Protein Science (1995), *4*:2392–2404. Cambridge University Press. Printed in the USA.

Significance of structural changes in proteins: Expected errors in refined protein structures

ROBERT M. STROUD AND ERIC B. FAUMAN[1]

Department of Biochemistry and Biophysics, University of California–San Francisco, San Francisco, California 94143-0448

(RECEIVED June 6, 1995; ACCEPTED September 11, 1995)

Abstract

A quantitative expression key to evaluating significant structural differences or induced shifts between any two protein structures is derived. Because crystallography leads to reports of a single (or sometimes dual) position for each atom, the significance of any structural change based on comparison of two structures depends critically on knowing the expected precision of each median atomic position reported, and on extracting it for each atom, from the information provided in the Protein Data Bank and in the publication. The differences between structures of protein molecules that should be identical, and that are normally distributed, indicating that they are not affected by crystal contacts, were analyzed with respect to many potential indicators of structure precision, so as to extract, essentially by "machine learning" principles, a generally applicable expression involving the highest correlates. Eighteen refined crystal structures from the Protein Data Bank, in which there are multiple molecules in the crystallographic asymmetric unit, were selected and compared. The thermal *B* factor, the connectivity of the atom, and the ratio of the number of reflections to the number of atoms used in refinement correlate best with the magnitude of the positional differences between regions of the structures that otherwise would be expected to be the same. These results are embodied in a six-parameter equation that can be applied to any crystallographically refined structure to estimate the expected uncertainty in position of each atom. Structure change in a macromolecule can thus be referenced to the expected uncertainty in atomic position as reflected in the variance between otherwise identical structures with the observed values of correlated parameters.

Keywords: accuracy; *B* factor; conformation change; crystallography; errors; positional difference; protein structure

Structural differences between macromolecules can best be evaluated as to significance by reference to the expected distribution of uncertainty, or positional variations in regions that are compared. Such variations differ widely for different regions of protein structure, as reflected in electron density maps and deduced thermal factors, and depend on connectivity of the atom and applied constraints, resolution of the analysis, number of observations, and method of refinement and other factors. Structures determined by NMR are often represented as a manifold that are consistent with the data because the errors in closely coupled distances and angles are cumulative for regions of sequence that are separated by longer through-bond distances. Here we focus on structures of proteins as determined by X-ray crystallography. We derive a readily accessible calculation that best predicts the expected positional uncertainty for any atom in any particular protein structure determination, from the information on the refined structure readily available in the Protein Data Bank format, or from publications of the structure analysis.

There is information that directly pertains to positional variance in the course of crystallographic refinement, the resolution

Reprint requests to: Robert M. Stroud, S-960 Department of Biochemistry and Biophysics, University of California–San Francisco, San Francisco, California 94143-0448; e-mail: stroud@msg.ucsf.edu.

[1] Present address: Biophysics Research Division, IST 1208 Box 2099, University of Michigan, Ann Arbor, Michigan 48109-2099.

Abbreviations: Δx, Δy, Δz, difference in position of a single atom between a pair of structures along the x, y, or z axis, respectively; χ, generalized one-dimensional axis that represents the average over all possible orientations; σ_χ, standard deviation in one dimension of the Gaussian portion of positional differences between a pair of structures; Δr, the difference distance in position between atoms in different structures $\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}$; σ_r, standard deviation of the Maxwellian distribution of positional differences in a pair of structures: $\sigma_r = \sqrt{3}\sigma_\chi$; $\Delta\mathfrak{R}$, distance between the observed position of an atom and its "true" position; $\sigma_{\mathfrak{R}}$, standard deviation of the Maxwellian distribution of $\Delta\mathfrak{R}$: $\sigma_{\mathfrak{R}} = (\sqrt{2}/2)\sigma_r$; ATOM, number of independently refined atom positions in the asymmetric unit; REFL, number of independent reflections used in refinement; $\epsilon_\chi(B,\text{ATOM/REFL})$, empirically derived estimate of σ_χ as a function of B factor and the ratio of ATOM/REFL for a given structure; N.E.S.(subset), normalized error score, defined as the deviation from $\epsilon_\chi(B,\text{ATOM/REFL})$ for a selected subset of atoms; N.E.S.$_i$(subset), normalized error score calculated using only structure i of the 18 structures used in the analysis; σ_{NES}(subset), standard deviation of the 18 values for N.E.S.$_i$(subset).

