·5% for the penicillopepsin crystal structure (Figure 7). This is significantly higher than one might expect from the precision of the intensity data, which is estimated to be a few percent. Models of thermal motion do not resolve this discrepancy (Burling & Brünger, 1994). Thus, it is conceivable that the present models for

solvation and thermal motion are incomplete. Alternatively, the intensity data might be affected by systematic errors of unknown origin. The multi-wavelength anomalous dispersion technique (Hendrickson, 1991) promises to shed some light on these issues. In principle it should allow one to collect phases for all observed reflections to an unprecedented degree of precision and to obtain model-free electron density maps (F. T. Burling, W. I. Weis & A. T. Brünger, unpublished results). High resolution and high-precision data will also be required to observe statistically significant differences between the distribution functions (compare Figure 13).

Molecular dynamics simulations can be tested and improved comparing observed and computed radial distribution functions. Qualitatively, there is reasonable agreement between our observations and molecular dynamics simulations (Komeiji *et al.*, 1993; Steinbach & Brooks, 1993; Lounnas *et al.*, 1992; Lounnas & Pettitt, 1994). High-precision diffraction data collected for many proteins will be needed to further improve the simulation techniques. Ultimately, computer simulations could complement the necessarily incomplete picture provided by X-ray crystallography and solution NMR to provide detailed insights about structural and dynamical features of protein-water interactions.

We thank W. I. Weis for unpublished results, F. T. Burling, W. L. DeLano, P. Gros, L. M. Rice, A. L. U. Roberts and H. W. Wyckoff for fruitful discussions, and A. R. Sieleki, M. N. G. James, J. N. Varghese and P. M. Coleman for providing diffraction data and coordinates. This work was supported by a grant from the National Science Foundation to A.T.B. (DIR 9021975).

References

Allen, M. P. & Tildesley, D. J. (1987). *Computer Simulation of Liquids*. pp. 54–58, Clarendon Press, Oxford.

Badger, J. (1993). Multiple hydration layers in cubic insulin crystals. *Biophys. J.* **65**, 1656–1659.

Badger, J. & Caspar, D. L. D. (1991). Water structure in cubic insulin crystal. *Proc. Nat. Acad. Sci., U.S.A.* **88**, 622–626.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). CHARMM: a program for macromolecular energy minimization, and dynamic calculations. *J. Comp. Chem.* **4**, 187–217.

Brünger, A. T. (1992*a*). Free *R* value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature (London)* **355**, 472–474.

Brünger, A. T. (1992*b*). *X-PLOR Version 3.1. A System for X-ray Crystallography and NMR*. Yale University Press, New Haven, CT.

Brünger, A. T. (1993). Assessment of phase accuracy by cross validation: the free *R* value. Methods and applications. *Acta Crystallogr. sect. D*, **49**, 24–36.

Brünger, A. T., Clore, G. M., Gronenborn, A. M., Saffrich, R. & Nilges, M. (1993). Assessment of the quality of solution nuclear magnetic resonance structures by complete cross-validation. *Science*, **261**, 328–331.

Bruune, R. M., Liepinsh, E., Otting, G., Wüthrich, K. & van Gunsteren, W. F. (1993). Hydration of proteins: a comparison of experimental residence times of water molecules solvating the bovine pancreatic trypsin inhibitor with theoretical model calculations. *J. Mol. Biol.* **231**, 1040–1048.

Burling, F. T. & Brünger, A. T. (1994). Thermal motion and conformational disorder in protein crystal structures: comparison of multi-conformer and time-averaging models. *Israel. J. Chem.* **34**, in the press.

Cheng, X. D. & Schoenborn, B. P. (1990). Hydration in protein crystals. A neutron diffraction analysis of carbonmonoxymyoglobin. *Acta Crystallogr. sect. B*, **46**, 195–208.

