*Acta Cryst.* (1987). B43, 544-547

# Use of the Occupancy Factor in the Refinement of Solvent Molecules in Protein Crystal Structures

BY CRAIG E. KUNDROT AND FREDERIC M. RICHARDS

*Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, USA*

(*Received* 9 *March* 1987; *accepted* 5 *June* 1987)

## Abstract

A model calculation has been performed to determine whether a variable occupancy factor should be used in the refinement of solvent molecules in protein crystal structures. The atomic structure factor of oxygen was modified by a temperature factor of 0 or 50 $Å^2$ and an occupancy factor of 1·0, 0·75, 0·50 or 0·25, and 'observed' structure factors were calculated for this, the 'target' atom, in a cubic unit cell. The structure factors for a 'model' atom were calculated by keeping the occupancy at a fixed value and by modifying the oxygen structure factor by a temperature factor, $B_m$. A 'best' $B_m$ was selected by minimizing the square of the differences of the target and model structure-factor amplitudes. The agreement between the electron density of the target and best model atoms is good in all cases except when the target-atom temperature factor was 0 $Å^2$ and the nominal resolution was 1 Å. This agreement suggests that, for data limited to a nominal resolution of not better than 1 Å, it is not appropriate to vary both occupancy and temperature factor for solvent molecules in protein structure refinements. Stereochemical considerations suggest that a fixed occupancy of less than 1·0 (*e.g.* 0·50) is likely to maximize the electron density fit.

## Introduction

The refinement of high-resolution protein crystal structures requires modelling of ordered solvent. Protein crystals are grown from supersaturated aqueous solutions and 30–60% of the crystal is typically solvent which is contained in large connected channels (Matthews, 1968). Some of the solvent near the protein surface is sufficiently well ordered to produce peaks in the electron density map. Many of these peaks occur in positions where a solvent molecule must act simultaneously as hydrogen-bond donor and acceptor, or in positions which remain the same under different crystallization conditions (Blake, Pulford & Artymiuk, 1983). These peaks are generally ascribed to water molecules. Occasionally, very high density or contiguous density in a particular geometry will suggest the presence of a solute molecule, *e.g.* a salt ion or an ethanol molecule *etc.* The vast majority of solvent peaks, however, are ascribed to water molecules.

The structure-factor amplitude of an ordered water molecule is usually taken to be the structure factor of atomic oxygen, modified by an isotropic temperature factor, $B$. This assumption is justified by the fact that the electron density in the vicinity of the oxygen nucleus is 600 times the density near the hydrogen nuclei (Kern & Karplus, 1972). Since the water molecules may not always be present at a given location a variable occupancy factor, $Q$, may also be used.

The inclusion of the occupancy factor has been the subject of much debate. The issue is whether this factor refines to a physically meaningful value. Most of the discussion has been concerned with the correlation between the final refined values of $B$ and $Q$ (Watenpaugh, Margulis, Sieker & Jensen, 1978; Sielecki & James, 1981), or with convergence of the refinement process to unique values of $B$ and $Q$ (Hendrickson, 1985). No consensus has emerged on when to include a variable occupancy factor and decisions are made on a case-by-case basis.

This note examines the problem from a different point of view. If the purpose of the refinement is to reproduce the observed electron density as well as possible, then one may ask, 'How different are the calculated electron densities of models with and without a variable occupancy factor?' The electron density distributions produced by two models for a water molecule are compared. In one model the occupancy and the temperature factor are both variable. In the other model, the temperature factor is variable but the occupancy is set at a fixed value. For a specified temperature factor and occupancy factor in the two-parameter model, the temperature factor of the one-parameter model is adjusted to give the best fit between the two sets of calculated reciprocal-space data. If the resulting electron density profiles are substantially different, the added flexibility provided by the second adjustable parameter in the two-parameter model is warranted. If the electron densities of the two models are very similar, however, the occupancy variable cannot be uniquely defined and should be fixed at a predetermined value.