of the structure, the constraints used, and other factors. Such variances can be estimated directly from the gradient and curvature in electron density maps (Chambers & Stroud, 1977), however, these are not readily restored from information deposited in the Protein Data Bank. Because various constraints toward ideality of geometry may be applied during the crystallographic analysis, positional deviations may be further reduced from these values as they would be determined for unconstrained atoms. To identify the most significant relationships among the more readily available descriptors, we compare structures that should otherwise be the same within the limitations of the methods used to determine their structures, and extract an empirical relationship that relates these differences in position to those most relevant factors necessary to produce an estimate of uncertainty. To obtain statistics, we identified structures within the Protein Data Bank that are expected to be identical and compared them with respect to a variety of possible parameters that might be expected to reflect inherent flexibility, or uncertainty of different sites. There are several cases of structures determined by more than one group (Chambers & Stroud, 1979; Clore & Gronenborn, 1991) or in different crystal forms (Kossiakoff et al., 1992). These show a relationship between differences in structure and the reported resolution, and thermal B factors (Chambers & Stroud, 1979). Differences between closely related protein structures also have been analyzed to extract probable errors (Chothia & Lesk, 1986). Here we derive a more general expression from multiple comparisons that can be applied to many structures.

Crystal structures that have more than one molecule independently arranged in the asymmetric unit represent a particularly rich source of information on accuracy and plasticity in crystallography because, within each structure, variables such as crystallizing conditions, primary sequence, crystal habit, data collection strategy, resolution of the intensities, and refinement methodology are the same for both molecules. We compare 18 pairs of structures with multiple, independently refined molecules in the asymmetric unit (Table 1). The differences in the structures were parameterized first with respect to B factor. After testing a number of potential indices of model quality for correlation with errors, a factor containing the ratio of parameters to observations used in the refinement was found to correlate best with the observed differences in position. The empirical formula can be applied to any refined macromolecular structure to obtain an estimate for the expected precision of each atom position. These variations will include the errors in each set of coordinates and expected plastic accommodations within the core regions of covalently identical protein molecules. The effect of other atomic attributes were evaluated for their relationship to positional differences and an overall equation was derived that allows a quantitative estimation of the uncertainty in position of any atomic position listed in the Protein Data Bank.

Results

Deriving the empirical error curve, $\epsilon_x(B)$

Eliminating true structural differences in extracting errors

There are real differences between multiple molecules in the asymmetric unit due to differences in the packing environment

Table 1. *Structures used in the analysis*

PDB code[b]	Name of protein	Resolution (Å)	R factor (%)	No. mol. in asym. unit	No. indep. reflect.	No. atoms in asym. unit	REFL/ATOM
1THB	T state of hemoglobin	1.50	19.6	2	87,000	4,874	17.8
2CCY	Cytochrome c'	1.67	18.8	2	30,533	2,146	14.2
4CHA	α-Chymotrypsin	1.68	23.4	2	35,274	3,591	9.8
4DFR	Dihydrofolate reductase	1.70	15.5	2	32,554	3,041	10.7
2HHB	Deoxy-hemoglobin	1.74	16.0	2	56,287	4,779	11.8
3CYT	Tuna cytochrome c (oxidized)	1.80	20.8	2	16,831	1,743	9.7
2AZA	Azurin	1.80	15.7	2	21,980	2,263	9.7
1GD1	Glyceraldehyde 3P dehydrogenase	1.80	17.7	4	93,120	10,984	8.5
1AZA	Azurin	2.00	19.0	2	15,614	2,133	7.3
1GP1	Glutathione peroxidase	2.00	18.6	2	26,564	3,102	8.6
1HMQ	Hemerythrin	2.00	17.3	4	40,422	4,296	9.4
2PKA	Kallikrein A	2.05	22.0	2	35,500	3,456	10.3
2PFK	Phosphofructo-kinase	2.40	16.8	4	59,481	9,371	6.3
1FCB	FlavocytochromeB2	2.40	18.8	2	61,365	6,948	8.8
4MDH	Malate dehydrogenase	2.50	16.7	2	22,910	5,675	4.0
4ATC	Aspartate transcarbamylase	2.60	24.0	2	26,912	7,620	3.5
1FC1	Immunoglobulin IGG	2.90	22.0	2	10,342	3,182	3.3
1HBS	Deoxyhemoglobin S	3.00	25.4	4	17,662	9,104	1.9

