Acta Cryst. (1993). D**49**, 24–36

Assessment of Phase Accuracy by Cross Validation: the Free *R* Value. Methods and Applications

BY AXEL T. BRÜNGER

The Howard Hughes Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, USA

(*Received* 6 *January* 1992; *accepted* 6 *July* 1992)

Abstract

Analogies between the free R statistic [Brünger (1992). *Nature* (*London*), **355**, 472–474] and the statistical methods of cross validation and bootstrap are discussed. Several new applications which make use of the previously observed correlation between the free R value and the phase accuracy of crystal structures are presented. One application concerns the relative weighting of individual restraint classes in macromolecular refinement. The free R value provides an objective statistical basis for the optimal choice of the weights. The results for the refinement of a penicillopepsin crystal structure at 1.8 Å resolution indicate that overall bond length and bond angle weights, derived from uncertainties observed in small-molecule crystal structures, appear to be transferable to macromolecules. In another application, the landscape of the R value around the crystal structure was investigated for unrestrained modeling of diffraction data with equal atomic scatterers. Others have suggested applications to *ab initio* phasing because of the simplicity of the liquid-like system of equal atomic scatterers. However, there are a large number of incorrect configurations of the scatterers whose R values at 1.8 Å resolution are close to that of the correct configuration given by the positions of the non-hydrogen atoms in the penicillopepsin crystal structure. A substantial number of the incorrect configurations have higher free R values than the correct one. It is therefore conceivable that the free R value could be used as a selection criterion to distinguish between certain incorrect configurations and configurations close to the correct one.

1. Introduction

The three-dimensional crystallographic structure determination of macromolecules involves fitting appropriate electron-density models to the observed diffraction data. Examples of this highly non-linear and multidimensional fitting process include phasing (Hauptman, 1991; Weeks, DeTitta, Miller & Hauptman, 1993), molecular replacement (Hoppe, 1957; Rossmann & Blow, 1962; Brünger, 1990), computer-graphics modeling guided by density maps (Jones, 1978) and restrained refinement (Hendrickson, 1985). As macromolecular diffraction data are usually not available to atomic resolution, these optimization procedures may become under-determined unless the diffraction data are augmented with information about the chemical and physical laws (such as atomicity, connectivity, geometry, stereochemistry or packing) that govern macromolecular structure. Diffraction data and prior knowledge are therefore often combined, for example in restrained least-squares refinement (Hendrickson, 1985).

The most common measure for the agreement between an electron-density model and the observed diffraction data is the R value, defined as

$$R = \frac{\sum_{\mathbf{h}} ||F_{\rm obs}(\mathbf{h})| - k|F_{\rm calc}(\mathbf{h})||}{\sum_{\mathbf{h}} |F_{\rm obs}(\mathbf{h})|} \qquad (1)$$

where $\mathbf{h} = (h, k, l)$ are the reciprocal lattice points of the crystal, k is a scale factor, $|F_{\rm obs}(\mathbf{h})|$ and $|F_{\rm calc}(\mathbf{h})|$ are the observed and calculated structure-factor amplitudes respectively. The R value is closely related to the crystallographic residual R' which has more convenient mathematical properties,

$$R' = \sum_{\mathbf{h}} [|F_{\rm obs}(\mathbf{h})| - k|F_{\rm calc}(\mathbf{h})|]^2. \qquad (2)$$

It can be shown that R' is a linear function of the negative logarithm of the likelihood of the atomic model, assuming that all observations are independent and normally distributed (Press, Flannery, Teukolsky & Vetterling, 1986). Least-squares refinement minimizes R', thus maximizing the likelihood of the atomic model. R' can be made arbitrarily small by increasing the number of model parameters and further refinement. The theory of linear hypothesis tests has been employed to decide whether the addition of parameters or the imposition of fixed relationships between parameters results in a *significant* improvement or a *significant* worsening of the agreement between the atomic model and the diffraction data (Hamilton, 1965). This theory strictly applies to the situation where the restraints can be expressed as holonomic boundary conditions, *e.g.* fixed bond lengths, and therefore not to non-linear restraints such as repulsive contact functions (Hendrickson, 1985) or empirical energy functions (Brooks, Bruccoleri, Olafson, States & Swaminathan, 1983).

In an analogy to testing statistical models by cross validation, we recently proposed the free R statistic (which measures the agreement between the atomic model and the diffraction data) for a 'test' set of reflections omitted in the

0907-4449/93/010024-13$06.00

fitting procedure (Brünger, 1992*a*). A high correlation was observed between the free R value and the phase accuracy of the atomic model; this was independent of the number of model parameters and restraints.

In this work we review briefly the statistical method of cross validation, placing particular emphasis on crystallographic applications. We then extend our previous work (Brünger, 1992*a*) to the weighting of individual restraint classes such as bond lengths and bond angles. We also illustrate the influence of noise in the data on the free R value. Finally, we study the landscape of the R value for liquid-like models of equal atomic scatterers in the neighborhood of the crystal structure.

2. Methods

2.1. *The test case: a crystal structure of penicillopepsin*

As in our previous paper (Brünger, 1992*a*), we used the penicillopepsin crystal structure from *Penicillium janthinellum* (Hsu, Delbare, James & Hofmann, 1977; James & Sielecki, 1983) as a prototype for a high-quality crystal structure at 1.8 Å resolution. The availability of good experimental phases obtained from multiple isomorphous replacement with eight heavy-atom derivatives and an overall mean figure of merit of 0.9 for all observed data to 2.8 Å resolution (Hsu *et al.*, 1977; James & Sielecki, 1983) made this an ideal test case. We used the figure-of-merit-weighted mean difference between the most probable multiple-isomorphous-replacement phases and the model phases ($|\overline{\Delta\Phi}|$) to estimate the accuracy of the model phases to 2.8 Å resolution.

2.2. *The computer program*

All calculations were carried out with the program *X-PLOR* (Brünger, 1992*b*). (Requests for *X-PLOR* should be made to the author.)

3. Theory

In this section we briefly review classical statistical regression theory, introduce the modern statistical tools of cross validation and bootstrap, and finally discuss their application to macromolecular crystallography.

3.1. *Classical statistical regression and significance tests*

Suppose we have a noisy data set consisting of n pairs of points

$$(x_i, y_i);\ i = 1, \ldots, n \tag{3}$$

where $y_i = (y_1, \ldots, y_n)$ are the observations and x_i are the predictor variables, for example the intensities $I(\mathbf{h})$ and Miller indices (h) of a crystallographic diffraction experiment.

Least-squares theory assumes a set of model functions $M(x_i, \beta), i = 1, \ldots, n$ as a function of m parameters $\beta = (\beta_1, \ldots, \beta_m)$. Ideally, one wants to find parameters $\beta^o = (\beta_1^o, \ldots, \beta_m^o)$ such that

$$y_i = M(x_i, \beta^o) + e_i \tag{4}$$

where e_i is an experimental error drawn at random from a population with zero mean. In the crystallographic case, the model parameters and functions are typically chosen to be the atomic coordinates of the crystal structure and the computed structure-factor amplitudes respectively.

It can be shown that the 'best' estimator of β^o is given by the minimum $\hat{\beta}^o$ of the residual

$$L(\beta) = \sum_{i=1}^{n} [y_i - M(x_i, \beta)]^2. \tag{5}$$

The estimator $\hat{\beta}^o$ is 'best' in the sense that it represents the maximum likelihood estimation of β^o (Press *et al.*, 1986).

Linear regression analysis is a special case of least-squares theory in which the model functions are linear functions of the parameters β,

$$M(x_i, \beta) = \sum_{j=1}^{m} \beta_j F_{ji} = (\mathrm{F}^* \beta)_i \tag{6}$$

where the asterisk denotes the transposition operation. The solution to this linear least-squares problem is the well-known 'normal' equation

$$\beta = (\mathrm{FF}^*)^{-1} \mathrm{F} y. \tag{7}$$

If the model functions M are not linear, the problem can be solved approximately by linearizing M,

$$M(x_i, \beta') \simeq M(x_i, \beta) + \sum_{j=1}^{m} \frac{\partial M(x_j, \beta)}{\partial \beta_j} (\beta_j' - \beta_j), \tag{8}$$

and by iterative application of the normal equation (7).