0108-7681/87/060544-04$01.50 

## Calculations

The two-parameter model is referred to as the 'target atom', and its functions are identified by the subscript $t$. The structure-factor amplitude $F_t(s)$ at $s = (\sin\theta)/\lambda$ is defined as

$$F_t(s) = Q_t F_o(s) \exp(-B_t s^2), \quad (1)$$

where $Q_t$ is the occupancy factor, $F_o(s)$ is the structure-factor amplitude of oxygen at $s$ (Cromer & Waber, 1974), and $B_t$ is the temperature factor. The functions for the one-parameter model, the 'model atom', are identified by the subscript $m$. The structure factor for this atom is defined as

$$F_m(s) = Q_m F_o(s) \exp(-B_m s^2), \quad (2)$$

where $Q_m$ has been set equal to a fixed value.

The task is to find the best value of $B_m$ for the model atom, given particular values of $B_t$ and $Q_t$ for the target atom. In a structure determination the value of $B_m$ will be fixed by the refinement procedure. In this study, therefore, the 'best' $B_m$ has been defined as that value which minimizes

$$Y = \sum_h wt(h)\{F_t[s(h)] - F_m[s(h)]\}^2, \quad (3)$$

where $wt(h)$ is the weighting factor for reflection $h$ and the sum ranges over all observed reflections. $Y$ is the function minimized in programs such as the Hendrickson & Konnert restrained least-squares refinement (see Hendrickson, 1985). A different criterion, such as minimizing the square of the difference of the electron densities, produces a different 'best' $B_m$ (Table 1).

To calculate the structure-factor amplitude at discrete points in reciprocal space, the atom was located at the origin in a unit cell with dimensions 50 × 50 × 50 Å (changing the value of the unit-cell dimensions does not affect the results). The lower resolution limit of the data was taken to be 10 Å while separate calculations were made for high-resolution limits of 2 and 1 Å. Equation (1) was evaluated for values of $h$ occurring along one reciprocal-lattice axis and $wt(h)$ was taken to be $h^2$ so that the $(F_t - F_m)^2$ terms were weighted according to the number of reciprocal-lattice points occurring at that resolution. No artificial error terms were added to the $F_t$ or $F_m$. Babinet's principle (Langridge, Marvin, Seeds, Wilson, Hooper, Wilkins & Hamilton, 1960) is sometimes used to include the scatter of the disordered solvent by modifying the atomic structure factors of all ordered atoms in the unit cell. The modified structure factor, $f'(s)$, is related to the unmodified structure factor, $f(s)$, by the equation

$$f'(s) = f(s) - v\rho \exp[-(4\pi v^{2/3} + B_b)s^2], \quad (4)$$

where $v$ is the volume of the atom, $\rho$ is the electron density of the solvent and $B_b$ is an artificial temperature factor introduced to smooth the protein–solvent boundary. Modifying the atomic structure factor of oxygen according to (4) (with $\rho = 0.30$ e Å$^{-3}$, $B_b = 200$ Å$^2$) generally changed the value of the best $B_m$ by 0·5 Å$^2$ and the value of $Y$ by less than 5%.

Table 1. *Values of 'best' $B_m$ in a single-parameter solvent model for different refinement criteria*

The calculations included data from 10 to 2 Å nominal resolution. Model atom occupancy = 1·0. Target atom temperature factor = 50 Å$^2$.

| Minimization criterion | Target atom occupancy ($Q_t$) 0·75 | 0·50 | 0·25 |
|---|---|---|---|
| $\int(\rho_t - \rho_m)^2\,dx$* | 67 | 108 | 250 |
| $\sum_h wt(h)\{F_t[s(h)] - F_m[s(h)]\}^2$ | 66 | 104 | 272 |
| $\left\{\sum_h \lvert F_t[s(h)] - F_m[s(h)]\rvert\right\} \Big/ \left\{\sum_h F_t[s(h)]\right\}$ | 60 | 77 | 146 |

* The integral is evaluated over the range 0 to 8 Å, with the atoms placed at the origin.

In refinements where a variable occupancy is included, model water molecules with temperature factors greater than 50 Å$^2$ or occupancy factors less than 0·3 are usually removed from the model (Hendrickson, 1985). Therefore, calculations were made for $B_t = 0$ and 50 Å$^2$ and $Q_t = 1.00$, 0·75, 0·50 and 0·25. The $B_t = 0$ Å$^2$ case is included as a limiting case.

