Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

How many water molecules can be detected by protein crystallography?

Oliviero Carugo[a,b]* and Domenico Bordo[c]

[a]Structural Biology Program, European Molecular Biology Laboratory, Meyerhofstrasse 1, Postfach 10.22 09, 69117 Heidelberg, Germany, [b]Department of General Chemistry of the University, via Taramelli 12, 27100 Pavia, Italy, and [c]Biostructure Unit, Advanced Biotechnology Center and IST, Largo R. Benzi 10, 16132 Genova, Italy

Correspondence e-mail:
carugo@embl-heidelberg.de

Received 1 July 1998
Accepted 17 September 1998

The number of water molecules which are expected to be experimentally located by protein crystallography was determined by multiple regression analysis on a test set of 873 known protein crystal structures determined at room temperature and on another set of 33 structures determined at low temperature. The dependence of the number of water molecules included in the protein models as a function of a number of significant regressors, such as resolution, fraction of crystal volume occupied by the solvent, number of residues in the asymmetric unit, fraction of apolar protein surface or secondary structure, has been studied. The number of water molecules included in crystallographic models depends primarily on the resolution at which the structure has been solved, while the temperature of the data collection has only marginal influence. On average, at 2.0 Å resolution one water molecule per residue is included in the model, while at 1.0 Å resolution about 1.6–1.7 are crystallographically located. At 2.0 Å resolution the well known rule-of-thumb of 'one water per protein residue' is confirmed, though the number of water molecules experimentally observed is strongly dependent on resolution. The results presented are useful in assessing the quality of a protein crystal structure, in selecting structural results to be compared and in evaluating the expected improvement on the solvent structure when increasing the crystallographic resolution.

1. Introduction

The location of water molecules in protein crystal structures has been the subject of a long-standing debate (see, for example, Levitt & Park, 1993; Frey, 1993), owing to their intrinsic crystallographic relevance and because solvent molecules mediate many key biological processes such as catalysis, protein–protein recognition and protein–substrate interaction. Usually, water molecules are observed forming hydrogen bonds with protein polar atoms. Crystallographic waters, however, are known to fall into fairly shallow energy minima and the activation-energy barrier for their displacement is quite small (Bryant, 1996; Pfeiffer *et al.*, 1998). Various spectroscopic techniques have been used to determine the relaxation times of water molecules at the protein surface (Finney, 1979). The residence time of surface water molecules in BPTI has been evaluated with NMR techniques as being of the order of hundreds of picoseconds (Otting *et al.*, 1991). The observation of water molecules in a calculated electron density requires that, despite the rapid exchange with bulk solvent, a given site keeps enough occupancy during data collection to contribute to the diffraction pattern. Furthermore, the intrinsic errors in the measurements of the diffraction intensities together with imprecise phase determination may result in artifacts in the electron-density map which, in the regions not occupied by covalently bonded protein atoms, may erroneously be interpreted as water molecules. In this regard, despite recent computational advancements in locating and refining the position and the occupancy of solvent molecules at the interface between protein and bulk solvent (Podjarny *et al.*, 1997; Schoenborn *et al.*, 1995), the positioning of water molecules remains largely subject to the crystallographer's personal judgement.

In the present communication, we examine experimentally known data in order to deduce on a statistical basis the expected number of crystallographically located water molecules. Such an analysis can be relevant for assessing the quality of protein crystal structures. In fact, although any assessment of the quality of a protein model based uniquely on stereochemical or statistical criteria cannot be accepted uncritically, it is certainly helpful to have a method to identify statistical outliers, which *per se* suggest a need to verify whether there is any reasonable explanation for the anomaly. The evaluation of the expected number of water molecules is also important in comparing distinct protein models, or to study specific structural features, such as the electrostatics or side-chain conformations. It could be misleading to consider structures with very different treatments of water molecules, in the same way as it is unreasonable to compare structures solved at very different resolution limits.