Clore, G. M., Bax, A., Wingfield, P. T. & Gronenborn, A. M. (1990). Identification and localization of bound internal water in the solution structure of interleukin 1β by heteronuclear three-dimensional ^{1}H rotating-frame Over-hauser ^{15}N-^{1}H multiple quantum coherence NMR spectroscopy. *Biochemistry*, **29**, 5671–5676.

Clore, G. M., Bax, A., Omichinski, J. G. & Gronenborn, A. M. (1994). Localization of bound water in the solution structure of a complex of the erythroid transcription factor GATA-1 with DNA. *Structure*, **2**, 89–94.

Fraser, R. D. B., MacRae, T. P. & Suzuki, E. (1978). An improved method for calculating the contribution of solvent to the X-ray diffraction pattern of biological molecules. *J. Appl. Crystallogr.* **11**, 696–694.

Hegde, R. S., Grossman, S. R., Laimins, L. A. & Sigler, P. B. (1992). The 1·7 Å structure of the bovine papillomavirus-1 E2 binding domain bound to its DNA target. *Nature (London)*, **359**, 505–512.

Hendrickson, W. A. (1991). Determination of macromolecular structures from anomalous diffraction of synchroton radiation. *Science*, **254**, 51–58.

Hsu, I.-N., Delbare, L. T. J., James, M. N. G. & Hofmann, T. (1977). Penicillopepsin from *Penicillium janthinellum* crystal structure at 2·8 Å and sequence homology with procine pepsin. *Nature (London)*, **266**, 140–145.

James, M. N. G. & Sielecki, A. R. (1983). Structure and refinement of penicillopepsin at 1·8 Å resolution. *J. Mol. Biol.* **163**, 299–361.

Jeffrey, G. A. & Saenger, W. (1991). *Hydrogen Bonding in Biological Structures*. Springer Verlag, Berlin.

Kleywegt, G. J. & Jones, T. A. (1994). Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr. sect. D*, **50**, 178–185.

Komeiji, Y., Uebayasi, M., Someya, J.-I. & Yamato, I. (1993). A molecular dynamics study of solvent behaviour around a protein. *Proteins: Struct. Funct. Genet.* **16**, 268–277.

Kossiakoff, A. A., Sintchak, M. D., Shpungin, J. & Presta, L.G. (1992). Analysis of solvent structure in proteins using neutron D_2-H_2O solvent maps: pattern of primary and secondary hydration of trypsin. *Proteins: Struct. Funct. Genet.* **12**, 223–236.

Kuhn, L. A., Siani, M. A., Pique, M. E., Fisher, C. L., Getzoff, E. D. & Tainer, J. A. (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

Levitt, M. & Park, B. H. (1993). Water: now you see it, now you don't. *Structure*, **1**, 223–226.

Levitt, M. & Sharon, R. (1988). Acurate simulation of protein dynamics in solution. *Proc. Nat. Acad. Sci., U.S.A.* **85**, 7557–7561.

Lounnas, V. & Pettitt, B. M. (1994). A connected-cluster of hydration around myoglobin: correlation between molecular dynamics simulations and experiment. *Proteins: Struct. Funct. Genet.* **18**, 133–147.

Lounnas, V., Pettitt, B. M., Findsen, L. & Subramanian, S. (1992). A microscopic view of protein solvation. *J. Phys. Chem.* **96**, 7157–7159.

Lounnas, V., Pettitt, B. M. & Phillips, G. N., Jr (1994). A global model of the protein-solvent interface. *Biophys. J.* **66**, 601–614.

Matthews, B. W. (1968). Solvent content of protein crystals. *J. Mol. Biol.* **33**, 491–497.

Mehrotra, P. K. & Beveridge, D. L. (1980). Structural analysis of molecular solutions based on quasi-component distribution functions. Application to $((H_2)CO)_{aq}$ at 25°C. *J. Amer. Chem. Soc.* **102**, 4287–4294.

Mezei, M. & Beveridge, D. L. (1986). Structural chemistry of the biomolecular hydration *via* computer simulation: the proximity criterion. *Methods Enymol.* **127**, 21–47.