## Results and discussion

The results of the calculations are summarized in Table 2. The real-space agreement is measured as

$$Z = \sum_x x^2[\rho_t(x, 0, 0) - \rho_m(x, 0, 0)]^2 \Big/ \sum_x x^2 \rho_t(x, 0, 0)^2, \quad (5)$$

where $\rho_t(x, 0, 0)$ and $\rho_m(x, 0, 0)$ are the target and model electron densities at $x = x$, $y = 0$, $z = 0$ and $x$ is sampled every 0·1 Å. The electron densities of the target and model atoms for some values of the parameters are plotted in Figs. 1 and 2.

The following trends are evident. The value of $B_m$ decreases as the high-resolution limit, $d_{min}$, decreases because the density of the target peak narrows with improved resolution. $Y$ increases as $d_{min}$ decreases because the discrepancy between $F_t$ and $F_m$ increases as the resolution range is extended. This is easily seen if $\ln[F_t(s)/F_m(s)]$ is plotted against $s^2$. The result is a straight line with a slope of $-(B_t - B_m)$ and intercept of $\ln(Q_t/Q_m)$. The real-space agreement decreases with increased $d_{min}$, as one would expect.

The electron-density plots suggest to us that in the most physically reasonable cases the two models produce very similar density, *i.e.* the differences would not produce significant peaks in a difference Fourier map. A significant peak is defined here as a peak

Table 2. *Occupancy of target molecule*

| | | 1·00 | | | 0·75 | | | 0·50 | | | 0·25 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $B_t$ (Å$^2$) | $d_{min}$ (Å) | Best $B_m$ (Å$^2$) | $Y$ | $Z$ | Best $B_m$ (Å$^2$) | $Y$ | $Z$ | Best $B_m$ (Å$^2$) | $Y$ | $Z$ | Best $B_m$ (Å$^2$) | $Y$ | $Z$ |
| $Q_m = 1·0$ | | | | | | | | | | | | | |
| 0 | 2·0 | 0 | 0·000 | 0·000 | 7 | 0·139 | 0·018 | 19 | 0·372 | 0·125 | 61 | 0·832 | 0·547 |
| 0 | 1·0 | 0 | 0·000 | 0·000 | 2 | 0·169 | 0·028 | 6 | 0·453 | 0·196 | 40 | 0·933 | 0·730 |
| 50 | 2·0 | 50 | 0·000 | 0·000 | 66 | 0·178 | 0·023 | 104 | 0·439 | 0·133 | 272 | 0·775 | 0·478 |
| 50 | 1·0 | 50 | 0·000 | 0·000 | 66 | 0·180 | 0·024 | 103 | 0·440 | 0·134 | 272 | 0·776 | 0·480 |
| $Q_m = 0·5$ | | | | | | | | | | | | | |
| 0 | 2·0 | −15 | 0·260 | 0·064 | −9 | 0·163 | 0·026 | 0 | 0·000 | 0·000 | 19 | 0·372 | 0·125 |
| 0 | 1·0 | −4 | 0·306 | 0·093 | −2 | 0·202 | 0·042 | 0 | 0·000 | 0·000 | 6 | 0·453 | 0·196 |
| 50 | 2·0 | 25 | 0·354 | 0·105 | 34 | 0·222 | 0·041 | 50 | 0·000 | 0·000 | 104 | 0·439 | 0·133 |
| 50 | 1·0 | 28 | 0·367 | 0·115 | 35 | 0·228 | 0·043 | 50 | 0·000 | 0·000 | 103 | 0·440 | 0·134 |

The functions $Y$ and $Z$ are defined in the text by equations (3) and (5), respectively.

whose height is greater than twice the standard deviation of the difference map. The differences are only significant in the limiting case of high resolution ($d_{min} = 1$ Å) and low temperature factor ($B = 0$ Å$^2$). The differences observed in the model calculations are exaggerated in the sense that the model employs higher-quality data (*i.e.* errorless) and a higher number of observations per variable than does a normal protein refinement. Therefore, the two-parameter model should only be appropriate in highest-resolution studies of very well ordered solvent structures. In the more common cases (*e.g.* $d_{min} \sim 2$ Å), the equivalency of the two models means that the variable occupancy in the two-parameter model is not uniquely or well defined and should not be used. The ability of the single-parameter model, with $Q_m = 1·0$ or 0·5, to fit the target density suggests that it may not even be meaningful to conduct a refinement with the possible values of $Q_m$ limited to just two values.