2. Methods

The data used in this analysis have been obtained from the Protein Data Bank (PDB; Bernstein *et al.*, 1977). To improve the homogeneity of the data, only structures deposited after 1990 and having more than 50 amino-acid residues were considered. The number of water molecules in a model has been defined as the sum of their relative occupancies. To avoid bias in the statistics, since the chemical nature of the heteroatoms (other than water molecules) is highly variable and ranges from small ions such as calcium to bigger cofactors such as NAD or haems, only models having less than 3% heteroatom content other than water were considered. Analogous but less accurate results were obtained with higher threshold values, mostly because of the inclusion of structures with large ligands (RNA, DNA, *etc.*) which could be considered more similar to the protein than to the solvent. The PDB structures were not filtered on the basis of the mutual sequence similarity, since it was assumed that the inclusion of water molecules in the model depends primarily on the crystallization conditions (*e.g.* pH, ionic strength), data-collection procedures and on the strategy adopted during the refinement. Nevertheless, some protein crystal structures highly redundant within the PDB, like lysozymes, constitute only about 2% of the data examined in the present paper. The temperature at which the data were collected was also considered. Structures determined at room temperature, which was assumed if not explicitly stated otherwise in the PDB entry, and structures determined at low temperature (less than 180 K) were considered separately. The statistical analyses presented in this work were performed on a test set of 873 crystal structures determined at room temperature, with a total of 398200 amino acids, 3155751 protein non-H atoms and 192551 water molecules, and on another set of 33 protein structures determined at low temperature, which accounted for a total of 22414 amino acids, 176130 protein non-H atoms and 11682 water molecules. The numbers of both protein and water non-H atoms were computed as the sum of their occupancies. In this regard, the atomic displacement parameters (usually referred to as *B* factors) were not considered, though they are certainly connected to the atomic occupancies, since they are affected by several approximations (see, for example, Ringe & Petsko, 1986; Karplus & Schulz, 1985; Carugo & Argos, 1998; Parthasarathy & Murthy, 1997) and are significantly influenced by the local stereochemistry (Carugo & Argos, 1998), with the consequence that any general relationship between *B* factor and atomic occupancy would be of limited statistical value. Detailed comparisons of protein crystal structures (Finer-Moore *et al.*, 1992; Daopin *et al.*, 1994) pointed out *B*-factor values over which water molecules cannot in general be reliably positioned. Nevertheless, the solvent modelling rests only on reliable electron-density maps and should be justified structurally or functionally or both. Therefore, the data reported in the PDB should be considered, from a statistical point of view, representative of what crystallographers can extrapolate from experimental data.

Amongst the variables that may influence the number of water molecules experimentally located and included in the PDB model, the following were examined: the resolution (R), the fraction of crystal volume occupied by the solvent ($V_{\rm sol}$), the number of residues within the asymmetric unit ($N_{\rm res}$), the fraction of the amino-acid sequence having helical (H), β-strand (E) or other (C) conformation and the fraction of apolar amino acids (Apo) at the protein surface. Only the PDB files explicitly reporting $V_{\rm sol}$ values (computed by the file authors) were considered. Secondary-structure assignments were performed with the program *STRIDE* (Frishman & Argos, 1995) and the residues with α, 3_{10} or π backbone conformation were all considered to be helical, whereas those with β or bridge local geometry were all considered β-strand. Amino acids were considered at the protein surface if their solvent-accessible area, computed with the program *ASC* (Eisenhaber *et al.*, 1995) using a probe radius of 1.25 Å as recommended by Hubbard & Argos (1996), was at least 50% of a reference value. This reference value was obtained by averaging the maximal solvent-accessible area for each residue type for 137 unrelated protein structures selected with *OBSTRUCT* [minimum resolution of 1.8 Å, having a maximum sequence identity of 20% with each other (Heringa *et al.*, 1992)].

The statistical analysis was carried out using software programs developed for this purpose. The number of water molecules ($N_{\rm HOH}$) observed in the asymmetric unit has been normalized to the number of protein atoms ($N_{\rm at}$).

Table 1
Values of the coefficients of the multiple regression model (with estimated standard errors on the last digit shown within parentheses).