Moews, P. C. & Kretsinger, R. H. (1975). Refinement of the structure of carp muscle calcium-binding parvalbumin by model building and difference Fourier analysis. *J. Mol. Biol.* **91**, 210–228.

Otting, G., Liepinsh, E. & Wüthrich, K. (1991). Protein hydration in aqueous solution. *Science*, **254**, 974–980.

Otwinowski, Z., Schevitz, R. W., Zhang, R.-G., Lawson, C. L., Jochimiak, A., Marmorstein, R. Q., Luisi, B. F. & Sigler, P. B. (1988). Crystal structure of *trp* repressor/operator complex at atomic resolution. *Nature (London)*, **355**, 321–329.

Parak, F., Hartmann, H., Schmidt, M., Corongiu, G. & Clementi, E. (1992). The hydration shell of myoglobin. *Eur. Biophys. J.* **21**, 313–320.

Phillips, S. E. W. (1980). Structure and refinement of oxymyoglobin at 1·6 Å resolution. *J. Mol. Biol.* **142**, 531–554.

Richards, F. M. (1985). Calculation of molecular volumes and areas for structures of known geometry. *Methods Enzymol.* **115**, 440–464.

Robinson, C. R. & Sligar, S. G. (1994). Hydrostatic pressure reverses osmotic pressure effects on the specificity of *Eco*RI-DNA interaction. *Biochemistry*, **33**, 3787–3793.

Saenger, W. (1987). Structure and dynamics of water surrounding biomolecules. *Annu. Rev. Biophys. Chem.* **16**, 93–114.

Schoenborn, B. P. (1988). Solvent effect in protein crystals. A neutron diffraction analysis of solvent and ion density. *J. Mol. Biol.* **201**, 741–749.

Shakked, Z., Guzekevich-Guerstein, G., Frolow, F., Rabinovih, D., Joachimiak, A. & Sigler, P. B. (1994). Determinants of repressor/operator recognition from the structure of the *trp* operator binding site. *Nature (London)*, **368**, 469–473.

Shiono, M. & Woolfson, M. M. (1992). Direct-space methods in phase extension and phase determination. I. Low-density elimination. *Acta Crystallogr. sect. A*, **48**, 451–456.

Squire, P. G. & Himmel, M. E. (1979). Hydrodynamics and protein hydration. *Arch. Biochem. Biophys.* **196**, 165–177.

Steenivasan, U. & Axelsen, P. H. (1992). Buried water in homologous serine proteases. *Biochemistry*, **31**, 12785–12791.

Steinbach, P. J. & Brooks, B. R. (1993). Protein hydration elucidated by molecular dynamics simulation. *Proc. Nat. Acad. Sci., U.S.A.* **90**, 9135–9139.

Teeter, M. M. (1984). Water structure of a hydrophobic protein at atomic resolution: pentagon rings of water molecules in crystals of crambin. *Proc. Nat. Acad. Sci., U.S.A.* **81**, 6014–6018.

Teeter, M. M. (1991). Water-protein interactions: theory and experiment. *Annu. Rev. Biophys. Chem.* **20**, 577–600.

Thanki, N., Thornton, J. M. & Goodfellow, J. M. (1989). Distributions of water around amino acid residues in proteins. *J. Mol. Biol.* **202**, 637–657.

Tulip, W. R., Varghese, J. N., Baker, A. T., van Donkelaar, A., Laver, W. G., Webster, R. G. & Colman, P. M. (1991). Refined atomic structures of N9 subtype influenza virus neuraminidase and escape mutants. *J. Mol. Biol.* **221**, 487–497.

Venable, R. M. & Pastor, R. W. (1988). Frictional models for stochastic simulations of proteins. *Biopolymers*, **27**, 1001–1014.

Wang, B.-C. (1985). Resolution of phase ambiguity in macromolecular crystallography. *Methods Enzymol.* **115**, 90–112.

Edited by B. W. Matthews

(Received 17 May 1994; accepted 22 July 1994)