A classical problem in regression analysis is the determination of the optimal number of parameters m in order to maximize the information contained in the model. Too few parameters will not fit the data satisfactorily whereas too many parameters might fit noise. At a first glance, one might choose the residual $L(\hat{\beta}^o)$ [equation (5)] to assess the information content but the problem with this choice is that the same data are used both for the least-squares estimator $\hat{\beta}^o$ and to assess the information content. In fact, $L(\hat{\beta}^o)$ decreases monotonically as the number of parameters m increase. The classical procedure to circumvent this problem is to compute the significance of a change in $L(\hat{\beta}^o)$ upon changing the number of parameters from m_1 to m_2. The so-called F value plays a pivotal role in estimating this significance,

$$F = \frac{L[\hat{\beta}_o(m_1)] - L[\hat{\beta}_o(m_2)]}{L[\hat{\beta}_o(m_2)]} \frac{n - m}{m_2 - m_1}. \tag{9}$$

The change is considered insignificant if the F value is 'close' to unity; whether this is the case can be determined from the theoretical distribution of F. Hamilton (1965) expressed the F test in terms of the more widely used R values in crystallography. While the F test enjoys great popularity, it is strictly limited to linear models and normally distributed experimental errors e_i.

3.2. *Cross validation*

Cross validation (Stone, 1974; Efron, 1982, 1988) overcomes the limitations of the classical F test. It is based on the idea that the information content of the model should be related to the residual L computed for some new data (*i.e.* data not involved in the determination of the least-squares estimator $\hat{\beta}^o$). It is common, however, to omit a single data point i from the original data set and to determine the least-squares estimator $\hat{\beta}^o(i)$ using equation (5) with i omitted rather than obtain new data. As the residual for the single data point is prone to large statistical fluctuations, it is more appropriate to obtain n least-squares estimators $\hat{\beta}^o(i)$ for each data point i in the original data set. The cross-validation residual is then defined as

$$L_{\mathrm{CV}} = \sum_{i=1}^{n} \{y_i - M[x_i, \hat{\beta}^o(i)]\}^2. \tag{10}$$

In the case of linear least squares, L_{CV} can be computed analytically from a single least-squares solution of an appropriately defined residual (Härdle, Hall & Marron, 1988). However in the general case, n least-squares evaluations are required to compute L_{CV}.

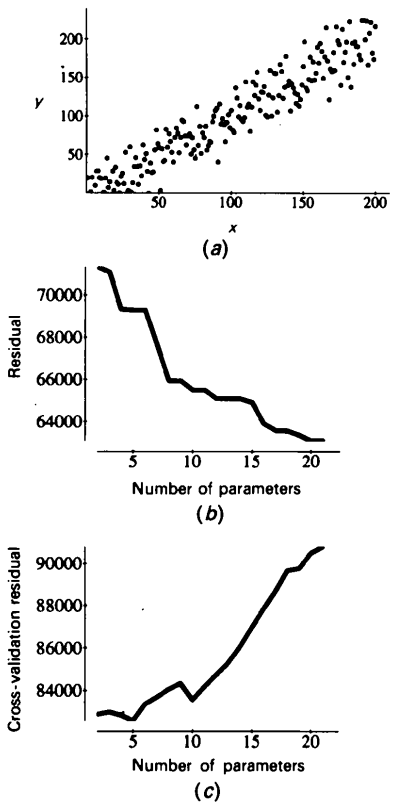


Fig. 1. Application of cross validation to the least-squares fitting of the noisy data set described in equation (11). (*a*) Actual data set used in the example. (*b*) Residuals $L(\hat{\beta}_o)$ obtained by linear least-squares fits [equation (7)] using the functions described in equation (12) for $2 \leq m \leq 22$; the residuals are plotted as a function of the parameters m used in the least-squares fit. (*c*) Cross-validation residual as defined in equation (10).

Fig. 1 illustrates this procedure with a simple example. Consider a hypothetical two-dimensional data set

$$[x, y(x) = x + N(0, 20)];\ x = 1, \ldots, 200 \tag{11}$$

where N is a normal (Gaussian) distribution with a mean of zero and standard deviation of 20 (Fig. 1*a*). The Gaussian noise is supposed to simulate measurement errors of the observations y. A set of functions

$$f_m(x) = a_1 + a_2 x + \sum_{i=3}^{m} a_m \sin(2^{m-2} x);\ m = 2, \ldots, 22 \tag{12}$$

were then fitted to this data set by linear least squares [equation (6)]. The least-squares residual $L(\hat{\beta}^o)$ is plotted in Fig. 1(*b*) as a function of the number of parameters m for $2 \leq m \leq 22$. As expected, the residual can be arbitrarily lowered by increasing the number of parameters m; yet by definition, the only reasonable fit is the linear function f_2.

Fig. 1(*c*) reports the cross-validation residuals L_{CV} for the least-squares fits using equation (6). Note that cross validation involved n least-squares evaluations for each fitted function, *i.e.* this required 21×200 evaluations of equation (6). We choose not to use the analytical formula of Härdle *et al.* (1988) as this is not applicable to the crystallographic problems discussed below. The cross-validation residual tends to increase as the number of parameters increase and thus cross validation suggests correctly that the functions f_m for larger m overfit the data. As a result of the small number of data points in the test set, the cross-validation residual L_{CV} in Fig. 1(*c*) is too noisy to determine precisely the optimal value of m. This situation could be improved by the bootstrap method (Efron, 1988). Artificial 'bootstrap' data sets are simulated by randomly drawing n data pairs from the original data set with replacement. Cross validation is then carried out for the bootstrap data sets. The distribution of the resulting cross-validation residuals can be used to compute the mean value of L_{CV} and the standard deviation.

Cross validation and bootstrap are examples of modern statistics where computer experiments produce probability distributions (Efron, 1988; Hinkley, 1988; Efron & Tibshirani, 1991). Raw computing power has replaced tedious and often impossible analytical calculations. The beauty of this approach is that it can be applied to any statistical modeling procedure, not just least-squares fitting.

Cross validation as described in equation (10) is clearly impractical for large data sets and so we propose the following modification. The original data set is partitioned into t disjoint 'test' sets T_r of equal size obtained by drawing them randomly from the original data set, *i.e.*,

$$\text{data} = \bigcup_{r=1}^{t} T_r \tag{13}$$

$$T_i \cap T_j = \emptyset.$$

Least-squares evaluations are then carried out with the test sets T_r omitted. We refer to the least-squares solutions as $\hat{\beta}^o(T_r)$. In analogy to equation (10), we define the cross-validation residual

$$L_{\mathrm{TCV}} = \sum_{r=1}^{t} \sum_{i \in T_r} \{y_i - M[x_i, \hat{\beta}^o(T_r)]\}^2. \qquad (14)$$

This procedure requires only as many least-squares evaluations as there are test sets. We suggested earlier (Brünger, 1992*a*) that for most crystallographic applications, it is sufficient to consider a single test set T and to compute the residual for the test set

$$L_{\mathrm{TCV}}(T) = \sum_{i \in T} \{y_i - M[x_i, \hat{\beta}_o(T)]\}^2 \qquad (15)$$

or the corresponding R value. The motivation behind equations (14) and (15) is to reduce the number of least-squares evaluations needed to compute the cross-validation residual. The statistical significance can be checked by repeating the procedure with different test sets. We expect that the method will work best with data sets which exhibit a high degree of redundancy, *i.e.* where the effect of omitting the test sets is small.

Fig. 2 illustrates cross validation for our example data set [equation (11)] using 10 test sets each containing 20 data points. One of the test sets T is shown in Fig. 2(*a*) and the remaining data (which we refer to as the working set) in Fig. 2(*b*). We borrowed the notion of test and training sets from the theory of neural networks (Hertz, Krogh & Palmer, 1991). The residuals $L_{\mathrm{TCV}}(T)$ for the test sets and the cross-validation residual L_{TCV} are shown in Figs. 2(*c*) and 2(*d*) respectively. The individual test residuals [equation (15)] show large statistical fluctuations as a result of the small size of the test set. Equation (14) therefore has to be applied in order to obtain statistically significant results. Fig. 2(*d*) clearly indicates that functions with $m > 2$ overfit the data. The statistical significance of L_{TCV} was estimated by repeating the whole procedure 20 times (Figs. 2*e* and 2*f*).

