in discreet locations, usually on the protein surface. Water molecules are placed according to a procedure that involves identifying electron density features that are not accounted for by protein. Are they necessarily all waters? Probably not: at very high resolution, some "waters" have been shown to be sodium or ammonium ions. But in the absence of other information, they are interpreted as water, usually based on criteria such as the height of the electron density peak and the position of the putative water molecule to protein atoms with which it can form hydrogen bonds. This does not rule out misidentification, but most ions that might be associated with a protein are sufficiently larger or have a shape or electronic properties that might rule out being a water molecule. Again, the number of water molecules that can be identified as associated with a given model will depend on the quality of the model, the quality of the data, and the resolution of the data. For instance, at 2Å resolution, the number of water molecules expected to be observable is very low and may include only those that are very tightly associated with the protein, often in the active or other functional site. Resolution of 1.8Å is usually required to place a significant number of water molecules on the surface with any precision. At this level of resolution, it is important to look for a certain ratio; namely there should be approximately as many water molecules as residues in the protein. Too many water molecules in a final model may mean that "extraneous" electron density is simply being fitted with water when in fact it may tell a different story. Because the R factor can be viewed as a measure of how much electron density is accounted for, such water molecules can drive an R factor down without adding accuracy to the model. The comparison between the R factor and R_{free} is therefore a good measure to assess the possibility of overfitting or misinterpretation (Figure 2.11, Table 2.2).

The second measure of misinterpretation relies on calculation of the biophysical data for the protein. If a model is well fitted to an electron density map, then it should reflect what we know about the properties of proteins. Those properties include the geometries of amino acids and secondary structures in terms of distances and angles. Atom-to-atom distances and angles for components of proteins are well known from small molecule structures and can be compared with those obtained for the model. The measure given is a root-mean-squared deviation (rmsd) for all such distances and angles. Such angles for the relationships of backbone atoms in specific secondary structures has been analyzed theoretically and is given in terms of allowed and disallowed values in a Ramachandran plot.[18] These can be calculated for the model and compared to the theoretical values (Figure 2.12).

A number of residues may fall outside of these criteria for different reasons. For instance, the values for glycine, because of the absence of a side chain, may fall outside the accepted ranges. Proline may fall outside these ranges because proline may exist in both *cis* and *trans* forms. Occasionally, particularly if the resolution is high enough to allow for a precise interpretation, any residue may fall outside acceptable ranges. When that happens it is worth paying some attention to such a residue, because there is usually a functional reason for it to do so. The final coordinates from the model, together with the data from which it was obtained, can be made available by deposition into the Protein Data Base.[19,20]

Positions of water molecules on the surface of the GCase model at 1.8Å resolution

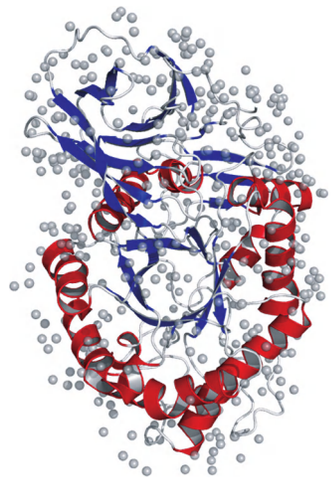

Figure 2.11. Ribbon diagram of the glucocerebrosidase model with the positions of bound water molecules. The surface of a protein model has extraneous electron density that is modeled as water molecules. The gray balls show the positions of such water molecules that have been placed in spherical electron densities on the surface of the protein. An electron density has been interpreted as a water molecule only if the resulting water position (only the oxygen atom is interpretable) is within hydrogen bonding distance of a protein atom that can donate or accept a hydrogen bond. Some of these water molecules can be considered part of the protein structure because they are found in the same position in every structure determination of that protein. Data were taken from PDB code 1OGS.[2]

These measures address the precision of the model. However, accuracy is the most important criterion for the quality of the model, and the only measure of accuracy is the agreement between the model of the protein and the biochemical data for its function. If the model does not explain those data, or at least agrees with them, the model is probably wrong, no matter how precise it may appear in terms of R factors, and so on. Does that mean that all correct models always explain all biochemical data? Certainly not, especially if the configuration of the model seen in the crystal represents only one possible form from a number of possible forms, only one of which is the functionally relevant one for the protein. But in general, the model should make biochemical sense in terms of what is already known about the protein. If it generally does, it is probably accurate (Figure 2.13).

STRENGTHS AND WEAKNESSES OF CRYSTALLOGRAPHY

From the above discussions it is clear that the crystallographic method has strengths and weaknesses. The greatest