Acta Cryst. (1996). D**52**, 1004-1011

Bayesian Difference Refinement

THOMAS C. TERWILLIGER[a]* AND JOEL BERENDZEN[b]

[a]*Structural Biology Group, Mail Stop M888, Los Alamos National Laboratory, Los Alamos, NM 87545, USA, and* [b]*Biophysics Group, Mail Stop D454, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. E-mail: terwilliger@lanl.gov*

(Received 10 October 1995; accepted 21 May 1996)

Abstract

Interest in a pair of highly isomorphous structures often focuses on the differences between them. In cases where substantial correlated model errors exist or where there are differences in the quality of the two experimental data sets (cases quite common in macromolecular crystallography), independent refinement of the two structures does not lead to the most accurate estimate of the differences between them. An alternative procedure that has proven effective in some such cases is difference refinement, in which the residual between observed and calculated differences in structure-factor amplitudes between the two structures is minimized. A Bayesian approach has been used to extend the range of applicability of difference refinement to cases where there is only partial correlation in model errors and where the overlap between the data sets is limited. The resulting method, Bayesian difference refinement, uses residuals to be minimized that vary smoothly between difference refinement and independent refinement. When the errors in the two structural models are very similar, difference refinement is used; when they are very different, independent refinement is used; and when they are partially correlated, a combination of the two is used. The procedure is very simple to apply and does not significantly increase the computational demands of refinement.

1. Introduction

In the analysis of macromolecular structures, crystallographers are often interested in differences between closely related pairs of structures, a 'native' and one or more 'variants'. This situation commonly occurs, for example, in the analysis of conformational changes upon metal substitution (Fermi, Perutz, Dickinson & Chien, 1982; Luisi & Shibayama, 1989), ligand binding (Perutz, Fermi, Abraham, Poyart & Bursaux, 1986), or mutation (Nagai *et al.*, 1987; Huang *et al.*, 1990; Matthews, 1993; Eriksson, Baase & Matthews, 1993). Where a group of two or more related structures are being analyzed, typically a single native structure is refined and then each of the other structures is refined and related to this native. Ordinarily, the native structure is very carefully refined and based on more complete and more accurately measured data than the variant structures in the group, so that it might be expected to be especially accurate. However, the models used to describe scattering from macromolecular crystals are generally inaccurate in describing the experimental data, with typical residuals four times the uncertainty in measurement. This unfortunate fact tends to degrade the accuracy of estimates of differences between structures obtained by independently refining the two structures, particularly if there are differences in the quality or coverage of the two data sets.

Difference refinement (Fermi *et al.*, 1982; Terwilliger & Berendzen, 1995) is an alternative approach to estimating these structural changes that is more robust in the face of these correlated model errors. The basis for difference refinement is that the parts of a native structure that are missing from the native model are often also missing from the models of the variants. Although the most probable set of parameters for the models describing a group of structures might be those obtained from a joint refinement based on a generalization of the approach taken here, it is often impractical to re-refine the native structure every time a new variant is to be compared with it. Difference refinement assumes that the model for the native structure is well determined, and so the native structure is held fixed. The final round of atomic refinement of the variant structure is then carried out by performing a weighted least-squares minimization of the residual between observed and calculated differences in amplitudes of structure factors, given by

$$\chi^2_{\mathrm{diff}} = \sum_h \frac{[(F'_{c_h} - F_{c_h}) - (F'_{o_h} - F_{o_h})]^2}{(\sigma_h^2 + \sigma_h'^2 + E^2_{\mathrm{diff}})}. \qquad (1)$$

$F'_{c_h} - F_{c_h}$ is the difference between the native and variant amplitudes of the calculated structure factors for a reflection with indices h, $F'_{o_h} - F_{o_h}$ is the difference between observed native and variant structure factors, σ_h and σ'_h are the instrumental uncertainties in the measured amplitudes of native and variant structure factors, respectively, and E_{diff} is an estimate of the residual model error in the group of related reflections (usually a shell of resolution) into which the hth

Acta Crystallographica Section D
ISSN 0907-4449 © 1996

reflection falls. In this minimization, either all the parameters of the structure to be refined or just those affecting the variant regions could be refined.

The functional in (1) can be written in a way that emphasizes the basis of difference refinement by defining F_{diff_h} as the quantity $F'_{o_h} - (F_{o_h} - F_{c_h})$ and $\sigma^2_{\text{diff}_h}$ as $\sigma^2_h + \sigma'^2_h + E^2_{\text{diff}_h}$. Then we can rewrite (1) as

$$\chi^2_{\text{diff}} = \sum_h \frac{(F'_{c_h} - F_{\text{diff}_h})^2}{\sigma^2_{\text{diff}_h}}. \qquad (2)$$

F_{diff_h} may be viewed as the variant structure factor corrected by an estimate of a model error term, $(F_{o_h} - F_{c_h})$.