3.3. *Cross validation in macromolecular crystallography*

The fitting of appropriate electron-density models to the observed diffraction data can be formulated as a least-squares problem [equation (2)]. The cross-validation method described in equation (10) is impractical as it would involve as many least-squares evaluations as there are observed unique reflections. Fortunately, crystallographic diffraction data exhibit a high degree of redun-

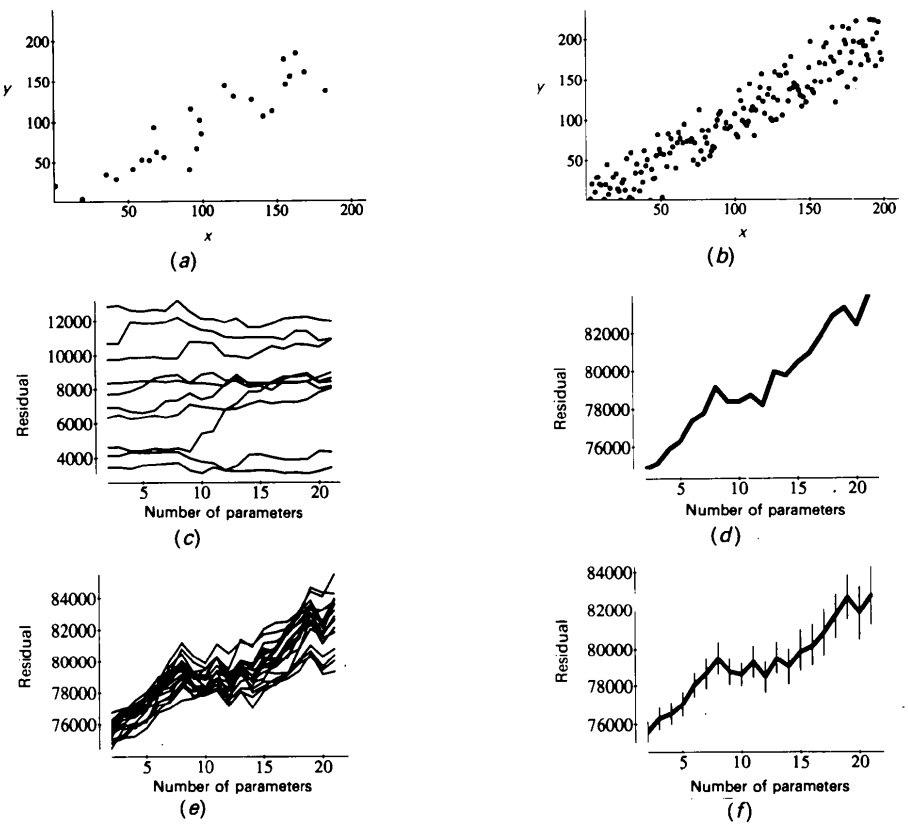


Fig. 2. Illustration of cross validation with test sets that partition the data set [equation (13)]. (*a*) Example of a test set obtained by randomly selecting 10% of the data from the data set. (*b*) The remaining 90% of the data set, referred to as the working set. (*c*) Residuals computed over the test sets after least-squares fits with the corresponding working sets [equation (15)]. (*d*) Cross-validation residual [equation (14)]. (*e*) 20 cross validations using different test sets for the data partitioning described in equation (13). (*f*) Mean and standard deviation (σ) for the 20 cross validations.

dancy, even for macromolecules. We can therefore afford to omit more than a single reflection for cross validation with test sets [equations (14) and (15)]. In practice, a certain fraction of reflections always have to be omitted as a result of measurement errors. Thus cross validation could be considered as using data-collection equipment that renders the reflections belonging to the test set unobservable.

In our previous work (Brünger, 1992*a*), we proposed the use of a single test set T for cross validation (Fig. 3). The remaining diffraction data comprises the working set A that is used for any crystallographic fitting procedure. We referred to the R value computed for the reflections belonging to T

$$R_T^{\text{free}} = \frac{\sum_{\mathbf{h}\in T} ||F_{\text{obs}}(\mathbf{h})| - k|F_{\text{calc}}(\mathbf{h})||}{\sum_{\mathbf{h}\in T} |F_{\text{obs}}(\mathbf{h})|} \qquad (16)$$

as the free R value. According to the principle of cross validation [equation (15)], the model has to be fitted to the working set A before the free R value can be evaluated. For instance, in the case of crystallographic refinement (Hendrickson, 1985) the model is fitted against

$$R'_A = \sum_{\mathbf{h}\in A} [|F_{\text{obs}}(\mathbf{h})| - k|F_{\text{calc}}(\mathbf{h})|]^2 \qquad (17)$$

The concept of partitioning the data into a test and a working set can be applied to any other statistical quantity describing the agreement between the atomic model and diffraction data, *e.g.* the standard linear correlation coefficient (Stout & Jensen, 1989). It is not restricted to refinement but can conceivably be applied to any optimization problem in crystallography such as density modification (Podjarny, Bhat & Zwick, 1987), molecular replacement (Hoppe, 1957; Rossmann & Blow, 1962; Brünger, 1990) or *ab initio* phasing (Hauptman, 1991; Weeks *et al.*, 1991).

In a different context, the partitioning of observed reflections into a basis set with known phases and a neighborhood set has been used for the multisolution strategy of phase determination by traditional direct methods (Woolfson, 1987) and by a combined maximization of entropy and likelihood (Bricogne, 1984; Bricogne & Gilmore, 1990). Recently, Karle (1991) proposed using 'rolling' working sets to aid the convergence behavior of least-squares minimization. Refined omit maps (Bhat & Cohen, 1984; Hodel, Kim & Brünger, 1992) can be viewed as the real-space analog to cross validation: part(s) of the model are omitted and then the remaining model is refined.

Cross validation with a single test set [equation (15)] as opposed to multiple sets [equation (14)] is computationally not more demanding than a single refinement round. Furthermore, R_T^{free} could be monitored during the process of model building and refinement by omitting the test set T right from the start and recording its behavior as the refinement progresses. Although the size of the test set T should be kept small in order to minimize the impact on the fitting procedure, it has to be large enough to produce a statistically well defined average for R_T^{free}. As we have previously shown for the penicillopepsin crystal structure, test sets that comprise about 10% of the observed unique reflections represent a good compromise between these two competing effects (Brünger, 1992*a*). To quantify this choice further, we calculated the mean and standard deviation of repeated R_T^{free} evaluations with different test sets T for the penicillopepsin structure. Table 1 reports a standard deviation of 0.5% for R_T^{free} which is probably acceptable for most purposes. It should be pointed out however that

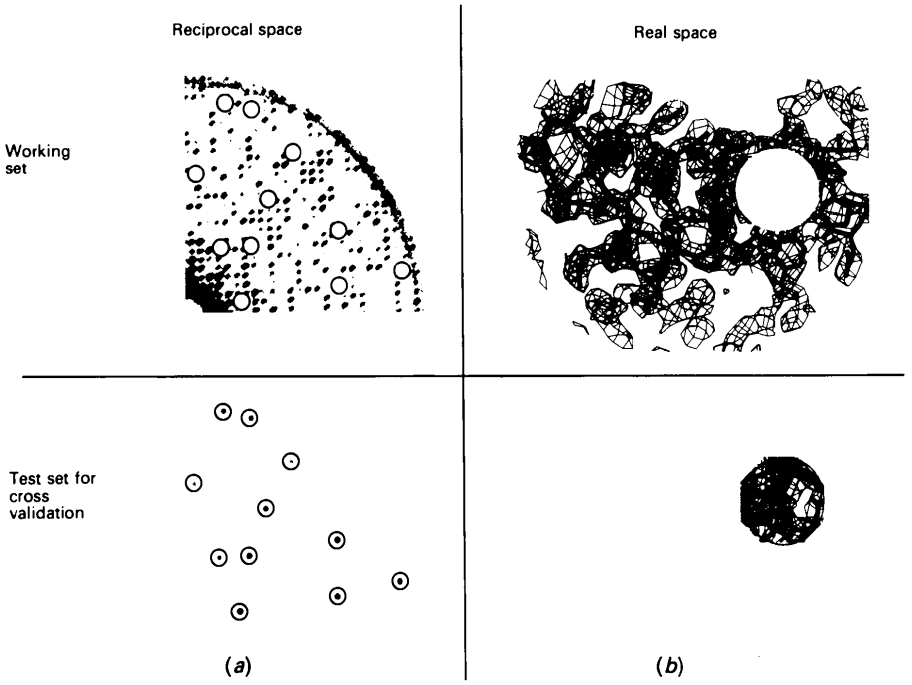


Fig. 3. The principle of cross validation with a single test set in crystallography. (*a*) Reciprocal space application; a number of observed reflections from an asymmetric unit are set aside for the test set and the agreement between the observed and the computed intensities of the test set is used for cross validation. (*b*) Real space application; a portion of the atomic model is omitted in the refinement process and the agreement between the refined density map and the model is checked in the omitted region.