Since the occupancy factor in the one-parameter model is not uniquely defined, it can be set to any constant value for the purposes of refinement. Stereochemical considerations may be used to select a value for $Q_m$. Recent crystallographic studies (Savage, 1986; Smith, Hendrickson, Honzatko & Sheriff, 1986) provide evidence for the existence of separate water networks, each of which alone makes good stereochemical sense, but which overlap one another so that they must be mutually exclusive at any instant in time. This suggests that a non-unit occupancy factor would be appropriate, even in lower-resolution studies where mutually exclusive networks may not be apparent.

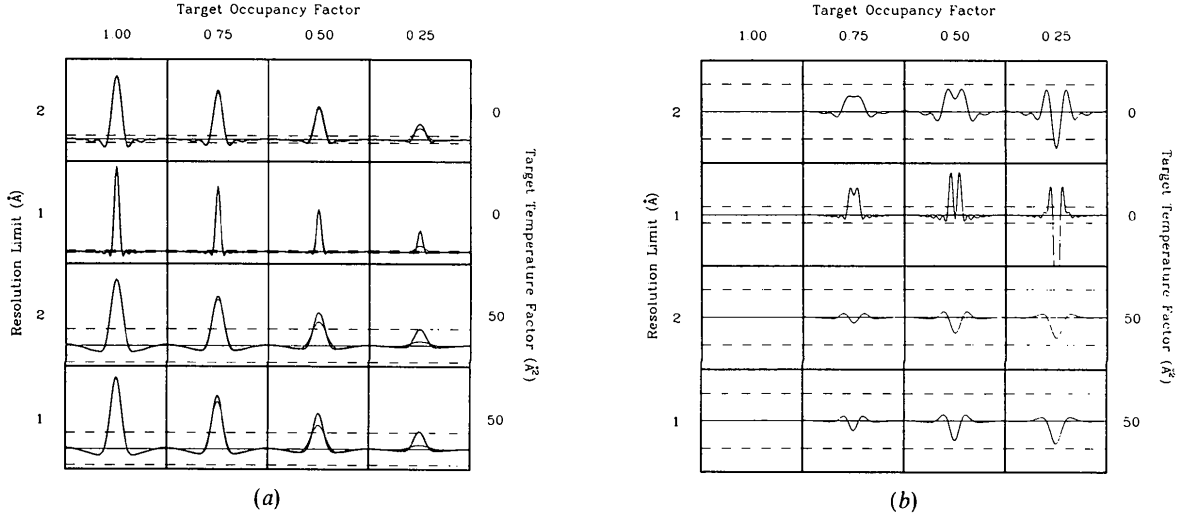


Fig. 1. Target and model electron densities for unit occupancy, $Q_m = 1$. The abcissa in all panels is the $x$ coordinate which varies from −8 to +8 Å. The ordinate in all panels is proportional to electron density. Different ordinate scales are used in each row of panels. The solid horizontal line indicates 0 e Å$^{-3}$ and the dashed horizontal lines indicate ±0·16 e Å$^{-3}$, *i.e.* twice the standard deviation of a typical protein difference Fourier map. (*a*) $\rho_t(x, 0, 0)$, thick line, and $\rho_m(x, 0, 0)$, thin line; (*b*) same data as (*a*) but the ordinate is $\Delta\rho(x, 0, 0) = \rho_m(x, 0, 0) - \rho_t(x, 0, 0)$.