Difference refinement, as given by (2), has been applied to several structure refinements and has been shown in model and real test cases to substantially improve estimates of coordinate shifts from a native to a closely related variant structure, when compared with independently refining the two structures (Terwilliger & Berendzen, 1995). Here we address two important problems that remain. The first is related to the fundamental assumption of difference refinement, namely that the model errors in the native and variant structures are highly correlated. We have shown earlier (Terwilliger & Berendzen, 1995) that if the model error is not highly correlated then difference refinement as given by (2) can actually lead to poorer estimates of difference between two structures than does independent refinement. We will show that it is possible to take this correlation into account so that the refinement procedure will lead to estimates that are at least as good as those obtained by independent refinement, even when the correlation is low.

The second outstanding problem is that since difference refinement is based on differences between both native and variant observed and calculated data, only reflections that have been measured for both structures can be included. One situation where this would be a substantial disadvantage is presented in the analysis of Laue diffraction data, where data sets frequently have completeness as low as 50%. Suppose that two Laue data sets, one for a native and one for a variant, were collected under slightly different geometries so that each data set comprised 50% of the data to a certain resolution, and the overlap between data sets was only 25%. Only the overlapping 25% would be usable in difference refinement of the variant, and half the available data would be thrown away, clearly an unsatisfactory solution. We will show how it is possible to include the reflections that have only been measured for the variant structure in a difference refinement procedure.

We address these problems here using a Bayesian approach. This method allows us to combine all the information we have about each structure to be refined, including information about the distribution and correlation of model errors in the different structures. Once we obtain a Bayesian expression for the probability distribution for parameters in our models, we will be able to choose the values of the parameters that maximize this probability by minimizing an expression related to the one shown in (2).

2. Bayesian estimation of differences between macromolecular structures

In this analysis we will assume that we have available a well determined model for a native structure and measured amplitudes of structure factors for this native structure as well as for a closely related variant structure. Our goal is to obtain the most likely values of the parameters in a crystallographic model describing the variant structure. We begin by developing a description of the native and variant structure factors that includes the crystallographic models used as well as sources of error that are uncorrelated and correlated between the two structures. We then use the observed data to estimate probability distributions describing these errors. Finally, we integrate over possible values of these error terms, fixing the parameters in the model for the native structure, to obtain an expression for the most likely values of the parameters in the variant structure. As in our previous treatment of difference refinement (Terwilliger & Berendzen, 1995), we will be approximating the component probability distributions and complex sums to first order (*e.g.* by using Gaussian distributions).

2.1. *Correlated and uncorrelated errors in macromolecular models*

Models of macromolecular structures are generally incomplete and cannot describe all aspects of the structure, no matter what values of the parameters are used. To describe this situation, let the quantity $\mathbf{F}_h$ represent the complex native structure factor for a reflection with indices indicated by h. We have a crystallographic model for this native structure that accounts for most, but not all, of the factors that lead to $\mathbf{F}_h$, and based on this model we can obtain a calculated structure factor, $\mathbf{F}_{c_h}$. Similarly, we write the structure factor for the variant structure as $\mathbf{F}'_h$ and the corresponding structure factor, calculated from a model, as $\mathbf{F}'_{c_h}$. The key idea of difference refinement is that the portion of the native structure factor that is not described by the native model, which we term the model error for the native structure, is correlated with the model error for the variant structure. We can write this explicitly as,

$$\mathbf{F_h} = \mathbf{F_{c_h}} + \mathbf{R_h} + \mathbf{S_h} \qquad (3)$$

$$\mathbf{F'_h} = \mathbf{F'_{c_h}} + \mathbf{R_h} + \mathbf{S'_h}, \qquad (4)$$

where we have separated the model error into two terms so that we can separately analyze those model-error terms that are identical in the two structures ($\mathbf{R}_h$) from those that are present in one structure but uncorrelated with model-error terms in the other structure ($\mathbf{S}_h$ and $\mathbf{S}'_h$).

In order to simplify our calculations, we now assume that the magnitudes of the model errors ($\mathbf{R_h} + \mathbf{S_h}$) and ($\mathbf{R_h} + \mathbf{S'_h}$) are small relative to the magnitudes of $\mathbf{F}_h$ and $\mathbf{F}'_h$, respectively. In this case we can write that F_h, the magnitude of $\mathbf{F}_h$, is approximately given by

$$F_h \simeq F_{c_h} + R_h + S_h, \tag{5}$$

where F_{c_h} is the magnitude of $\mathbf{F}_{\mathbf{c}_\mathbf{h}}$, and R_h and S_h are the components of $\mathbf{R}_h$ and $\mathbf{S}_h$ along the direction of $\mathbf{F}_{\mathbf{c}_\mathbf{h}}$. Similarly, for the variant structure we can write

$$F'_h \simeq F'_{c_h} + R_h + S'_h. \tag{6}$$

Finally, we can write an expression for the experimentally observed structure-factor amplitudes, F_{o_h} and F'_{o_h}, in terms of the calculated amplitudes, F_{c_h} and F'_{c_h},

$$F_{o_h} \simeq F_{c_h} + R_h + S_h + \varepsilon_h \tag{7}$$

$$F'_{o_h} \simeq F'_{c_h} + R_h + S'_h + \varepsilon'_h, \tag{8}$$

where ε_h and ε'_h are the errors in measurement of the native and variant structure factors, respectively.