[a] The four-letter code refers to the PDB designation for each structure, for which the references are: 1THB (Waller & Liddington, 1990); 2CCY (Finzel et al., 1985); 4CHA (Tsukada & Blow, 1985); 4DFR (Bolin et al., 1982); 2HHB (Fermi et al., 1984); 3CYT (Takano & Dickerson, 1980); 2AZA (Baker, 1988); 1GD1 (Skarzynski et al., 1987); 1AZA (Norris et al., 1983); 1GP1 (Epp et al., 1983); 1HMQ (Stenkamp et al., 1982); 2PKA (Bode et al., 1983); 2PFK (Rypniewski & Evans, 1989); 1FCB (Xia & Mathews, 1990); 4MDH (Birktoft et al., 1989); 4ATC (Ke et al., 1984); 1FC1 (Deisenhofer, 1981); 1HBS (Padlan & Love, 1985).

of the independent molecules. To examine differences that have random distribution, all those differences that followed a normal or Gaussian distribution were extracted. The normally distributed differences between these pairs of structures are regarded here as due to the random errors in the structures. Crystal contacts and other systematic differences were screened by fitting a Gaussian to the observed distribution of one-dimensional differences.

In the first step, each structure was examined for any possible dependence of errors upon B factors. In each structure comparison, a scatter plot was constructed of Δr (the difference in atomic position in the pair of structures) versus the mean B factor assigned to the atom in the two structures (Fig. 1). Running bins in B factor were constructed with a width of ± 2 Å^2 and an increment of 0.1 Å^2. The distribution of Δr values within a single B factor bin can be displayed as a histogram, as in Figure 2. This histogram has the form of a Maxwellian distribution (Chamber & Stroud 1979):

$$P(\Delta\Re) = \frac{\Delta\Re^2}{\sigma_{\Re}^3}\sqrt{\frac{54}{\pi}}\, e^{-(3/2)(\Delta\Re^2/\sigma_{\Re}^2)} \qquad (1)$$

where $P(\Delta r)$ is the probability of obtaining a given value of Δr, and σ_r is the three-dimensional standard deviation of Δr.

In order to work with a Gaussian distribution representing positional variation in one dimension, rather than a Maxwellian distribution representing the scalar positional variation in three dimensions (Chambers & Stroud, 1979), each value of Δr was replaced with its one-dimensional probability distribution for Δx. That is, a given value of Δr has a certain probability of having a Δx component with a given value (derived in Appendix 1). The probability of a specific value of Δx is:

$$P(\Delta X) = \begin{cases} \dfrac{1}{\Delta R}; & 0 \le \Delta X < \Delta R \\ 0; & \Delta R \le \Delta X. \end{cases} \qquad (2)$$