Table 1. *Statistics for cross validation*

Statistics are given for R, R_T^{free}, the mean phase error ($\overline{|\Delta\Phi|}$) and the deviation of bond lengths (Δ_{bonds}) and bond angles (Δ_{angles}) from the ideal for simulated annealing refinements of penicillopepsin (Hsu *et al.*, 1977; James & Sielecki, 1983) at 6–1.8 Å resolution. The mean and the standard deviation (σ) were computed for a set of ten refinements using different working sets comprising 90% of the observed unique reflections. R_T^{free} was computed for the corresponding test sets. The weight w_x [equation (18)] was set to 200 000. The penicillopepsin crystal structure without water molecules and unit occupancy values was used as the starting point. Each refinement consisted of a slow-cooling protocol (Brünger, Krukowski & Erickson, 1990) starting at 1000 K, overall *B*-factor refinement and restrained individual *B*-factor refinement with target values for the temperature-factor deviations (Hendrickson, 1985) of 1.5, 2, 2 and 2.5 Å^2 for bonded backbone, angle-related backbone, bonded side-chain and angle-related side-chain atoms respectively.

Property	Mean	σ
R (%)	20.77	0.09
$\overline{\|\Delta\Phi\|}$ (°)	35.55	0.23
R_T^{free} (%)	26.41	0.47
Δ_{bonds} (Å)	0.007	0.0001
Δ_{angles} (°)	0.92	0.01

this is not a general rule, and the accuracy of R_T^{free} depends on the number of observed unique reflections; it is a function of the resolution of the diffraction data and the size of the asymmetric unit of the crystal. Averaging of R_T^{free} for multiple test sets may be required to decrease statistical fluctuations [*cf.* equation (14)]. It is clearly desirable to refine the final atomic model against all observed data after all R_T^{free} evaluations have been completed to remove any effect due to the omission of the test set T.

In this and our previous work (Brünger, 1992*a*), we elected to obtain the test sets by random selection from the unique set of observed reflections. It is conceivable that selections of reflections with high 'leverage' (Prince & Nicholson, 1985) could increase the sensitivity of the free R value. This will be the subject of future investigations.

4. Results and discussion

4.1. *Bias removal*

The principle of cross validation states that the free R value is evaluated *after* the model has been refined with the test set omitted from the diffraction data. Thus if a crystal structure has been refined with all diffraction data included, a procedure is required to remove the 'memory' of the test set. A refinement method with a large radius of convergence is necessary to accomplish this. Fig. 4 compares the ability of conjugate gradient minimization and simulated annealing to remove bias towards the test set. In both cases, the starting structure was the refined crystal structure of penicillopepsin (James & Sielecki, 1983) but with the crystal waters removed and the atomic B factors set to 15 Å^2. The positional refinements were followed by restrained *B*-factor refinements.

R and R_T^{free} begin to deviate immediately after the first few refinement steps. In the case of conjugate gradient minimization, R_T^{free} gradually increases while R and the mean phase difference ($\overline{|\Delta\Phi|}$) remain approximately constant (Fig. 4*a*); the rate of increase slows over the course of the minimization. In the case of simulated annealing, both R and R_T^{free} increase sharply as a result of 'heating' of the structure (Fig. 4*b*) and a slight worsening of the phase error ($\overline{|\Delta\Phi|}$) in the model is observed. All three quantities then show a slow average decrease as the simulated annealing refinement proceeds. The value of R_T^{free} after simulated annealing is around 30% (Fig. 4*b*) compared to just 28.4% (Fig. 4*a*) after 120 steps of conjugate gradient minimization. Subsequent thermal *B*-factor refinement improves R, R_T^{free} and $\overline{|\Delta\Phi|}$. One can conclude that an extended conjugate gradient refinement round should be sufficient to obtain a rough estimate of R_T^{free} but that a round of simulated annealing is required for better convergence of R_T^{free}.

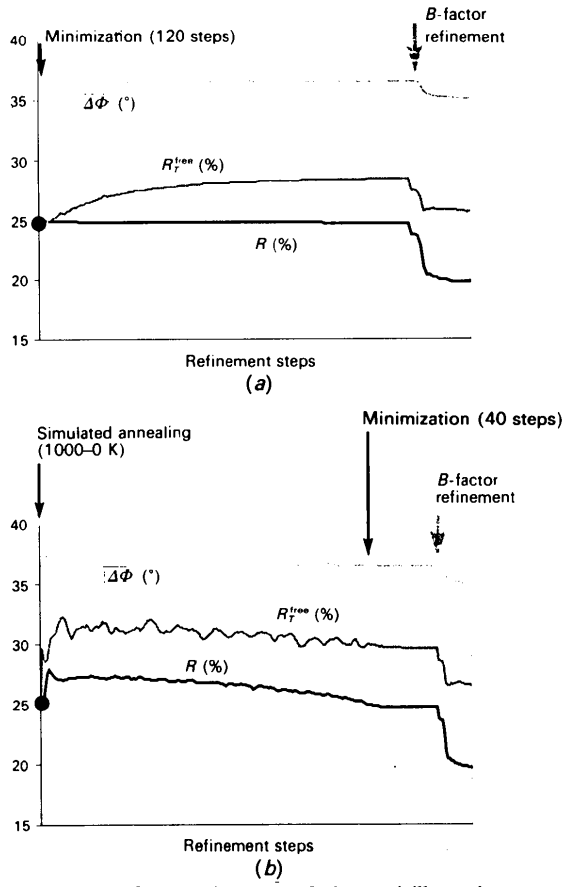


Fig. 4. Course of the refinement of the penicillopepsin structure against the working set which consists of a random selection of 90% of the observed unique reflections. The starting point for both refinements is the penicillopepsin structure refined against the full data set with all B factors set to 15 Å^2. (*a*) Minimization consisting of 120 steps of conjugate gradient refinement using the method of Powell (1977), followed by 20-step restrained *B*-factor refinement. (*b*) Simulated annealing using the slow-cooling protocol of Brünger *et al.* (1990) starting at 1000 K, cooling to 0 K, followed by 40-step conjugate gradient minimization, overall *B*-factor refinement and 20-step restrained *B*-factor refinement. The R and R_T^{free} values and the mean phase error ($\overline{|\Delta\Phi|}$) are shown. All refinements were carried out with the protein structure without ordered water molecules at 6–1.8 Å resolution.

4.2. *Objective choice of restraint weights for refinement*

We have already shown (Brünger, 1992*a*) that the R_T^{free} method can be used to optimize the overall weighting between diffraction data and chemical restraints in crystallographic refinements. For instance, restrained positional refinement (Jack & Levitt, 1978; Hendrickson, 1985; Brünger, 1991) makes use of a target function

$$E^{\text{pot}} = w_x R' + E_{\text{chemical}} \qquad (18)$$

where the weight w_x, *a priori* unknown, relates the residual R' to the chemical and geometric restraints E_{chemical}. If too small a value is chosen for w_x, too much emphasis is placed on the geometry provided in the dictionary of the refinement program, resulting in a poor fit to the diffraction data. If the chosen value is too large, the structure will become 'over-refined'; although the conventional R value is very small, the geometry of the structure becomes severely distorted. w_x is typically chosen according to predefined notions about the allowed geometric distortions of the structure. This choice is rather subjective and published macromolecular crystal structures show a large distribution of deviations from ideal geometry. Optimizing R_T^{free} as function of w_x is a more objective method to determine w_x. Furthermore, we have observed that this choice of w_x also maximizes the model's phase accuracy (Brünger, 1992*a*).

Fig. 5 illustrates the optimization of w_x for a different parameter set than the one previously used (Brünger, 1992*a*). We used the novel parameter set of Engh & Huber (1991) which is contained in the files 'tophcsdx.pro' and 'parhcsdx.pro' in *X-PLOR* (Brünger, 1992*b*). This parameter set differs from older parameter sets in *X-PLOR* [which were derived from the *CHARMM* energy functions (Brooks *et al.*, 1983)] in terms of bond length and bond angle constants. The target values for bond lengths and bond angles were derived from small-molecule crystal structures deposited in the Cambridge Structural Database (Allen, Kennard & Taylor, 1983). The weights or 'energy' constants were derived from the standard deviations of the bond lengths and bond angles assuming that the standard deviation represents expected deviations from harmonic potential minima at 293 K. Additional atom types had to be introduced to achieve a good fit to the experimental data. From a comparison of Fig. 5 with Fig. 3 of Brünger (1992*a*), it appears that the Engh & Huber (1991) parameters show wider minima for both R_T^{free} and $\overline{|\Delta\Phi|}$. Furthermore, the deviations from ideal geometry are significantly smaller for the optimal choice of w_x (the r.m.s. deviations of bond lengths and bond angles are 0.008 Å and 1° respectively). Thus the Engh & Huber parameters allow one to fit a model with surprisingly small deviations from ideal geometry.