2.2. *Probability distribution for parameters in the model for the variant structure*

To obtain a probability distribution for the parameters θ' in the model describing the variant structure, we begin by using Bayes' rule (Box & Tiao, 1973) to write an expression for the posterior probability distribution for θ' given that we have made measurements $\{F_{o_h}\}$ and $\{F'_{o_h}\}$ of the native and variant structure factors, where the brackets indicate that we are referring to the entire data sets. In this analysis, we will assume that the parameters in the model for the native structure are already accurately known. (In a more complete treatment, beyond the scope of this work, this assumption could be relaxed.) We write that,

$$p(\theta'|\{F_{o_h}, F'_{o_h}\}) \propto p(\{F_{o_h}, F'_{o_h}\}|\theta')p_o(\theta'). \tag{9}$$

The prior probability distribution for the parameters in the model for the variant structure, $p_o(\theta')$ includes all our expectations of the bond angles, distances, and other restraints that are commonly included in macromolecular refinement (Hendrickson & Konnert, 1980; Konnert, 1976; Sussman, Holbrook, Church & Kim, 1977; Tronrud, Ten Eyck & Matthews, 1987).

We can obtain an expression for the probability distribution $p(\{F_{o_h}, F'_{o_h}\}|\theta')$ on the right hand side of (9) in several steps. Using (5)–(8) we can calculate the related probability distribution $p(\{F_{o_h}, F'_{o_h}\}|\theta', \{R_h, S_h, S'_h\})$, assuming that the measurement errors ε_h and ε'_h are normally distributed,

$$p(\{F_{o_h}, F'_{o_h}\}|\theta', \{R_h, S_h, S'_h\}) \propto \prod_h \mathcal{N}(F_{o_h} - F_{c_h}, \sigma_h^2)\mathcal{N}(F'_{o_h} - F'_{c_h}, \sigma_h'^2), \tag{10}$$

where $\mathcal{N}(x, \sigma^2) = 1/\sigma(2\pi)^{1/2}\exp(-x^2/2\sigma^2)$ represents a normal distribution with variance σ^2, and σ_h and σ'_h are the uncertainties in measurement of F_{o_h}, and F'_{o_h}, respectively. We will obtain information about probability distributions for R_h, S_h, and S'_h below. Meanwhile we can obtain an estimate of $p(\{F_{o_h}, F'_{o_h}\}|\theta')$ by integrating (10) over the 'nuisance' variables R_h, S_h and S'_h (Box, 1980). Assuming that (see below) R_h, S_h, and S'_h are independent of F_{c_h} and F'_{c_h} we can write

$$p(\{F_{o_h}, F'_{o_h}\}|\theta') \propto \int p(\{F_{o_h}, F'_{o_h}\}|\theta', \{R_h, S_h, S'_h\})p_o(\{R_h\})p_o(\{S_h\})p_o(\{S'_h\}) \times \{\mathrm{d}R_h\}\{\mathrm{d}S_h\}\{\mathrm{d}S'_h\}, \tag{11}$$

where $p_o(\{R_h\}), p_o(\{S_h\})$, and $p_o(\{S'_h\})$ are estimates of the prior probability distributions for R_h, S_h and S'_h, and the brackets again indicate that the entire data set is to be considered.

2.3. *Prior probability distributions for the errors*

To make estimates of the prior distributions $p_o(\{R_h\})$, $p_o(\{S_h\})$, and $p_o(\{S'_h\})$ we assume that R_h, S_h, and S'_h are small relative to F_{c_h} and F'_{c_h} and also uncorrelated with them. As discussed by Read (1986), this assumption is not strictly true and the model errors are generally negatively correlated with the calculated structure factors. If the model errors are small, however, this correlation will be very small and may for our purposes be ignored.