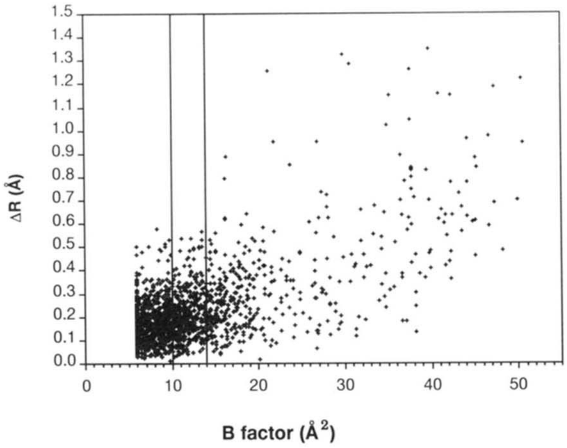

Fig. 1. Scatter plot of the difference in position (ΔR) of an atom in the two molecules in the asymmetric unit of the 1GP1 structure, after superpositioning, as a function of the average atomic B factor assigned to the atom in the two molecules. The vertical bars indicate the atoms with a mean B factor of 12 ± 2 Å^2.

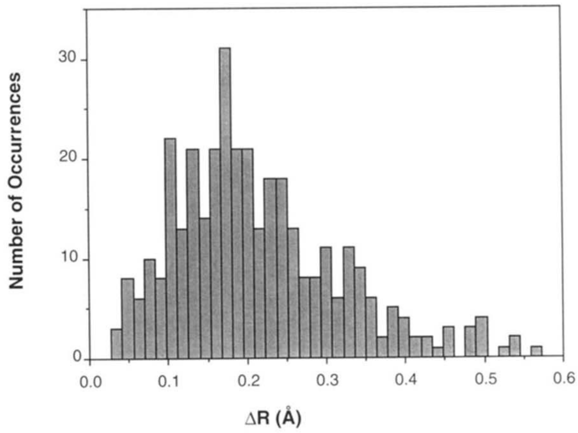

Fig. 2. Histogram of the distribution of ΔR values for atoms in the 1GP1 structure with a mean atom B factor of 12 ± 2 Å^2. The bin size is 1/17 of the RMS value of the ΔRs.

The generic one-dimensional axis is denoted as χ, which can be thought of as the X axis rotated over all possible orientations. Thus, this conversion from three dimensions to one dimension simultaneously provides the advantage of dealing in a one-dimensional variable where Gaussian (Fig. 3) rather than Maxwellian (Fig. 2) distributions hold, while avoiding the arbitrariness of any given orientation of the coordinate basis vectors.

For each range of B factor (e.g., 10–14 Å^2 as in Figs. 2 and 3), a histogram was constructed of frequency versus one-dimensional difference in position (Fig. 3). The standard deviation of the Δr values, σ_r, can be estimated as the RMS value of the Δr. Because σ_χ, the one-dimensional standard deviation, is related to σ_r by $\sigma_r = \sqrt{3}\sigma_\chi$, the abscissa of the histogram was binned in divisions of 1/17 of the RMS value of the Δr values at the given B factor, or roughly 1/10 of the expected σ_χ.

A Gaussian distribution of the form:

$$P(\Delta\chi) = Ae^{-(1/2)(\Delta\chi/\sigma\chi)^2} \qquad (3)$$

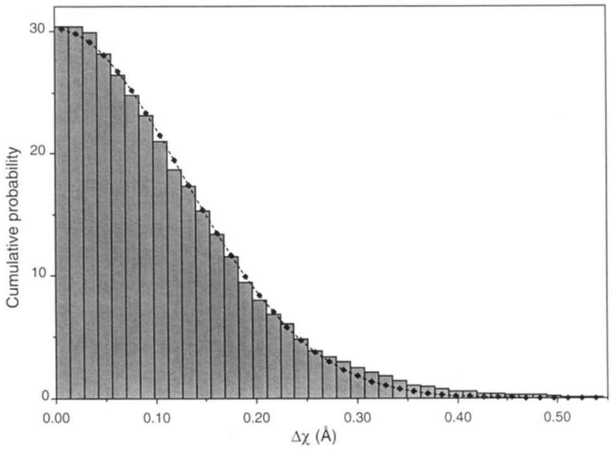

Fig. 3. Histogram of $\Delta\chi$ values obtained from the ΔR values in Figure 2. The dashed line indicates the best fit Gaussian curve, where the standard deviation is $\sigma_\chi(12.0)$.