Fig. 5 compares conjugate gradient minimization and simulated annealing in their ability to optimize R_T^{free} as a function of w_x. The minimum of R_T^{free} is in approximately the same position for both methods, although a small shift towards smaller values of w_x is observed for minimization (Fig. 5*a*). Simulated annealing produces a much more pronounced increase of R_T^{free} for small values of w_x. In this case, too little emphasis is put on the diffraction data and the simulated annealing process explores 'incorrect' conformations that cannot be reached by conjugate gradient minimization because they are separated by small energy barriers from the correct conformation. In practical terms, minimization can be used as a rough guide for R_T^{free} optimization although simulated annealing will provide more pronounced results. We therefore used simulated annealing to obtain R_T^{free} in the remainder of this paper. It should be noted that the empirical 'check' procedure in *X-PLOR* (Brünger, 1992*b*) typically produces values for w_x that are between two and three times too large compared with those given by the more objective method.

The overall weight w_x [equation (18)] is a special case of more general weights that apply to subsets of diffraction data or particular classes of restraints (such as bond lengths, bond angles and dihedral angles). One could imagine optimizing R_T^{free} as a function of all these individual weights. In general, this is a rather computer-intensive optimization problem as each point in multidimensional space would require a complete round of refinement to evaluate R_T^{free}.

We carried out several optimizations of R_T^{free} to assess the relative weightings of bond length, bond angle, di-

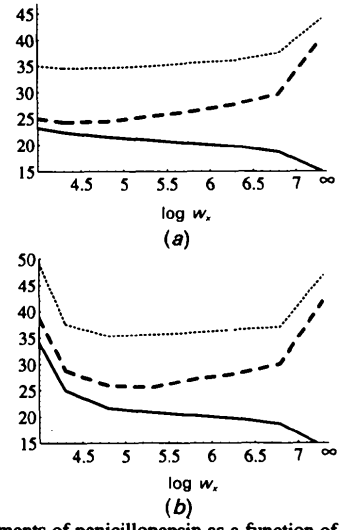


Fig. 5. Refinements of penicillopepsin as a function of $\log w_x$ [equation (18)] using cross validation. R_T^{free} (dashed line) was computed for the test set T which was obtained by a 10% random selection. R (solid line) was computed for the working set A. $\overline{|\Delta\Phi|}$ (dotted line) is the figure-of-merit-weighted mean phase difference between model phases and the most probable multiple-isomorphous-replacement phases at 6–2.8 Å resolution. $w_x = \infty$ represents the completely unrestrained case. (*a*) Minimization consisting of 120 steps of conjugate gradient steps using the method of Powell (1977), followed by 20-step restrained *B*-factor refinement. (*b*) Simulated annealing using the slow-cooling protocol of Brünger *et al.* (1990) starting at 1000 K, cooling to 0 K, followed by 40-step conjugate gradient minimization, overall *B*-factor refinement and 20-step restrained *B*-factor refinement. All refinements were carried out with the protein structure without ordered water molecules at 6–1.8 Å resolution.

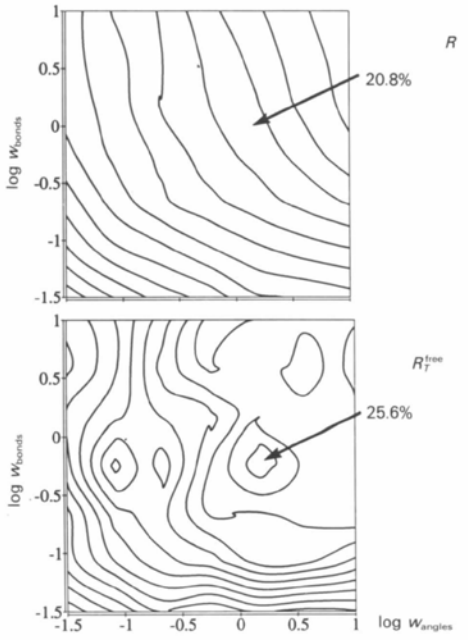
hedral angle and van der Waals restraints in the Engh & Huber (1991) parameter set. A two-dimensional optimization of R_T^{free} as a function of w_{bonds} and w_{angles} [equation (19)] is shown in Fig. 6 and a two-dimensional optimization of w_{conf} and w_{vdw} [equation (20)] in Fig. 7. Finally, a one-dimensional optimization of R_T^{free} as a function of w_{dihe} [equation (21)] is plotted in Fig. 8.

$$E^{\text{pot}} = w_x R' + w_{\text{bonds}} E_{\text{bonds}} + w_{\text{angles}} E_{\text{angles}} + E_{\text{vdw}} + E_{\text{dihe}} + E_{\text{improper}} \quad (19)$$

$$E^{\text{pot}} = w_x R' + w_{\text{vdw}} E_{\text{vdw}} + w_{\text{conf}}(E_{\text{bonds}} + E_{\text{angles}} + E_{\text{dihe}} + E_{\text{improper}}) \quad (20)$$

$$E^{\text{pot}} = w_x R' + E_{\text{bonds}} + E_{\text{angles}} + E_{\text{vdw}} + w_{\text{dihe}}(E_{\text{dihe}} + E_{\text{improper}}). \quad (21)$$

The w_x weight is always set to its optimal value as determined in Fig. 5, *i.e.* $w_x = 200\,000$. The functional forms of the various terms of the E^{pot} function are described below.



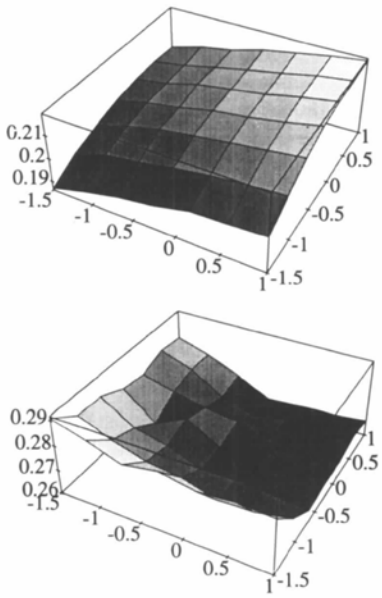
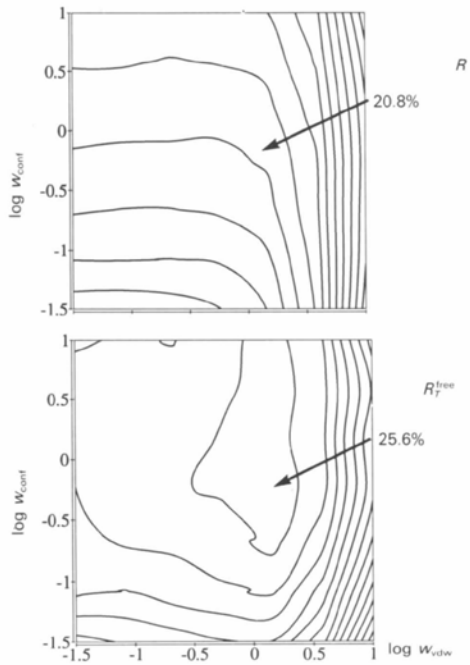


Fig. 6. Refinements of penicillopepsin as a function of $\log w_{\text{bonds}}$ and $\log w_{\text{angles}}$ [equation (19)] using cross validation. Contour plots and three-dimensional projections of R_T^{free} and R after simulated annealing refinement and restrained *B*-factor refinement are shown; these were computed for the test set T and the working set A. Details of the refinement procedure are identical to Fig. 5. The weights w_{conf} and w_{vdw} were modified by using the 'constraints interaction weight' command in *X-PLOR*.