So long as the structure factors $\mathbf{R}_h$, $\mathbf{S}_h$ and $\mathbf{S}'_h$ are due to scattering at a number of locations in the unit cell of the native and variant crystals, their prior probability distribution can be quite reasonably described by Wilson statistics (Wilson, 1949). The components R_h, S_h and S'_h along the direction of the calculated native structure factor will then have a normal prior probability distribution with a variance dependent on the resolution of the reflection. We can write that

$$p_o(R_h) = \mathcal{N}(R_h, \alpha E^2), \tag{12}$$

where for centric reflections the value of α is the expected intensity factor (Stewart & Karle, 1976) and for acentric reflections the value of α is half the expected intensity factor (Terwilliger & Eisenberg, 1987), and E^2 is a measure of the total correlated model error. Similar analyses may be applied to S_h, and S'_h, leading to,

$$p_o(S_h) = \mathcal{N}(S_h, \alpha A^2). \tag{13}$$

$$p_o(S'_h) = \mathcal{N}(S'_h, \alpha A'^2), \tag{14}$$

where the model errors that are uncorrelated between the native and variant structures are given by αA^2 and $\alpha A'^2$, respectively.

We use a procedure similar to the one we previously developed for estimation of model errors in difference refinement (Terwilliger & Berendzen, 1995) to estimate the total correlated error E^2 and the uncorrelated errors A^2 and A'^2. From (7) and (8), if we had a good estimate of the parameters in the model for the variant structure, θ', we could use the part of $F_{o_h} - F_{c_h}$ that is correlated with $F'_{o_h} - F'_{c_h}$ to estimate the mean-square value of R_h in a range of resolution. That is,

$$\langle (F_{o_h} - F_{c_h})(F'_{o_h} - F'_{c_h})\rangle \simeq \langle R_h^2\rangle \simeq \alpha E^2, \tag{15}$$

where centric and acentric reflections are treated separately, α is as defined above, and the averages are taken over reflections in a range of resolution. Similar arguments lead to the relations,

$$\langle (F_{o_h} - F_{c_h})^2\rangle \simeq \langle R_h^2 + S_h^2 + \varepsilon_h^2\rangle \simeq \alpha E^2 + \alpha A^2 + \langle \sigma_h^2\rangle, \tag{16}$$

and

$$\langle (F'_{o_h} - F'_{c_h})^2\rangle \simeq \langle R_h^2 + S_h'^2 + \varepsilon_h'^2\rangle \simeq \alpha E^2 + \alpha A'^2 + \langle \sigma_h'^2\rangle. \tag{17}$$

(15), (16) and (17) can be used to estimate the parameters E^2, A^2 and A'^2.

2.4. *Bayesian difference refinement*

Substituting (10) into (11), integrating over the nuisance variables, substituting the result into (9), and taking the negative logarithm of both sides, yields an expression (neglecting a factor of two) for the log likelihood of a particular set of parameters θ' describing the model for the variant structure,

$$R = -\ln P(\theta') = \sum_h \frac{(F'_{c_h} - F_{B\text{diff}_h})^2}{\sigma^2_{B\text{diff}}} - \ln p_o(\theta'), \tag{18}$$

where $F_{B\text{diff}_h}$ is given by

$$F_{B\text{diff}_h} = F'_{o_h} - \beta(F_{o_h} - F_{c_h}). \tag{19}$$

Note that (18) and (19) are very similar to (2), except that the estimate of the model error made from the observed and calculated amplitudes of the native structure factor is multiplied by the factor β before subtracting it from the observed amplitude of the variant structure factor and (18) includes the prior probability distribution $p_o(\theta')$. The value of the factor β is determined by the ratio of the sum of the mean-square correlated model error term, E^2, to the sum of the mean-square model error term unique to the native structure and the variance in the measurement of the amplitude of the native structure, $A^2 + \sigma_h^2$,

$$\beta = E^2/(E^2 + A^2 + \sigma_h^2). \tag{20}$$

In large part, the factor β reflects the correlation between the model error terms for the native and variant structures, but it also reflects the correlation of the errors in the parameters in the models. If there were no errors in the model for the native structure, and if there were no correlation between the model-error terms ($E^2 = 0$) then β would be zero and the model-error term for the native structure would not be applied at all to the variant structure. If the correlation and model errors were high, then $F_{B\text{diff}}$ reduces to F_{diff}, and Bayesian difference refinement is equivalent to difference refinement.

Finally, the weighting factor in (18) is given by,

$$\sigma^2_{B\text{diff}_h} = \sigma_h'^2 + A_h'^2 + 1/[1/(\sigma_h^2 + A^2) + 1/E^2]. \tag{21}$$

That is, the uncertainty in $F_{B\text{diff}_h}$ is related to the experimental uncertainties in the measurement of the amplitudes of the native and variant structure factors, σ_h and σ'_h, the model errors unique to the native and variant structures, A_h and A'_h, and to the correlated model error, E.

