18.4.5.4. Ordered solvent water

A protein crystal typically contains ~50% aqueous solvent. This is roughly divided into two separate zones. The first is a set of highly ordered sites close to the surface of the protein. The second, lying remote from the protein surface, is essentially composed of fluid water, with no order between different unit cells.

At room temperature, the solvent sites around the surface are assumed to be in dynamic equilibrium with the surrounding fluid, as for a protein in solution. Nevertheless, the observation of apparently ordered solvent sites on the surface indicates that these are occupied most of the time. The waters are organized in hydrogen-bonded networks, both to the protein and with one another. The most highly ordered water sites lie in the first solvent shell, where at least one contact is made directly to the protein. For the second and subsequent shells the degree of order diminishes: such shells form an intermediate grey level between the ordered protein and the totally disordered fluid. Indeed, the flexible residues on the surface form part of the continuum between a solid and liquid phase.

In the ordered region, the solvent structure can be modelled by discrete sites whose positional parameters and ADPs can be refined. For sites with low ADPs, the refinement is stable and their behaviour is well defined. As the ADPs increase, or more likely the associated occupancy in a particular site falls, the behaviour deteriorates, until finally the existence of the site becomes dubious. There is no hard cutoff for the reality of a weak solvent site. However, the number and significance of solvent sites are increased by atomic resolution data. Despite the fact that the waters contribute only weakly to the high-resolution terms, the improved accuracy of the rest of the structure and the reduction of noise due to the high resolution mean that their positions become better defined.

Indeed, the occupancy of some solvent sites can be refined if the resolution is sufficient, or at least their fractional occupancy can be estimated and kept fixed (Walsh *et al.*, 1998). This leads to the possibility of defining overlapping water networks with alternative hydrogen-bonding schemes. This can be a most time-consuming step in atomic resolution refinement, and a trade-off finally has to be made between the relevance of any improvement in the model and the time spent.

18.4.5.5. Automatic location of water sites

The protein itself has a clearly defined chemical structure, and the number of atoms to be positioned and how they are bonded to one another are known at the start of model building. The solvent region is in marked contrast to this, as the number of ordered water sites is not known *a priori*, and the distances between them are less well defined, their occupancy is uncertain, and there may be overlapping networks of partially occupied solvent sites. Those of low occupancy lie at the level of significance of the Fourier maps.

Selection of partially occupied solvent sites poses a most cumbersome problem in the modelling over and above that of the macromolecule itself, and can be highly subjective and very time consuming. Improved resolution of the data reveals additional weak or partially occupied solvent sites, which generally do not behave well during refinement. Water atoms modelled into relatively weak peaks in electron density tend to drift out of the density during refinement due to the inaccurate gradients that define their positions.

Given the huge number of water sites in question, automatic and at least semi-objective protocols are required. Several procedures have been developed for the automated identification of water sites during refinement [*inter alia ARP* (Lamzin & Wilson, 1997) and *SHELXL* (Sheldrick & Schneider, 1997)] and others allow selective inspection of such sites using graphics [such as *O* (Jones *et al.*, 1991) and *COOT* (Emsley & Cowtan, 2004)]. These depend on a combination of peak height in the density map and geometric considerations. However, these programs are currently optimized for structures at more typical resolutions, and future efforts could be made to adapt them for atomic resolution structures with overlapping water networks and other high levels of detail.

18.4.5.6. Bulk solvent and the low-resolution reflections

As stated in Section 18.4.5.3 and first reviewed by Matthews (1968) and more recently by Andersson & Hovmöller (1998), macromolecular crystals contain substantial regions of totally disordered, or bulk, aqueous solvent, in addition to those solvent molecules bound to the surface. The average electron density of the crystal volume occupied by protein is 1.35 g cm^{-3} (according to Matthews) or 1.22 g cm^{-3} (according to Andersson & Hovmöller), while that of water is 1.0 g cm^{-3}. This is because the atoms are more closely packed within the protein, as they are connected by covalent bonds, while in solvent regions they form sets of hydrogen-bonded networks.