was then fit to the histogram through nonlinear, unweighted least-squares minimization (Fig. 3). This extracts the true normal distribution component of the differences in atomic positions. The standard deviation of this distribution is termed $\sigma_\chi(B)$, i.e., the one-dimensional standard deviation in atomic position associated with a given B factor. The scatter plot of ΔR versus B factor (Fig. 1) is thus replaced with a curve of $\sigma_\chi(B)$ versus B factor (Fig. 4).

Consistent with the hypothesis that $\sigma_\chi(B)$ reflects the errors in the crystal structure is the observation that the extracted differences diminish as the resolution increases. This is true, for example, in the case of azurin (see below), for which two structures, a medium-resolution and a high-resolution structure, were used. The differences between the two molecules decreased with the addition of the high-resolution data. Other measures of crystal structure accuracy, such as dihedral angle quality and energy of hydrogen bonds, also improve with increasing resolution (Morris et al., 1992).

Errors are correlated with atom B factor

To generate a smooth dependence of $\sigma_\chi(B)$ upon B, it was fitted to an exponential of the form

$$\sigma_\chi(B) = a + b*e^{(B/c)} \tag{4}$$

where a, b, and c are the refinable parameters (Fig. 4). This functional form for the dependence on B was selected over the parabolic form previously used ($\sigma_\chi(B) = a + b*B + c*B^2$ [Perry et al., 1990]) because the exponential used here is monotonically increasing, whereas the second-order dependence was not, and because the exponential fit to the $\sigma_\chi(B)$ curves generally resulted in a smaller least-squares value than the corresponding parabolic equation fit to the same curve. The values of a, b, and c (Equation 4) for each protein were refined by nonlinear least-squares minimization to data from all atoms with B factors between 0 Å^2 and 40 Å^2. A B factor cutoff of 40 Å^2 was imposed because in most cases there were not enough data points ($n < 100$) to obtain reliable estimates of $\sigma_\chi(B)$ for values greater than 40 Å^2. This is not a great limitation, because 90% of the 70,000 atoms used in the analysis had B factors less than 40 Å^2. In some cases, proteins did not have enough atoms for the curve to extend all the way to 40 Å^2. In these cases, the exponential was fit to the reduced range and the limitation on range was noted. In addition, any limitation on the range for very small B factors was also noted for each structure. The B factor limitations for each structure are apparent in Figure 7. In this manner, each plot of $\sigma_\chi(B)$ versus B factor was represented by the three parameters a, b, and c.

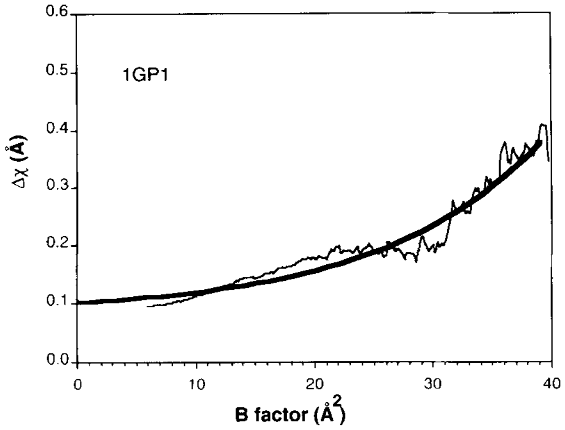

Fig. 4. Differences between the two molecules in the asymmetric unit of the 1GP1 structure as a function of B factor. The thin curve represents 341 values for $\sigma_\chi(B)$ (from $B = 6.0$ Å^2 to $B = 40.0$ Å^2 in steps of 0.1 Å^2). The thick curve is the best-fit three-parameter exponential function to these data points.