Note that (19) provides an estimate of $F_{B\text{diff}_h}$ for each reflection where the amplitude of the variant structure factor, F'_{o_h} is measured, regardless of whether the amplitude of the corresponding native structure factor has been measured. For a reflection where the native value has not been measured ($\sigma_h^2 = \infty$) then $\beta = 0$ in 19, and (21) reduces to, $\sigma^2_{B\text{diff}_h} = \sigma_h'^2 + A_h'^2 + E^2$. That is, for reflections where the amplitude of the native structure factor has not been measured, the model error that is unique to the variant model and that which is correlated with the native model contribute equally to the uncertainty in $F_{B\text{diff}_h}$. For reflections where both native and variant have been measured, the correlated model error E^2 can contribute much more weakly [through the last term in (21)]. The result of this is that reflections for which both native and variant data have been measured contribute much more information on the variant structure than those where only variant data are measured, but all the measured variant data can be included in refinement.

3. Comparison of independent refinement, difference refinement and Bayesian difference refinement using test data

We recently constructed several test cases for evaluating difference refinement based on known native and variant structures consisting of peptides with 51 atoms and two water molecules (Terwilliger & Berendzen, 1995). The native and variant structures were each generated using short molecular dynamics simulations with *X-PLOR* (Brünger, Kuriyan & Karplus, 1987). For the test case we will consider here, the known native and variant structures differed by an r.m.s. distance of

0.1 Å. The water molecules were used to simulate that part of the known structures that was not included in the crystallographic model. That is, they were used in generating simulated native and variant data sets, but were not included in the modeling. For the case considered here, these water molecules were placed in different positions in the known native and variant structures, with the shifts in these coordinates from native to variant ranging from 0 to 1.6 Å. In this way, the effects of decreasing correlation of model errors on the refinement process could be assessed.

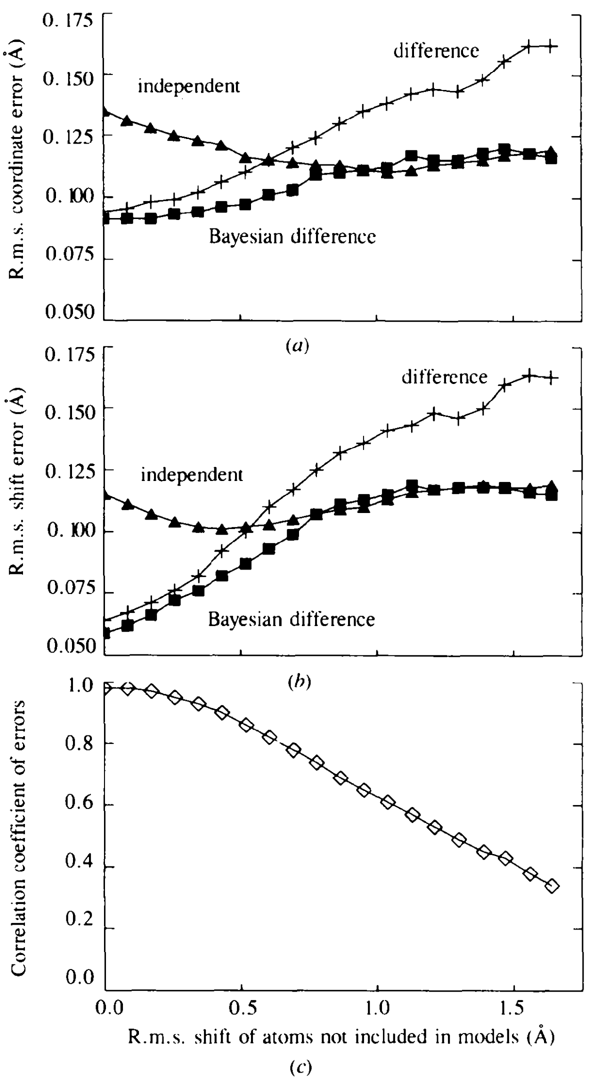

Fig. 1. The effects of decreasing correlation of modeling errors on independent refinement (crosses), difference refinement (triangles), and Bayesian difference refinement (squares) with varying correlated errors. Simulated native and a variant data sets were prepared as described in the text. The shift between the known native and variant structures was 0.1 Å r.m.s.; measurement errors were 5%. 50% of the variant structure data was not used in refinement. Two unmodelled water molecules were added at different positions in the native and variant structures, and the x axis shows the r.m.s. shift between their positions in the wild-type and variant structures. The standard R factors for the refinement were all approximately 20% (Terwilliger & Berendzen, 1995). (*a*) R.m.s. errors in the model variant atomic coordinates. (*b*) R.m.s. errors in the displacements from model native to model variant structures. (*c*) The correlation coefficient of model residuals, β.