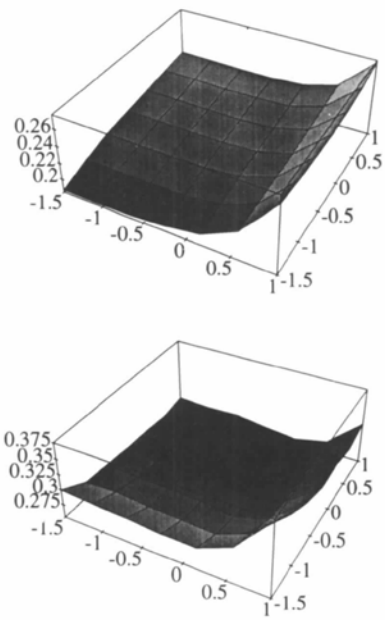


Fig. 7. Refinements of penicillopepsin as a function of $\log w_{\text{conf}}$ and $\log w_{\text{vdw}}$ [equation (20)] using cross validation. Contour plots and three-dimensional projections of R_T^{free} and R after simulated annealing refinement and restrained *B*-factor refinement are shown; these were computed for the test set T and the working set A. Details of the refinement procedure are identical to Fig. 5. The weights w_{conf} and w_{vdw} were modified by using the 'constraints interaction weight' command in *X-PLOR*.

The bond length and bond angle terms are given by

$$E_{\text{bonds}} = \sum_{b \in \text{bonds}} k_b (r - r_b^o)^2 \qquad (22)$$

$$E_{\text{angles}} = \sum_{a \in \text{angles}} k_a (\theta - \theta_a^o)^2 \qquad (23)$$

where k_a, k_b are constants, r is the actual bond length, r_b^o is the equilibrium bond length, θ is the actual bond angle and θ_a^o is the equilibrium bond angle. The functional form of these terms is identical for empirical energy calculations (Brooks *et al.*, 1983) and least-squares programs (Hendrickson, 1985). Fig. 6 reports a global minimum of R_T^{free} for values of w_{bonds} and w_{angles} that are close to unity. As the constants k_a, k_b, r_b^o, θ_a^o [equation (19)] are derived from observed small-molecule geometries (Engh & Huber, 1991), one can interpret this result in terms of the transferability of the uncertainties in bond length and bond angle geometry from small-molecule structures to macromolecular structures. R_T^{free} is less sensitive to over-weighting the bond or angle terms than to under-weighting (Fig. 6). The optimal choice corresponds to r.m.s. bond length and bond angle deviations of 0.008 Å and 1° respectively from the ideal.

The E_{vdw} term in equation (20) is a sum over Lennard-Jones potentials

$$E_{\text{vdw}} = \sum_{ij} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{R_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{R_{ij}} \right)^{6} \right] \qquad (24)$$

where the sum extends over all non-bonded atom pairs ij between atoms i and j. R_{ij} is the distance between the two atoms, ε_{ij} the minimum value of E_{vdw} and σ_{ij} the intersection of the Lennard-Jones potential with the R_{ij} axis. Fig. 7 shows a fairly broad minimal region. A significant increase of R_T^{free} occurs for an *over*-weighted van der Waals term or an *under*-weighted conformational term. This indicates that, unless it is too heavily weighted, the phase accuracy of the model is relatively insensitive to the exact choice of the van der Waals term. This result could be viewed as an empirical justification for the use of the simple repulsive functions described by Hendrickson (1985)

$$E_{\text{vdw}} = \sum_{ij} (1/s_{ij})(2^{1/6}\sigma_{ij} - R_{ij})^4. \qquad (25)$$

The dihedral angle terms $E_{\text{dihe}} + E_{\text{improper}}$ are used to restrain the planarity of certain groups of atoms, such as aromatic rings and peptide bonds, and to maintain the chirality of asymmetric carbon atoms,

$$E_{\text{dihe}} = \sum_{d \in \text{dihedrals}} k_d [1 + \cos(n_d \varphi + \delta_d)]$$

$$E_{\text{improper}} = \sum_{i \in \text{impropers}} k_i (\varphi - \delta_i)^2, \qquad (26)$$

where φ is the actual torsion angle and $k_d, k_i, \delta_d, \delta_i, n_d$ are appropriately chosen constants. The functional form of these restraints is obtained from the *CHARMM* empirical energy function (Brooks *et al.*, 1983), and differs from that used in restrained least-squares programs (Hendrickson, 1985). However, the effect of both functional forms is similar provided that the terms are properly weighted with respect to the remaining conformational terms. Fig. 8 shows R_T^{free} as a function of w_{dihe} [equation (21)]. As the differences involved are small, we tested their significance by carrying out ten independent R_T^{free} evaluations with different test sets. The minimum of R_T^{free} is approximately located at $w_{\text{dihe}} = 10^{0.75} \simeq 5.62$. This suggests that the dihedral angle terms are somewhat underweighted in the Engh & Huber (1991) parameter set.

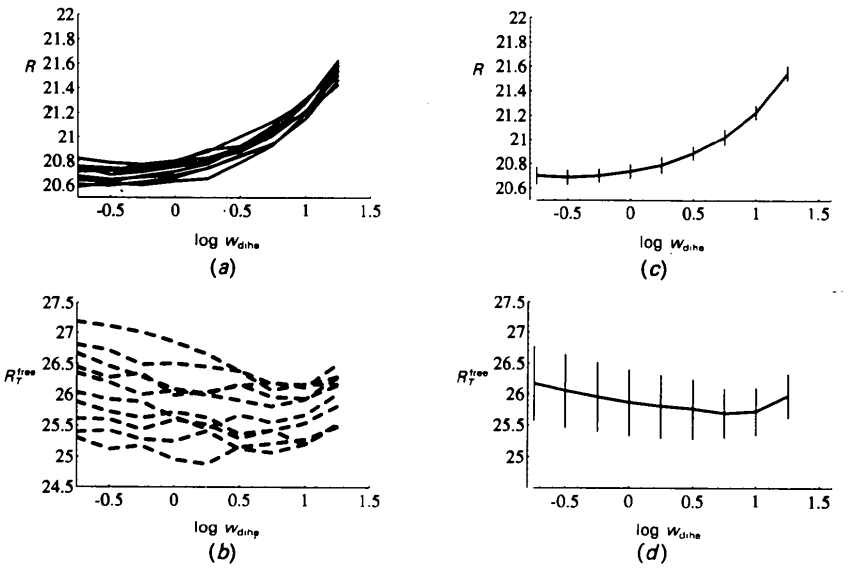


Fig. 8. Refinements of penicillopepsin as a function of $\log w_{\text{dihe}}$ [equation (21)] using cross validation with ten different working sets A and corresponding test sets T. Details of the refinement procedure are identical to Fig. 5. (*a*) R as a function of $\log w_{\text{dihe}}$ computed for the ten working sets. (*b*) R_T^{free} as a function of $\log w_{\text{dihe}}$ computed for the ten corresponding test sets. (*c*) Mean and standard deviation of R. (*d*) Mean and standard deviation of R_T^{free}.

4.3. *Influence of noise*

The deviation between R and R_T^{free} can be caused by noise in the data, incompleteness of the atomic model, or an unfavorable observable to parameter ratio. Fig. 9 illustrates the influence of noise on the difference between R and R_T^{free} as a function of the weight w_x [equation (18)]. An 'ideal' data set computed from a single structure of penicillopepsin (Fig. 9*a*) is compared to the ideal data set with the added noise (Fig. 9*b*) and to an ideal data set computed from a thermal ensemble of structures (Fig. 9*c*). For large values of w_x, the difference between R and R_T^{free} is much larger for the penicillopepsin data set (Fig. 5) and the ideal data set with noise (Fig. 9*b*) than for the ideal data sets without noise of a single structure (Fig. 9*a*) or of a thermal ensemble (Fig. 9*c*). Thus, the presence of Gaussian noise in the diffraction data is a possible explanation of the observed behavior of R and R_T^{free} (Fig. 5). This clearly does not rule out other explanations such as systematic errors during data collection, data set merging and reduction, or incompleteness of the atomic model, and the subject will need further investigation.