Coefficient	Variable	Value
a	Constant	0.33 (3)
b_1	R	−0.069 (8)
b_2	$V_{\rm sol}$	−0.0019 (3)
b_3	$N_{\rm res}$	−0.00031 (8)
b_4	Apo	0.0012 (4)
b_5	E	−0.0009 (2)

3. Results and discussion

Within the room-temperature data set, the observed ratio between water and protein atoms $N_{\rm HOH}/N_{\rm at}$, previously assumed to be a dependent variable, ranged from 0.0 to 0.86. The seven independent variables, R, $V_{\rm sol}$, $N_{\rm res}$, H, E, C, and Apo had well spread and unimodally distributed values, ranging from 1.1 to 3.5, 0.21 to 0.78, 50 to 4420, 0.0 to 0.97, 0 to 0.67, 0.03 to 0.82 and from 0.16 to 0.63, respectively, as shown in Fig. 1.

The dependent variable modestly correlated with each of the regressors. The Pearson linear correlation coefficient (shown in parentheses) decreased in the order R (0.446), $V_{\rm sol}$ (−0.366), $N_{\rm res}$ (−0.226), Apo (−0.198), E (–0.044), C (0.034) and H (0.018), and only the first five were statistically different from 0.0 (at the 90% level of confidence). As expected, fewer water molecules were detected in structures determined at lower resolution. Remarkably, an increase in $V_{\rm sol}$ is accompanied by a diminished number of crystallographic water molecules, which is likely to be because of the lowered crystal rigidity and packing energy. The decrease associated with an increase in the number of amino-acids is because of the relationship between the protein surface (S), to which the number of water molecules can be assumed to be roughly proportional, and the protein volume (V), approximately proportional to the number of animo acids, which is $S \simeq V^{2/3}$. The weak dependence between the number of water molecules and the fraction of polar/apolar surface was unexpected, since polar water molecules interact with polar protein atoms. Neither the secondary structure composition nor the fold type appear to influence the value of $N_{\rm HOH}/N_{\rm at}$.

Multiple regression analysis was performed on the room-temperature test set

$$N_{\rm HOH}/N_{\rm at} = a + \sum b_i x_i,$$

where x_i were the various independent variables described above (R, $V_{\rm sol}$, $N_{\rm res}$, H, E, C, Apo) and a and b_i were the coefficients determined by least-squares methods. The subset of regressor variables sufficient to describe the overall data variance was determined by successively including in the model the independent variable with the higher Pearson linear correlation coefficient. Fig. 2 shows that the multiple correlation coefficient as well as its square value (which measures the proportion of the variability that has been accounted for by the regression equation) increases until the fifth regressor variable is included in the model. Analogously, the mean-square error of the partial regression decreases until the fifth variable is added and remains stable when the last two regressor variables are included into the model. Within this regression model, the first five independent variables, R, $V_{\rm sol}$, $N_{\rm res}$, Apo, and E are, therefore, sufficient to describe the overall variability of the data. The coefficients a and b_i reported in Table 1 are statistically different from 0.0 and can be used to estimate $N_{\rm HOH}/N_{\rm at}$ within a standard error (SE) defined as

$$\begin{aligned}{\rm SE} = \{s[(1/n) + \textstyle\sum\sum p_{ij} \\ \times (x_i - \langle x_i\rangle)(x_j - \langle x_j\rangle)]\}^{1/2},\end{aligned}$$

where x_i and x_j (for i and j from 1 to 5) are the actual values of the regressor variables. The appropriate parameters are given in Table 2.

Most of the multivariate variance is explained by the resolution (R) alone, which displays the largest regression coefficient. The inclusion of the other regressors shows a modest increase (from 0.45 to 0.52) in the multiple regression coefficient. Therefore, it might be convenient to predict the value of $N_{\rm HOH}/N_{\rm at}$ only on the basis of the resolution. For the set of crystal struc-