In these tests, simulated data were generated for 1210 structure factors from 8 to 2 Å and a 5% normally distributed measurement error was added to each observed amplitude. The model for the native structure was obtained by least-squares restrained refinement with a modified version of ***PROLSQ*** (Konnert, 1976; Hendrickson & Konnert, 1980) using the known native structure as the starting model (Terwilliger & Berendzen, 1995). The refined native structure differed from the known native structure by 0.083 Å r.m.s. Variant structures were refined in the same way, but either by independent refinement (using exactly the same method as for the native), by difference refinement, or by Bayesian difference refinement. To simulate a case where complete data was not available for the variant structure, 50% of the data for the variant structure were not used in the refinement process.

Previously, we found that difference refinement yields considerably smaller errors in coordinate shifts from native → variant than does independent refinement when the correlation of model errors is high. As this correlation decreases, however, the errors in coordinate shifts obtained with difference refinement becomes even larger than those obtained with independent refinement. We now add Bayesian difference refinement to the comparison. Fig. 1(*a*) compares the overall coordinate errors in the variant structures obtained with independent refinement, difference refinement, and Bayesian difference refinement as a function of the r.m.s. shift coordinates of the water molecules from native → variant structures that were not included in the refinements. Fig. 1(*b*) compares the errors in coordinate shifts from native → variant for the same three methods, and Fig. 1(*c*) shows the values of the factor β.

Figs. 1(*a*) and 1(*b*) show that when the positions of the water molecules not included in the refinements are increasingly shifted between the known native and variant structures, the relative r.m.s. errors in the variant coordinates obtained by difference refinement and by independent refinement do change considerably. When the shift in water molecule position is zero, for example, difference refinement yields an r.m.s. coordinate error of 0.094 Å and an r.m.s. error in coordinate shifts from native → variant of 0.064 Å compared to an r.m.s. coordinate error of 0.135 Å and an r.m.s. error in coordinate shifts of 0.115 Å for independent refinement. Note that the principal reason why difference refinement yields a more accurate structure than independent refinement in this case is that only half the reflection data are used in the refinement of the variant structure using either refinement method, but difference refinement includes information on the native structure, which was refined using all the data. When the coordinate shift for the water molecules is 0.6 Å, however, both methods yield r.m.s. coordinate errors of 0.115 Å and the two methods yield similar errors in coordinate shifts of 0.103 to 0.110 Å. When the coordinate shift of the water molecules is 1.6 Å, difference refinement yields an r.m.s. coordinate error of 0.162 Å and an error in coordinate shifts of 0.163 Å while independent refinement yields an r.m.s. coordinate errors and r.m.s. errors in coordinate shifts of only 0.119 Å. That is, when the model errors in native and variant structures have little correlation, difference refinement is very inaccurate relative to independent refinement.

As anticipated from the theoretical treatment we have presented, Bayesian difference refinement combines the best aspects of independent and difference refinement. When the model errors for native and variant structures are highly correlated and difference refinement is most effective, the factor β in (20) is nearly unity (Fig. 1*c*) and Bayesian difference refinement is essentially equivalent to difference refinement. The r.m.s. coordinate error when the shift in water molecule positions is zero is just 0.091 Å and the r.m.s. error in coordinate shifts is 0.059 Å. Conversely, when the model errors have little correlation, the factor β is small and Bayesian difference refinement is very nearly the same as independent refinement. For moderate correlations of model errors in the range $0.8 < \beta < 0.9$, however, Bayesian difference refinement is superior to the other two methods. When the r.m.s. shift in coordinates of water molecules is 0.606 Å for example, the r.m.s. coordinate error using Bayesian difference refinement is 14% better than that obtained by either independent or difference refinement.

4. Comparison of Bayesian difference refinement with independent refinement using data for a mutant of gene V protein

The purpose of using Bayesian difference refinement is to obtain accurate estimates of coordinate shifts from one structure to another when these shifts are quite small. One case where this is likely to be particularly useful is in comparing the structure of a mutant protein containing one or a few amino-acid substitutions with that of the wild-type protein. Such a case is illustrated in Fig. 2, which compares independent refinement of a mutant structure with Bayesian difference refinement of the same structure. In this example, the wild-type structure of gene V protein has been refined using data to a resolution of 1.8 Å, and data on the Ile47→Val mutant to a resolution of 1.8 Å is used to refine the structure of the mutant protein. In order to fully test the refinement procedures, the structure of the mutant is refined using simulated annealing (Brünger *et al.*, 1987).