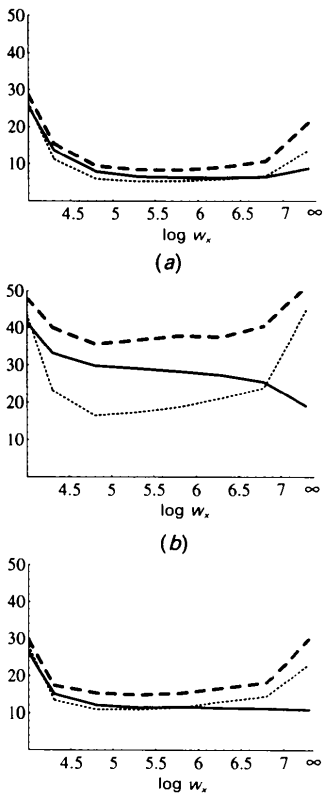


Fig. 9. Simulated annealing refinements of penicillopepsin as a function of $\log w_x$ [equation (18)] using cross validation for ideal data sets. Plots for R_T^{free} (dashed line), R (solid line) and $|\overline{\Delta\Phi}|$ (dotted line) after simulated annealing refinement and restrained B-factor refinement are shown. Details of the refinement procedure are identical to Fig. 5. (*a*) Using ideal diffraction data computed from the penicillopepsin crystal structure (Hsu *et al.*, 1977; James & Sielecki, 1983) with the F_{obs} amplitudes set to the F_{calc} amplitudes. (*b*) Using ideal diffraction data with Gaussian noise added to the structure-factor amplitudes and the standard deviation of the Gaussian distribution set to 10 units. (*c*) Using averaged structure factors computed from an ensemble of penicillopepsin structures; the ensemble was generated by a 5 ps molecular-dynamics simulation at 300 K with weak harmonic restraints applied to the initial C^{α} positions. Computation of the ideal data and all refinements were carried out with the protein structure without ordered water molecules at 6–1.8 Å resolution.

Thermal motion of the atoms appears to be modelled appropriately through thermal B-factor refinement resulting in the relatively small deviation between R and R_T^{free} in Fig. 9(*c*). Note that the R_T^{free} method probes global features of the modelling and refinement procedure. It is probably not sensitive enough to detect the inappropriate modelling of a few alternate conformations of surface side chains through thermal B factors (Kuriyan, Petsko, Levy & Karplus, 1986).

4.4. *Testing of unrestrained atomic models*

Unrestrained liquid-like models consisting of equal atomic scatterers have been suggested by Subbiah (1991) and Karle (1991) in the context of *ab initio* phasing. Sampling of a large number of configurations might be achievable as a result of the simplicity of these models. Successful application of liquid-like models is resolution dependent. Subbiah (1991) used predominantly low-resolution data (> 8 Å) while Karle (1991) used atomic resolution data (~ 1 Å). The application to resolution ranges typical for macromolecular crystal structures poses a problem; for example, Figs. 5 and 9 show that unrestrained refinement ($w_x \rightarrow \infty$) can lead to small R values at 1.8 Å resolution but relatively large phase errors ($|\overline{\Delta\Phi}|$). Furthermore, we have shown (Brünger, 1992*a*) that a random configuration of equal atoms can be refined to a very low conventional R value at 1.8 Å resolution. The question arises whether the 'landscape' of R for a configuration of unrestrained equal atoms has a single well defined global minimum (Fig. 10*a*) that is 'close' to the actual distribution of atoms in the crystal, or whether many false minima exist with equal or even lower R values than that of the correct distribution (Fig. 10*b*).

We probed the landscape of R by setting up unrestrained equal atoms, initially positioned at the heavy-atom

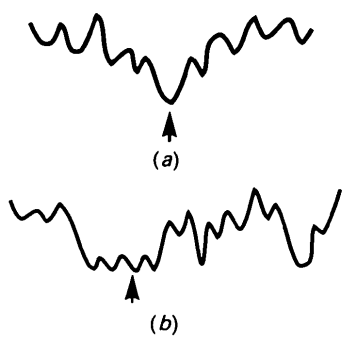


Fig. 10. Illustration of two possible landscapes for the R value of an unrestrained collection of equal atoms refined against the diffraction data as a function of a fictitious coordinate. The arrow indicates the correct configuration with maximum phase accuracy. (*a*) The R landscape provides a unique solution to the phase problem. (*b*) False minima exist with similar low R values and incorrect phases.

positions of the penicillopepsin crystal structure, and carrying out simulated annealing refinements at various temperatures. If the situation corresponded to Fig. 10(*a*), we would expect either to revert to the global minimum after simulated annealing or to be stuck in one of the higher R minima. Furthermore, we would expect to be trapped in increasingly higher minima as the initial simulated annealing temperature increased. In contrast if the situation corresponded to Fig. 10(*b*), we would expect to find an increasingly large number of 'false' minima with low R values corresponding to incorrect configurations of the unrestrained atoms.

Fig. 11 shows R, R_T^{free} and $\overline{|\Delta\Phi|}$ after simulated annealing refinements at temperatures between 3000 and 0 K. The penicillopepsin data set (Hsu *et al.*, 1977; James & Sielecki, 1983) and several ideal data sets with and without noise were used. The atomic r.m.s. differences between the unrestrained atoms and the crystal structure range from 0.6 Å for simulated annealing at 0 K to 4.5 Å at 3000 K. In our notation, simulated annealing at 0 K corresponds to conjugate gradient minimization. The shift of 0.6 Å for conjugate gradient minimization is remarkably large. Even when using the noise-free ideal data set, conjugate gradient minimization at 1.8 Å produces an r.m.s. shift in the positions of the atomic scatterers of about 0.3 Å. This implies that the exact atomic positions of the crystal structure do not correspond to a local minimum of the crystallographic residual for the unrestrained model of equal atoms. It appears that the R landscape for both the real and the ideal data set with noise is probably best described by Fig. 10(*b*); the higher the initial simulated annealing temperature, the larger the phase error of the refined distribution of unrestrained atoms while R stays approximately constant. The noise-free ideal data sets of both the single structure and the thermal ensemble exhibit a slight increase of R as the temperature is increased.

We define the 'correct' configuration of the atomic scatterers as the one that is obtained for a conjugate gradient refinement starting from the crystal structure coordinates. Despite an r.m.s. difference of 0.6 Å from the true non-hydrogen-atom positions of the crystal structure, the phase accuracy of this configuration is 45° (Fig. 11*a*) which would probably be sufficient for a successful solution of the crystal structure. R_T^{free} assumes a minimum for this configuration regardless of the presence of noise in the data (Fig. 11). Furthermore, R_T^{free} is correlated with the phase accuracy of the unrestrained equal-atom model. Thus, R_T^{free} provides a much better measure than R to distinguish between incorrect and correct configurations of the atomic scatterers.

5. Concluding remarks

We have shown in this paper and our previous one (Brünger, 1992*a*) that cross validation is a powerful concept for macromolecular crystallography which allows one to define more objective criteria for the correctness and accuracy of model representations of crystal structures. The free R value is a particular implementation of cross vali-

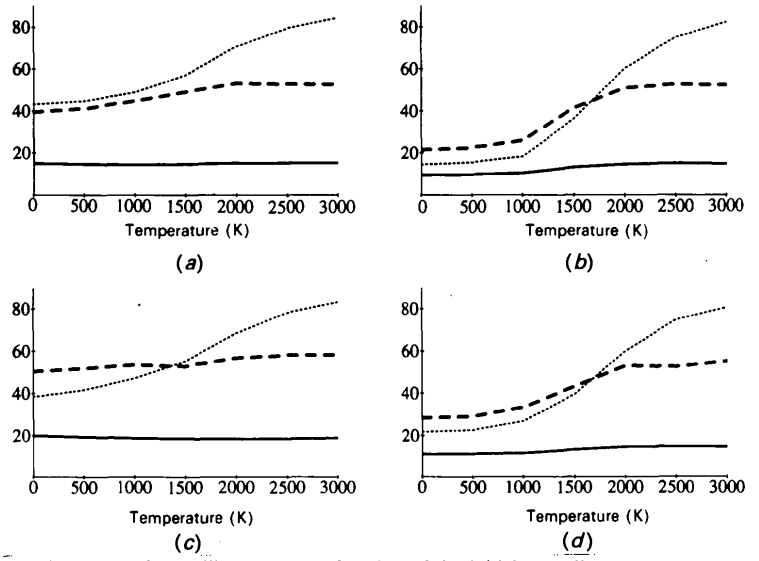


Fig. 11. Simulated annealing refinements of penicillopepsin as a function of the initial annealing temperature (Brünger *et al.*, 1990) using cross validation for various data sets. Plots are shown for R_T^{free} (dashed line), R (solid line) and $\overline{|\Delta\Phi|}$ (dotted line) after simulated annealing refinement and restrained B-factor refinement. Details of the refinement procedure are identical to Fig. 5. (*a*) Using the penicillopepsin diffraction data (Hsu *et al.*, 1977; James & Sielecki, 1983). (*b*) Using ideal diffraction data computed from the penicillopepsin crystal structure (Hsu *et al.*, 1977; James & Sielecki, 1983) with the F_{obs} amplitudes set to the F_{calc} amplitudes. (*c*) Using ideal diffraction data with Gaussian noise added to the structure-factor amplitudes and the standard deviation of the Gaussian distribution set to 10 units. (*d*) Using averaged structure factors computed from an ensemble of penicillopepsin structures; the ensemble was generated by a 5 ps molecular-dynamics simulation at 300 K with weak harmonic restraints applied to the initial C^{α} positions. Computation of the ideal data and all refinements were carried out with the protein structure without ordered water molecules at 6–1.8 Å resolution.

dation. Owing to the generality of the concept, it can in fact be applied to any statistic used both as a target in modeling the diffraction data and as a criterion to assess the quality of the modeling. Figures of merit or correlation coefficients in either reciprocal or real space are examples that could benefit from cross validation. Repeated application of cross validation can, in principle, be used to determine the statistical significance and probability distribution of the statistic without needing to derive difficult analytical formulae. Clearly, repeated cross validation involving refinements of high-resolution crystal structures is only possible for the crystallographer who has access to large computing resources. However, the intrinsic parallelism of repeated cross validation means that massively parallel computers should alleviate this problem in the near future.