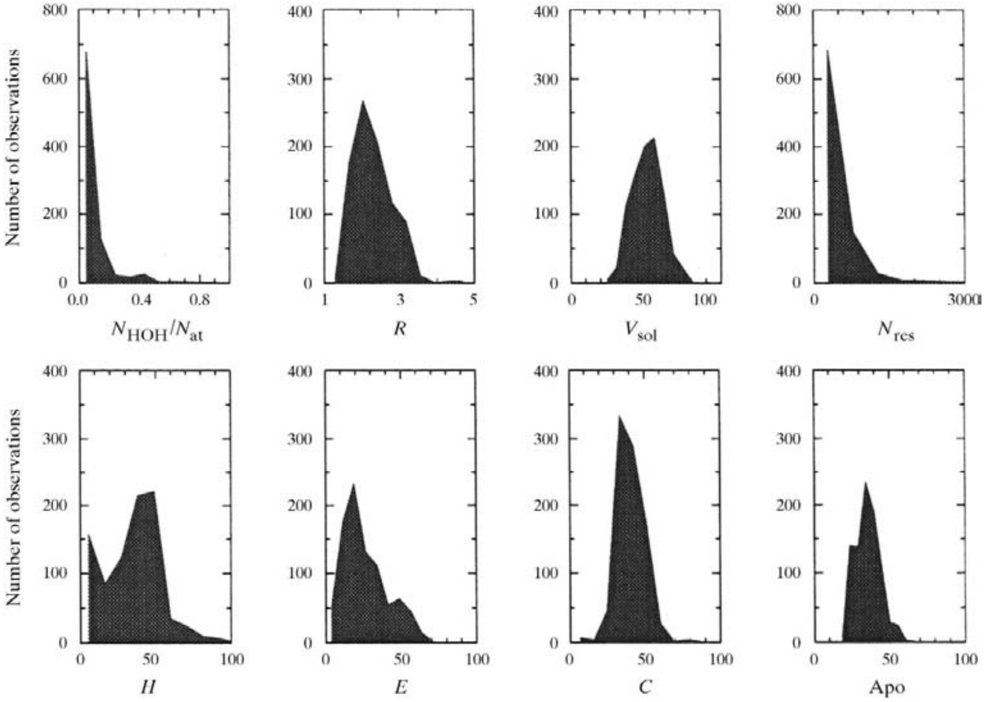

Figure 1
Histograms of the distribution of the examined dependent and independent variables (see text).

Table 2
Values of the parameters needed to compute the standard error (*SE*) of the estimated $N_{\rm HOH}/N_{\rm at}$ values.

Parameter	Meaning				Value
s	Sum of square error on degrees of freedom				0.00782
n	Number of PDB files in the set				873
$\langle x_0 \rangle$	Constant				1
$\langle x_1 \rangle$	Mean value of Res in the PDB files set				2.297
$\langle x_2 \rangle$	Mean value of Vs in the PDB files set				51.095
$\langle x_3 \rangle$	Mean value of Num in the PDB files set				423.152
$\langle x_4 \rangle$	Mean value of Apo in the PDB files set				35.116
$\langle x_5 \rangle$	Mean value of Str in the PDB files set				25.634
p	Inverse normalized matrix $\times 10^6$				
86775.36	−12383.78	−438.46	1.76	−898.32	−155.28
	7323.05	−152.88	−1.27	94.81	22.64
		15.71	−0.04	0.01	0.06
			0.01	−0.02	0.01
				20.41	−1.18
					5.54

tures at room temperature, the useful relationship (Pearson linear correlation coefficient, −0.446) is

$$N_{\rm HOH}/N_{\rm at} = 0.301 - 9.5 \times 10^{-2} R$$

(R is expressed in Å), with an estimated standard error (SE)

$$SE = 92 \times 10^{-3} [1.14 \times 10^{-3} + (R - 2.3)^2 \times 5 \times 10^{-3}]^{1/2}.$$