Previous correlations have shown that expected positional errors should and do increase with increasing B factor (Cruickshank, 1949; Chambers & Stroud, 1979; Bott & Frane, 1990; Perry et al., 1990), although the current report is the first to use the empirically derived three-parameter exponential form: $a + b*e^{(B/c)}$. In particular, Cruickshank derived the following formula for the one-dimensional standard deviation of uncertainty in atomic position:

$$\sigma_x = \frac{\sigma(A_h)}{A_{hh}}, \tag{5}$$

where σ_x is the standard deviation in the x direction for an orthorhombic space group, $\sigma(A_h)$ is the standard deviation of the first derivative (with respect to x) of the electron density in an F_o map, and A_{hh} is the second derivative (with respect to x) of the electron density, or the curvature, at the atom center. Because an atom with a larger B factor will have a smaller curvature, the Cruickshank formula predicts that atoms with larger B factors will have larger positional errors. Although no analytic expression was attempted for this B factor dependence, an error curve generated by the Cruickshank formula can be fit very well by a three-parameter exponential, as we use here.

Correlation of errors with measures of model quality

The dependence of errors on B factor for each structure is contained in the values of a, b, and c for that structure. However, the values of a, b, and c are different for each structure, the $\sigma_\chi(B)$ curves are all different (see Fig. 7), and it is apparent that a single exponential curve does not suffice for all the structures. Other factors, such as resolution, which do not affect the atomic B factors, do affect the accuracy of the atomic positions.

In order to generate a family of exponential curves, we sought a parameter relating the different curves obtained for the different structures. To examine the dependence of differences between structures on the quality of the model, the value of $\sigma_\chi(B)$ for each structure *at a given B factor* was plotted versus each of 90 different potential indices of model quality. These included such parameters as resolution, R factor ($R = \sum|(|F_o| - |F_c|)|/\sum|F_o|$), number of independent reflections, number of refined atom positions, and functions of these. For each index of model quality, the correlation coefficient (r) was evaluated from a linear least-squares fit to the plot to identify the parameters that correlate best. The functions of most appropriate parameters were expressed as a linear dependence on B factor, described by a Slope(B) and an Intercept(B), which were determined by least-squares refinement and are each functions of B factor.

Due to the limitations noted above, the $\sigma_\chi(B)$ versus B factor curves for all 18 structures exist simultaneously only in the range 14–31 Å^2, which accounts for 64% of the atoms used. The index of quality with the highest average correlation coefficient in this range is $e^{(-2\mathrm{ATOM/REFL})}$, where REFL is the total number of reflections used in refinement, and ATOM is the total number of atoms in the asymmetric unit subject to refinement (Fig. 5). This index yielded an average correlation coefficient of 0.89, and ranged from a maximum of 0.94 at a B factor of 14 Å^2 to a minimum of 0.76 at a B factor of 31 Å^2. For 18 data points, a correlation coefficient of 0.6 is statistically significant at the 99% confidence limit. That is, a random collection of 18 points has only a 1% chance of yielding a correlation coefficient greater than 0.6.

In this study, the errors in a crystal structure were more closely correlated with the ratio of the number of atoms to the number of reflections than to the resolution of the structure (average correlation coefficient of 0.79). As demonstrated in Appendix 2, however, this ratio is proportional to the cube of the resolution, multiplied by the protein fraction in the unit cell. Thus, trial of the many different indicators permitted testing of more complex relationships or interrelationships between parameters.

The dependence of positional accuracy on the ratio ATOM/REFL can be understood in terms of the overdeterminancy of the crystal structure refinement, that is, the ratio of observations to parameters. In an unrestrained crystallographic refinement, the observations are the intensities of independent reflections and the free parameters are the x, y, z positions and B factors of the atoms in the asymmetric unit. In macromolecular crystallography, the number of atoms involved often prohibits completely unrestrained refinement, so restraints and constraints are applied. The number of structure analyses where restraints have been removed is few, though these structures, like trypsin when refined by difference Fourier methods and without constraints (Chambers & Stroud, 1977), offer a rich source of data. These data ultimately become built into useful restraints and constraints on structure and coupling of B factors designed to impose expected structural features. The various uses of these restraints and constraints during refinement of a particular structure make it impossible to calculate the precise overdeterminancy in macromolecular refinement, but it is still related to the number of reflections and the number of atoms. The actual overdeterminancy will be related to the exact number and nature of the restraints and constraints employed in the refinement.