To model both the solvent and protein regions of the crystal appropriately, it is necessary to have a satisfactory representation of the bulk solvent. The high R factors generally observed for most proteins for the low-resolution shells are in part symptomatic of the poor modelling of this feature or of systematic errors in the recording of the intensities of the low-angle reflections. For atomic resolution structures, the R factor can fall to values as low as 6–7% around 3–5 Å resolution. However, in lower-resolution shells it then rises steadily, often reaching values in the range of 20–40% below 10 Å. These observations indicate serious deficiencies in our current models or data.

The poorest approach is to ignore bulk solvent and assign zero electron density to those regions where there are no discrete atomic sites, as this leads to a severe discontinuum. An improved approach is to assign a constant value of the electron density to all points of the Fourier transform that are not covered by the discrete, ordered sites. This provides substantial reduction in the R factor for low-resolution shells of the order of 10% and requires the introduction of only one extra parameter to the least-squares minimization. An improvement of this simplistic model is the introduction of a second parameter, $B_{\rm sol}$, described by

$$\text{scale} = k_0 \exp(-B_0 s^2/4)[1 - k_{\rm sol} \exp(-B_{\rm sol} s^2/4)], \quad (18.4.5.1)$$

where k_0 and B_0 are the scale factors for the protein, and $k_{\rm sol}$ and $B_{\rm sol}$ are the equivalent parameters for the bulk solvent (Tronrud, 1997). In effect, this provides a resolution-dependent smoothing of the interface contribution, rather than an overall term applied equally to all the data. The physical basis of this is discussed by Tronrud and implemented in several programs, for example *SHELXL*, *REFMAC* and *PHENIX.REFINE* (Fig. 18.4.5.2).

Another approach (Jiang & Brunger, 1994) to account for the effect of solvent is as follows: (1) calculate the mask covering the atoms in the crystal; (2) include a constant value for the electron density in the region not covered by the mask; (3) calculate structure factors from the solvent region; and (4) compute a total

structure by applying the appropriate scale to the solvent contribution:

$$F_{\text{total}} = F_{\text{prot}} + k_m \exp(-B_m s^2/4) F_{\text{mask}}, \qquad (18.4.5.2)$$

where k_m and B_m are the scale and temperature factor for the solvent contribution, and F_{total}, F_{prot} and F_{mask} are the complex structure factors corresponding to the combined contribution from the protein and solvent region. The scale and B values are usually refined iteratively to find the best match to the observed structure factors, and are implemented in *CNS*, *REFMAC* and *PHENIX.REFINE*.

Nevertheless, there remain severe problems in the modelling of the interface. The border between the two regions is not abrupt, as there is a smooth and continuous change from the region with fully occupied, discrete sites to one which is truly fluid, but this passes through a volume with an increasing level of dynamic disorder and associated partial occupancy. Modelling of this region poses major problems, as described above, and the definition of disordered sites with low occupancy remains difficult even at atomic resolution. At which stage the occupancy and associated ADP can be defined with confidence is not yet an objective decision. In addition, refinement and modelling at this level of detail is very time consuming in terms of human intervention.