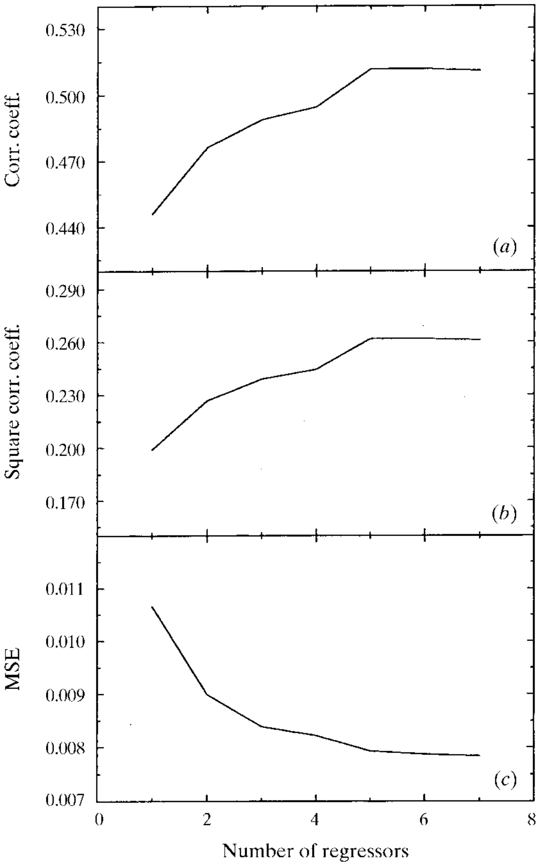

Figure 2
Dependence of the multiple regression coefficient, of its square value and of the mean-square error (MSE) on the number of regressor variables included in the regression model.

The same regression analysis, performed on the set of crystallographic structures determined at low temperature, yielded statistically comparable results. Remarkably, the Pearson correlation coefficient between $N_{\rm HOH}/N_{\rm at}$ and resolution was found to be larger in absolute value (−0.748) than for the room-temperature data, though the limited sampling at low temperature might be the only reason for this. The useful relationship between $N_{\rm HOH}/N_{\rm at}$ and resolution, closely similar to that computed with room-temperature data, is

$$N_{\rm HOH}/N_{\rm at} = 0.334 - 0.11R,$$

with estimated standard error (SE),

$$SE = 43 \times 10^{-3} [30 \times 10^{-3} + (R - 2.2)^2 \times 0.167]^{1/2}.$$

The statistical trend of $N_{\rm HOH}/N_{\rm at}$, illustrated in Fig. 3, shows that no significant differences are observed between the expected values of $N_{\rm HOH}/N_{\rm at}$ at room and at low temperature. The dependence on the resolution is instead rather strong. About 10–12 water molecules can be located per 100 protein atoms at resolution 2.0 Å, which is about double that expected at resolution 1.0 Å. Since the mean number of atoms per residue is between 7 and 8, this implies that slightly less than one water molecule per residue can be crystallographically located in protein models at 2.0 Å resolution, and about 1.6–1.7 water molecules at 1.0 Å resolution. At 2.0 Å resolution, the well known rule-of-thumb of 'one water per

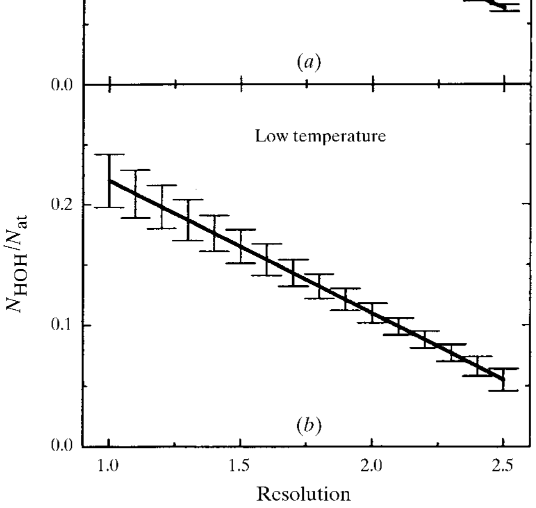

Figure 3
Expected $N_{\rm HOH}/N_{\rm at}$values in function of the resolution values for room-temperature and low-temperature protein crystal structures. Standard errors are displayed as vertical bars.

protein residue' is confirmed, though the number of water molecules experimentally observed is strongly dependent on resolution.