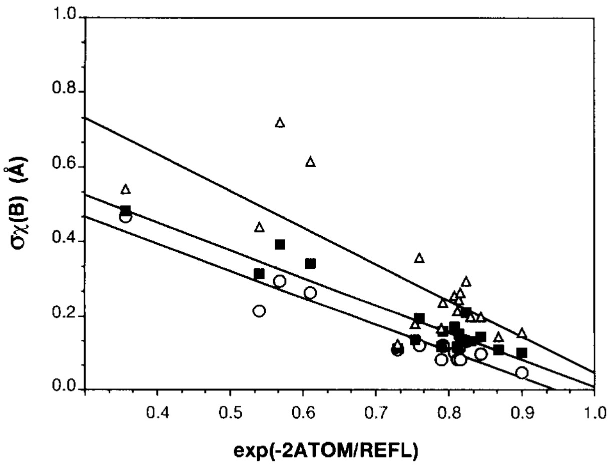

Fig. 5. The values of $\sigma_\chi(B)$ (from the smooth curve approximation) for the 18 structures in the study at three distinct B factors plotted as a function of exp(−2ATOM/REFL) for each structure. Circles, B factor = 10 Å^2; squares, B factor = 20 Å^2; triangles, B factor = 30 Å^2. Also indicated is the best-fit line to the data points at each B factor.

The presence of the solvent fraction in the relationship between resolution and ATOM/REFL suggests that, given the same protein in two different space groups, the one with the higher solvent content would yield the more accurate structure. This is because a higher solvent content implies a larger unit cell, consequently a greater number of reflections at a given resolution. In practice, however, crystals with a higher solvent content often tend to diffract to a lower maximum resolution, and both factors enter into the equation. Thus, use of the test for parameters that have the highest correlation is designed to reveal occult interrelationships such as these.

Perhaps surprisingly, the R factor of the structure did not correlate significantly with the level of errors observed (average correlation coefficient of 0.57, and below 0.60 in all B factor bins). This could be due in part to the different conventions used for reporting R factor. For example, some crystallographers apply a 2σ cutoff, that is, they may remove observations for which the condition $F/\sigma_F > 2.0$ is not met, which will produce a lower R factor than if no cutoff is used. Also, all the structures used were final reported structures, and the R factor is most useful in evaluating the progress of crystallographic refinement. Correspondingly, this implies that the results obtained in this study are only applicable to other structures that have been completely refined, and not structures still in refinement.

Calculation of expected errors correlated with B factor

Slope(B) and Intercept(B) are derived by linear least-squares analysis. The $\sigma_\chi(B)$ values for each structure are three-parameter exponential functions (Equation 4) and, as a result, Slope(B) and Intercept(B), which are related to the $\sigma_\chi(B)$ in a linear manner, can also be described by three-parameter exponential functions.

A plot of the Slope(B) versus B factor fits best to a three-parameter exponential curve:

$$\mathrm{Slope}(B) = k1 + k2 * e^{(B/k3)}, \tag{6}$$

where $k1 = -0.687$, $k2 = -0.00223$, and $k3 = 6.16$.

Likewise, a plot of Intercept(B) versus B factor yields an additional three parameters:

$$\mathrm{Intercept}(B) = k4 + k5 * e^{(B/k6)}, \tag{7}$$

where $k4 = 0.642$, $k5 = 0.00852$, and $k6 = 7.88$.

Thus, all the information relating the B factor of an atom to the accuracy of its position is contained in these six parameters.