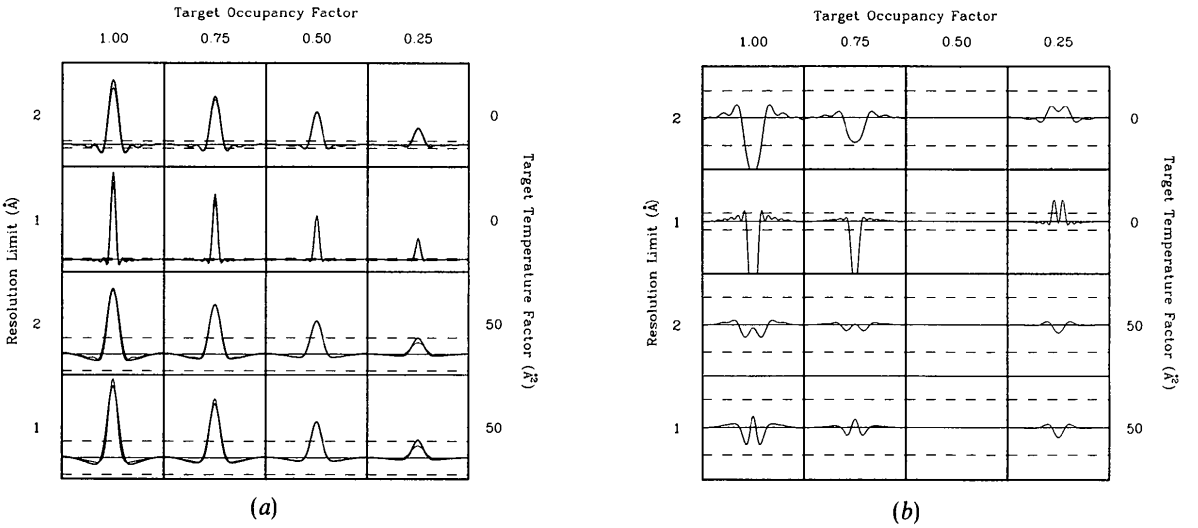


Fig. 2. Target and model electron densities for $Q_m = 0{\cdot}5$. See the legend of Fig. 1 for a description of the ordinates and abcissas. (*a*) $\rho_t(x, 0, 0)$, thick line and $\rho_m(x, 0, 0)$, thin line; (*b*) $\Delta\rho(x, 0, 0) = \rho_m(x, 0, 0) - \rho_t(t, 0, 0)$.

Fig. 2 and Table 1 show that a $Q_m = 0{\cdot}5$ can also be used to fit the target electron density. The assumption of $Q = 0{\cdot}5$, or slightly higher, should approximate most physically reasonable cases and produce satisfactory calculated electron densities.

Lastly, we note that the refinement of the solvent-molecule parameters is analogous to the refinement of heavy-atom parameters in the isomorphous replacement method. The heavy-atom refinements often involve data with $d_{min} > 2$ Å, and it is not possible, from a single difference data set, to obtain well defined parameter values if both $B$ and $Q$ are variable (Dickerson, Weinzierl & Palmer, 1968). The heavy-atom problem is frequently improved by using multiple data sets where the occupancy is changed by varying the free-ligand concentration. This option is not available for water sites.

We thank M. Bannon for help in preparing the manuscript. We particularly thank Dr Adrian Goldman, Dr Mark Sanderson and Dr Hal Wyckoff for helpful discussions on this manuscript and the basic problem. This work was supported by a grant from the Institute of General Medical Sciences to FMR, GM-22778.

## References

BLAKE, C. C. F., PULFORD, W. C. A. & ARTYMIUK, P. J. (1983). *J. Mol. Biol.* **167**, 693-723.

CROMER, D. T. & WABER, J. T. (1974). *International Tables for X-ray Crystallography*, Vol. IV, pp. 71-151. Birmingham: Kynoch Press. (Present distributor D. Reidel, Dordrecht.)

DICKERSON, R. E., WEINZIERL, J. E. & PALMER, R. A. (1968). *Acta Cryst.* B**24**, 997-1003.

HENDRICKSON, W. A. (1985). *Methods Enzymol.* **115**, 252-270.

KERN, C. W. & KARPLUS, M. (1972). *Water: A Comprehensive Treatise*, edited by F. FRANKS, Vol. 1, pp. 21-91. New York: Plenum.

LANGRIDGE, R., MARVIN, D. A., SEEDS, W. E., WILSON, H. R., HOOPER, C. W., WILKINS, M. H. F. & HAMILTON, L. D. (1960). *J. Mol. Biol.* **181**, 423-447.

MATTHEWS, B. W. (1968). *J. Mol. Biol.* **33**, 491-497.

SAVAGE, H. (1986). *Biophys. J.* **50**, 947-965.

SIELECKI, A. R. & JAMES, M. N. G. (1981). *Refinement of Protein Structures*, edited by P. A. MACHIN, J. W. CAMPBELL & M. ELDER, pp. 78-87. Daresbury Laboratory, Warrington: Science and Engineering Research Council.

SMITH, J., HENDRICKSON, W. A., HONZATKO, R. B. & SHERIFF, S. (1986). *Biochemistry*, **25**, 5018-5027.

WATENPAUGH, K. D., MARGULIS, T. N., SIEKER, L. C. & JENSEN, L. H. (1978). *J. Mol. Biol.* **122**, 175-190.