In this paper we illustrate the application of the R_T^{free} method to verify the recent refinement parameter set of Engh & Huber (1991). We wanted to obtain the optimal relative weighting between bond length, bond angle, dihedral angle and van der Waals restraints. We found that the distribution of bond lengths and bond angles as found in the Cambridge Structural Database is in fact optimal for the penicillopepsin crystal structure at 1.8 Å resolution. The deviations from ideal geometry are surprisingly small (0.008 Å and 1° for bond lengths and bond angles respectively). However, this is a statement about the fit of a single structure to the time- and space-averaged diffraction data. It does not imply that the thermal motions of the macromolecule are small. We found that dihedral angle restraints are probably under-weighted and will have to be revised. The weighting of the van der Waals restraints is not critical over a wide range of values.

Our results suggest a project to verify parameter sets used in crystallographic refinement. Optimization of weights by minimizing R_T^{free} could be carried out for those crystal structures that have been deposited in the Brookhaven Data Bank (Bernstein *et al.*, 1977) together with their diffraction data sets. If the protein structure data base is sufficiently large, we expect that one could address very detailed questions, *e.g.* the relative weighting of subclasses of restraints such as the C—N bond length in peptide groups. In this way, one could optimize the chemical restraints such that crystal structures refined with them would exhibit maximum phase accuracy.

Unrestrained refinement of equal atoms against the 1.8 Å resolution data of penicillopepsin causes large phase errors, while the conventional R value is lower than that for the restrained crystal structure. In fact, many distributions exist with equally good R values but with arbitrarily large phase errors. This poses a problem for applications of liquid-like models of equal atomic scatterers (Subbiah, 1991; Karle, 1991) for *ab initio* phasing at resolution ranges typical for macromolecules. We generated a large number of configurations in the neighborhood of the crystal structure with atomic r.m.s. differences between 0.6 and 4.4 Å. Our results indicate that R_T^{free} is correlated with the phase accuracy of these configurations and that the correct configuration has the lowest R_T^{free} value, and we suggest using R_T^{free} as a 'score' to check configurations of the liquid. We anticipate that by choosing configurations with the lowest R_T^{free} values, one could significantly reduce the number of incorrect configurations. At the present time, the computational requirements of the R_T^{free} evaluation clearly preclude a naïve brute-force scoring of all possible configurations of the liquid in the asymmetric unit for macromolecules.

The author thanks M. N. G. James and A. R. Sielecki for providing the diffraction data and coordinates of the penicillopepsin structure, E. Fauman, M. Nilges, J. Jiangsheng, P. S. Sigler, T. Simonson and W. Weis for discussions, the National Science Foundation and the Pittsburgh Supercomputer Center for support (grants Nos. DIR9021975 and DMB8900008P), and Dr J. Erickson for allowing generous use of the CRAY-YMP at the National Cancer Institute in Frederick, Maryland.

References

ALLEN, F. H., KENNARD, O. & TAYLOR, R. (1983). *Acc. Chem. Res.* **16**, 146-153.

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.

BHAT, T. N. & COHEN, G. H. (1984). *J. Appl. Cryst.* **17**, 244-248.

BRICOGNE, G. (1984). *Acta Cryst.* A**40**, 410-445.

BRICOGNE, G. & GILMORE, C. J. (1984). *Acta Cryst.* A**46**, 284-297.

BROOKS, B. R., BRUCCOLERI, R. E., OLAFSON, B. D., STATES, D. J. & SWAMINATHAN, S. (1983). *J. Comput. Chem.* **4**, 187-217.

BRÜNGER, A. T. (1990). *Acta Cryst.* A**46**, 46-57.

BRÜNGER, A. T. (1991). *Ann. Rev. Phys. Chem.* **42**, 197-223.

BRÜNGER, A. T. (1992*a*). *Nature (London)*, **355**, 472-474.

BRÜNGER, A. T. (1992*b*). *X-PLOR.* Version 3.0. Yale Univ., New Haven, USA.

BRÜNGER, A. T., KRUKOWSKI, A. & ERICKSON, J. (1990). *Acta Cryst.* A**46**, 585-93.

EFRON, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*, CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 38. Philadelphia: Society for Industrial and Applied Mathematics.

EFRON, B. (1988). *SIAM Rev.* **30**, 421-449.

EFRON, B. & TIBSHIRANI, R. (1988). *Science*, **253**, 390-395.

ENGH, R. & HUBER, R. (1991). *Acta Cryst.* A**47**, 392-400.

HAMILTON, W. C. (1965). *Acta Cryst.* **18**, 502-510.

HÄRDLE, W., HALL, P. & MARRON, J. S. (1988). *J. Am. Stat. Assoc.* **83**, 86-95.

HAUPTMAN, H. A. (1991). In *Crystallographic Computing* 5: *From Chemistry to Biology*, edited by D. MORAS, A. D. PODJARNY & J. C. THIERRY. Oxford Univ. Press.

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

HERTZ, J., KROGH, A. & PALMER, R. G. (1991). *Introduction to the Theory of Neural Computation. Lecture Notes*, Vol. I, *Santa Fe Institute in the Science of Complexity.* Redwood City: Addison-Wesley.

HINKLEY, D. V. (1988). *J. R. Stat. Soc. Ser. B*, **50**, 321-337.

HODEL, A., KIM, S.-H. & BRÜNGER, A. T. (1992). *Acta Cryst.* A**48**. In the press.

HOPPE, W. (1957). *Acta Cryst.* **10**, 750-751.

HSU, I.-N., DELBARE, L. T. J., JAMES, M. N. G. & HOFMANN, T. (1977). *Nature (London)*, **266**, 140-145.

JACK, A. & LEVITT, M. (1978). *Acta Cryst.* A**34**, 931-935.

JAMES, M. N. G. & SIELECKI, A. R. (1983). *J. Mol. Biol.* **163**, 299-361.

JONES, T. A. (1978). *J. Appl. Cryst.* **11**, 268-272.
KARLE, J. (1991). *Proc. Natl Acad. Sci. USA*, **88**, 10099-10103.
KURIYAN, J., PETSKO, G. A., LEVY, R. M. & KARPLUS, M. (1986). *J. Mol. Biol.* **190**, 227-254.
PODJARNY, A. D., BHAT, T. N. & ZWICK, M. (1987). *Annu. Rev. Biophys. Biophys. Chem.* **16**, 351-373.
POWELL, M. J. D. (1977). *Math. Program.* **12**, 241-254.
PRESS, W. H., FLANNERY, B. P, TEUKOLSKY, S. A. & VETTERLING, W. T. (1986). *Numerical Recipes, The Art of Scientific Computing*, pp. 498-504. Cambridge Univ. Press.
PRINCE, E. & NICHOLSON, W. L. (1985). In *Structure and Statistics in Crystallography*, edited by A. J. C. WILSON. New York: Adenine Press.
ROSSMANN, M. G. & BLOW, D. M. (1962). *Acta Cryst.* **15**, 24-31.
STONE, M. (1974). *J. R. Stat. Soc. Ser. B*, **36**, 111-147.
STOUT, G. H. & JENSEN, L. H. (1989). In *X-ray Structure Determination, a Practical Guide*, pp. 229-230. New York: John Wiley.
SUBBIAH, S. (1991). *Science*, **252**, 128-133.
WEEKS, C. M., DETITTA, G. T., MILLER, R. & HAUPTMAN, H. A. (1993). *Acta Cryst.* D**49**, 179-181.
WOOLFSON, M. M. (1987). *Acta Cryst.* A**43**, 593-612.