These results support some detailed analysis of the water structure in protein crystals. For example, Earnest *et al.* (1991) (data not used in the regression analysis and extracted from the PDB) observed $N_{\rm HOH}/N_{\rm at}$ ratios of 0.08 and 0.13 in the crystal structures of rat trypsin at room temperature (2.3 Å resolution) and at 120 K (1.59 Å resolution), in agreement with the predicted values of 0.083 (3) and 0.159 (13). It should nevertheless be remembered that the present results, like any other stereochemical and statistical trend, cannot substitute for critical evaluation of the electron-density maps.

The present analysis can have various applications. For example, the simplified statistical model presented here allows the prediction of the expected number of water molecules included, at a given resolution, in a protein structure. Models including an abnormally high (or low) number of crystallographic water molecules should either point to some peculiarity of the protein structure or, alternatively, bring the attention of the crystallographer to the refinement procedure in order to assess its correctness. In particular, the value of $N_{\rm HOH}/N_{\rm at}$ in distinct crystallographic models of the same protein should not deviate more than 2.6SE (at a 99% probability level of confidence) from the expected values. At room temperature, for example, if $R = 2.0$ Å, the expected values for $N_{\rm HOH}/N_{\rm at}$ and SE are 0.111 and 0.004, respectively, and the two structures should therefore have $N_{\rm HOH}/N_{\rm at}$ between 0.101 and 0.121. If the protein contains about 1500 atoms (*i.e.* about 200 residues) this means that there should be 151–181 water molecules. It is also possible to predict the expected increase of information before obtaining suitable crystals by increasing the resolution. For example, by improving the resolution from 2.0 to 1.5 Å, $N_{\rm HOH}/N_{\rm at}$ would increase from 0.111 to 0.159, thus allowing the experimental location of 50% more water molecules.

References

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Bryant, R. G. (1996). *Annu. Rev. Biophys. Biomol. Struct.* **25**, 29–53.

Carugo, O. & Argos, P. (1998). *Proteins Struct. Funct. Genet.* **31**, 201–213.

Daopin, S., Davies, D. R., Schlunegger, M. P. & Gruetter, M. G. (1994). *Acta Cryst.* D**50**, 85–92.

Earnest, T., Fauman, E., Craik, C. S. & Stroud, R. (1991). *Proteins Struct. Funct. Genet.* **10**, 171–187.

Eisenhaber, F., Lijnzaad, P., Argos, P., Sander, C. & Scharf, M. (1995). *J. Comput. Chem.* **16**, 273–284.

Finer-Moore, J. S., Kossiakoff, A. A., Hurley, J. H., Earnest, T. & Stroud, R. M. (1992). *Proteins Struct. Funct. Genet.* **12**, 203–222.

Finney, J. L. (1979). *Water, a Comprehensive Treatise*, Vol. 6, edited by F. Franks, pp. 47–122. New York: Plenum Press.

Frey, M. (1993). *Top. Mol. Struct. Biol.* **17**, 100–146.

Frishman, D. & Argos, P. (1995). *Proteins*, **23**, 566–579.

Heringa, J., Sommerfeldt, H., Higgins, D. & Argos, P. (1992). *CABIOS*, **8**, 599–600.

Hubbard, S. J. & Argos, P. (1996). *Protein Eng.* **10**, 1011–1015.

Karplus, P. A. & Schulz, G. E. (1985). *Naturwissenschaften*, **72**, 212–213.

Levitt, M. & Park, B. H. (1993). *Structure*, **1**, 223–226.

Otting, G., Liepinsh, E. & Wuetricht, K. (1991). *Science*, **254**, 974–980.

Parthasarathy, S. & Murthy, M. R. N. (1997). *Protein Sci.* **6**, 2561–2567.

Pfeiffer, S., Spitzner, N., Lohr, F. & Rueterjans, H. (1998). *J. Biomol. NMR*, **11**, 1–15.

Podjarny, A. D., Howard, E. I., Urzhumtsev, A. & Grigera, J. R. (1997). *Proteins*, **28**, 303–312.

Ringe, D. & Petsko, G. A. (1986). *Methods Enzymol.* **131**, 389–433.

Schoenborn, B. P., Garcia, A. & Knott, R. (1995). *Prog. Biophys. Mol. Biol.* **64**, 105–119.