When the structure of the mutant protein is refined independently to the wild-type protein, side chains that are not very well defined in either structure (such as

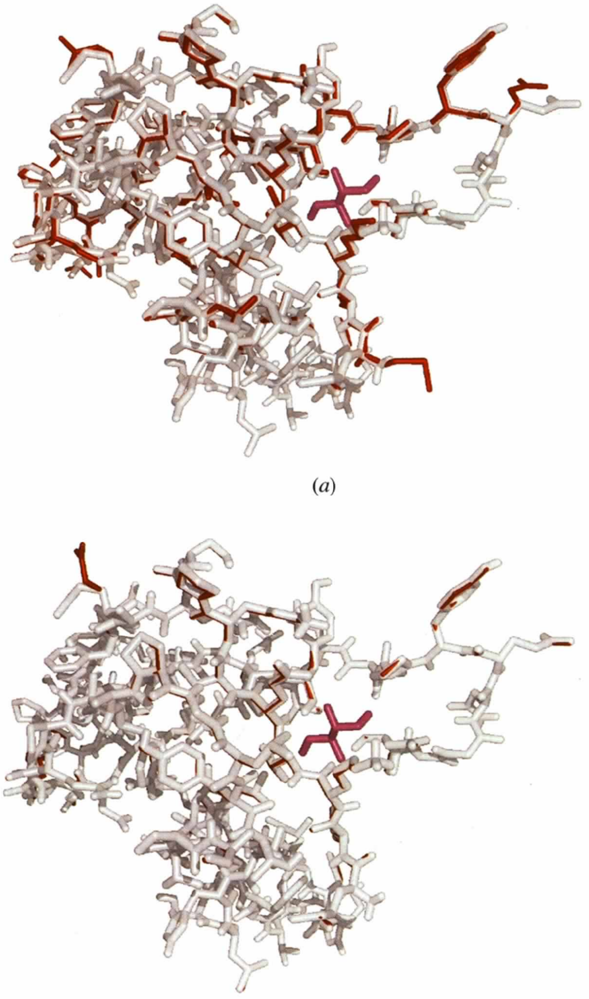

Fig. 2. Comparison of independent refinement and Bayesian difference refinement for a gene V protein mutant. The structure of the Ile47→Val mutant of gene V protein (Zhang, Skinner, Sandberg, Wang & Terwilliger, 1996) was refined with simulated annealing using *X-PLOR* (Brünger *et al.*, 1987) using the data from 5 to 1.8 Å. This structure has previously been refined at a resolution of 1.8 Å without using simulated annealing (Zhang *et al.*, 1996). In the test refinement, the coordinates of water molecules were restrained with harmonic restraints. All coordinates and thermal factors were refined. Identical procedures were used for independent and Bayesian difference refinement except for the choice of observed or $F_{\rm diff}$ structure factors. In each case the refined structure of the mutant protein is superimposed on that of the wild-type gene V protein (Skinner *et al.*, 1994). Ile47 is shown in magenta, the wild-type structure is white, the Ile→Val mutant structure is red. (*a*) Independe:.t refinement. (*b*) Bayesian difference refinement.

many of those on the protein surface) will tend to shift somewhat arbitrarily between the two structures. This may be seen in Fig. 2(*a*), where when independent refinement is used, the mutation in the center of the figure at Val35 appears to affect many atoms that are located all around the structure, particularly on the surface. If, on the other hand, Bayesian difference refinement is used, those shifts which are reflected in the differences in measured data tend to occur, while those side chains that are poorly defined tend to remain in place. Fig. 2(*b*) shows that using Bayesian difference refinement only atoms near the site of mutation are substantially affected by the mutation. In effect, Bayesian difference refinement makes those changes that are indicated by difference Fourier analyses and tends to leave other atoms in place.

5. Discussion

In deriving Bayesian difference refinement, we have assumed that the native structure is at least as well determined as the variant, that the model errors in the native and variant structure are partly correlated, that the measurement and model-error distributions may be approximated as normal distributions, and that the model errors are uncorrelated with the calculated structure factors. Under these assumptions, which can be made good to first order in realistic macromolecular cases, minimization of the Bayesian difference residual (18) will lead to the most probable values of the parameters of the model describing the changes from native → variant structures. For example, suppose that the native structure contained some substantial coordinate errors that could have been reduced by further refinement of that structure. Now suppose that we refine a very closely related variant structure that is, in fact, essentially identical to the native structure in this region. Carrying out Bayesian difference refinement on the variant structure will cause the same coordinate errors to be made in the variant structure that were made in the native structure. The difference refinement method will correctly yield very small shifts from native to variant in this region, but the absolute structure of the variant will be in error, matching the native. This result is of course desirable in the most common case when it is the changes between native and variant structures that are of interest.