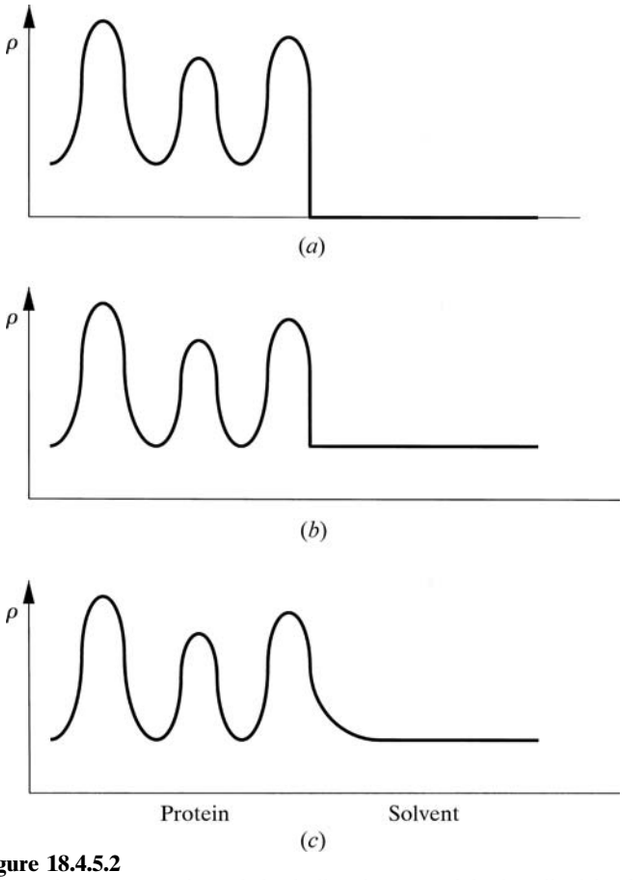

Figure 18.4.5.2
Schematic representation of the bulk-solvent models described in the text. (*a*) No bulk-solvent correction, *i.e.* solvent density set to zero. (*b*) Constant level of solvent outside the macromolecule and ordered water envelope. Here, sharp edge effects remain. (*c*) The model as in (*b*), but smoothed at the edge of a macromolecule, equivalent to the application of a B value to the solvent model.

18.4.5.7. Metal ions and other ligands in the solvent

In general, proteins are crystallized from aqueous solutions which contain various additives, such as anions or cations (especially metals), organic solvents, including those used as cryoprotectants, and other ligands. Some of these may bind in specific or indeed non-specific sites in the ordered solvent shell, in addition to any functional binding sites of the protein. To identify such entities at limited resolution is often impossible, as the range of expected ADPs is large and there is very poor discrimination in the appearance of such sites and of water in the electron density. Atomic resolution assists in resolving ambiguities, as the interatomic distances, ADPs and occupancies are all better defined.

For metal ions, two additional criteria can be invoked. Firstly, the coordination geometry, with well defined bond lengths and angles, provides an indication of the identity of the ion, as different metals have different preferred ligand environments (Harding, 1999, 2006). The bond-valence approach is also applicable (Müller *et al.*, 2003). In addition, the value of the refined ADP and/or occupancy is helpful. Secondly, the anomalous signal in the data should reveal the presence of metal and some other non-water sites in the solvent through computation of the anomalous difference synthesis (Dauter & Dauter, 1999). This emphasizes the need to retain the anomalous signal during the collection and reduction of native data. While these approaches can be applied at lower resolution, they both become much more powerful at atomic resolution.

The presence of bound organic ligands has become especially relevant since the advent of cryogenic freezing. Compounds such as ethylene glycol and glycerol possess a number of functional hydrogen-bonding groups that can attach to sites on the protein in a defined way. Indeed, these may often bind in the active sites of enzymes such as glycosyl hydrolases, where they mimic the hydroxyl groups of the sugar substrate. It is most important to identify such moieties properly, particularly if substrate studies are to be planned successfully.

18.4.5.8. Deformation density

X-ray structures are generally modelled using the spherical-atom approximation for the scattering, which ignores the deviation from sphericity of the outer bonding and lone-pair electrons. Extensive studies over a long period have confirmed that the so-called deformation density, representing deviation from this spherical model, can be determined experimentally using data to very high resolution, usually from 0.8 to 0.5 Å. An excellent review of this field has been provided by Coppens (1997). The observed deviations can be compared with those expected from the available theories of chemical bonding and the densities derived therefrom.

The application of atomic resolution analysis to proteins allowed the observation of the deformation density in macromolecules (Lamzin *et al.*, 1999). Data for two proteins were analysed: crambin (molecular weight 6 kDa) at 0.67 Å resolution and a subtilisin (molecular weight 30 kDa) at 0.9 Å resolution. Significant and interpretable deformation density could not be observed for the individual residues. However, on averaging the density over 40 peptide units for crambin and more than 250 for the subtilisin, the deformation density within the peptide unit was clearly visible and could be related to the expected bonding features in these units. This shows the real power of atomic resolution crystallography, which can reveal features containing no more than 0.2 e Å^{-3}.

Deformation density studies are now being applied to many polypeptides (Jelsch *et al.*, 1998; Koritsanszky *et al.*, 2002; Pichon-