The new feature of Bayesian difference refinement is that the estimate of the model-error term based on the native structure, $F_{o_h} - F_{c_h}$, is multiplied by the factor β before using it as a correction for the amplitude of the variant structure factor (19). An intuitive explanation as to why this might be a good idea is as follows. Suppose the native and variant structures are very similar, and the model-error terms for the two structures are very similar as well. In this case, correcting the amplitude of the variant structure factor by the entire model-error term estimated from the native structure would lead to a very good estimate of that part of the amplitude of the variant structure factor that can be represented by our model. Refining our variant model by difference refinement would lead to a very good estimate of coordinate shifts. Now suppose instead that the model-error terms for the native and variant structures had no similarity whatsoever. In this case, 'correcting' the amplitude of the variant structure factor with a model-error term estimated from the native structure would only increase the error in our estimate of shifts. The factor β, which depends largely on the correlation of model errors between the two structures, allows the subtraction of a reasonable fraction of the model term estimated from the native structure. Our tests using model data (Fig. 1) show that the factor β allows Bayesian difference refinement to combine the best aspects of difference refinement and of independent refinement into a single method.

The second new feature of Bayesian difference refinement is that it provides a means of including measured structure-factor information for reflections where the amplitude of the variant structure factor has been measured but no measured data is available for the native structure. The factor β for these reflections will be zero and σ_h^2 will tend towards infinity. Thus, the relative weighting of these observations will be lower than for those where native data is available, but they will be included in the refinement.

A third feature of Bayesian difference refinement is that the appropriate weighting scheme is clear [(21)]. This weighting is similar to that given in (2), but it explicitly incorporates the uncertainty in amplitudes of structure factors calculated from the model for the native structure and treats model errors that are correlated between the structures separately from those that are unique to each structure.

6. Conclusions

Bayesian difference refinement appears to be a very simple and useful method for incorporating knowledge about the deficiencies in a crystallographic model for one structure into the refinement of a related 'variant' structure. Using the Bayesian approach, the extent of correlation of model errors in two structures can be obtained and applied to the refinement of the variant structure. In our tests using model data, Bayesian difference refinement yielded variant structures and coordinate shifts from native → variant structures that were as accurate as, or more accurate than, difference refinement and independent refinement.

The authors would like to thank the NIH and the LDRD program of Los Alamos National Laboratory for generous support. Bayesian difference refinement is available from TT as part of *HEAVY* version 4.

References

Box, G. E. P. (1980). *J. Roy. Statist. Soc. A*, **143**, 383-430.

Box, G. E. P. & Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. New York: John Wiley.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Science*, **235**, 458-460.

Eriksson, A. E., Baase, W. A. & Matthews, B. W. (1993). *J. Mol. Biol.* **229**, 747-769.

Fermi, G., Perutz, M. F., Dickinson, L. C. & Chien, J. C. W. (1982). *J. Mol. Biol.* **155**, 495-505.

Hendrickson, W. A. & Konnert, J. H. (1980). *Computing in Crystallography*, edited by R. Diamond, S. Rameshan, K. Venkatesan, p. 13.01. Bangalore: Indian Academy of Science.

Huang, Y., Pagnier, J., Magne, P., Baklouti, F., Kister, J., Delaunay, J., Poyart, C., Fermi, G. & Perutz, M. F. (1990). *Biochemistry*, **29**, 7020-7023.

Konnert, J. H. (1976). *Acta Cryst.* A**32**, 614-617.

Luisi, B. & Shibayama, N. (1989). *J. Mol. Biol.* **206**, 723-736.

Matthews, B. W. (1993). *Annu. Rev. of Biochem.* **62**, 139-160.

Nagai, K., Luisi, B., Shih, D., Miyazaki, G., Imai, K., Poyart, C., De Young, A., Kwiatkowsky, L., Noble, R. W., Lin, S.-H. & Yu, N.-T. (1987). *Nature (London)*, **329**, 858-860.

Perutz, M. F., Fermi, G., Abraham, D. J., Poyart, C. & Bursaux, E. (1986). *J. Am. Chem. Soc.* **108**, 1064-1078.

Read, R. J. (1986). *Acta Cryst.* A**42**, 140-149.

Skinner, M. M., Zhang, H., Leschnitzer, D. H., Bellamy, H., Sweet, R. M., Gray, C. M., Konings, R. N. H., Wang, A. H.-J. & Terwilliger, T. C. (1994). *Proc. Natl Acad. Sci. USA*, **91**, 2071-2075.

Stewart, J. M. & Karle, J. (1976). *Acta Cryst.* A**32**, 1005-1007.

Sussman, J. L., Holbrook, S. R., Church, G. M. & Kim, S.-H. (1977). *Acta Cryst.* A**33**, 800-804.

Terwilliger, T. C. & Berendzen, J. (1995). *Acta Cryst.* D**51**, 609-618.

Terwilliger, T. C. & Eisenberg, D. S. (1987). *Acta Cryst.* A**43**, 6-13.

Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. (1987). *Acta Cryst.* A**43**, 489-501.

Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318-321.

Zhang, H., Skinner, M. M., Sandberg, W. S., Wang, A. H.-J. & Terwilliger, T. C. (1996). *J. Mol. Biol.* In the press.