The expected error at each B factor, $\epsilon_\chi(B)$, for a particular protein is then a function of the ratio of ATOM/REFL:

$$\epsilon_\chi(B,\mathrm{ATOM/REFL}) = \mathrm{Intercept}(B) + \mathrm{Slope}(B) * e^{(-2*\mathrm{ATOM/REFL})}, \tag{8}$$

where Intercept(B) and Slope(B) are defined by Equations 6 and 7.

For the analysis that follows, $\epsilon_\chi(B)$ for a single structure was recast as an exponential function of three variables:

$$\epsilon_\chi(B) = p1 + p2*e^{(B/p3)} \tag{9}$$

as described in Appendix 3.

This gives rise to a family of exponential curves of expected error versus B factor each at a different value of ATOM/REFL (Fig. 6). Figure 7A–R shows the observed dependence of positional differences – or "errors" – in structure, $\sigma_\chi(B)$, for the 18 structures along with the predicted error curves, $\epsilon_\chi(B)$, calculated according to the above equation.

Tests of the function in predicting expected "errors"

Because the curves were constructed primarily from the 60% of atoms with B factors between 14 and 31 Å^2, the resulting error curves were tested to see how well they predicted the differences between the structures used to derive them. To evaluate how well the function $\epsilon_\chi(B,\text{ATOM/REFL})$ explained all differences in atomic positions between the pairs of structures, a Z-score was defined for each atom as:

$$Z\text{-score} = \frac{\Delta x}{\epsilon_\chi(B)}, \frac{\Delta y}{\epsilon_\chi(B)} \text{ or } \frac{\Delta z}{\epsilon_\chi(B)}, \tag{10}$$

where Δx, Δy, and Δz are the differences in the position of an atom in the two molecules in the asymmetric unit along each of the orthogonal axes.

If the error curves, $\epsilon_\chi(B)$, really do reflect a normal distribution of differences, the distribution of Z-scores should be Gaussian with a standard deviation of 1.0. As shown in Table 2, 13 of the 18 structures had an overall standard deviation within 20% of 1.0, and only one structure (2PFK) deviates from its $\epsilon_\chi(B)$ curve by more than 50%. Thus, the variation in positional uncertainty of most of the 29,280 atom pairs in the study is substantially contained within the six parameters ($k1$–$k6$) used to construct ϵ_χ.

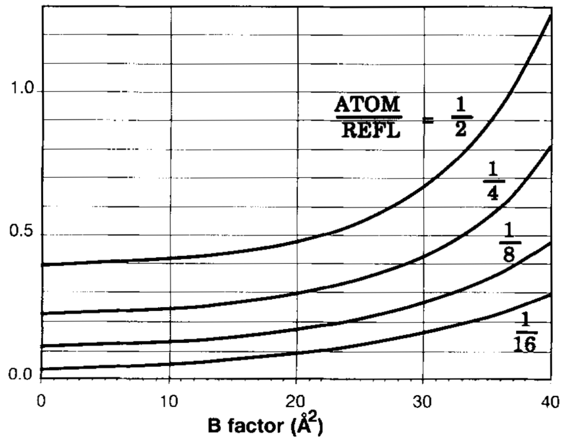

Fig. 6. Family of $\epsilon_\chi(B)$ curves. Each value of ATOM/REFL yields a distinct member of this family. The curves shown span the range of values seen in this study. From top to bottom, the values of ATOM/REFL are 1/2, 1/4, 1/8, and 1/16, respectively.

Table 2. *Internal control and comparison of $\epsilon_\chi(B)$ to the Luzzati formula*

Structure	SD of Z-score	⟨Δℛ⟩ Luzzati	⟨Δℛ⟩ from ϵ_χ
1THB	1.08	0.25	0.12
2CCY	0.95	0.20	0.14
4CHA	0.83	–	0.12
4DFR	0.91	0.15	0.18
2HHB	1.08	0.18	0.14
3CYT	0.81	0.20	0.16
2AZA	1.04	0.15	0.15
1GD1	0.67	0.18	0.15
1AZA	0.98	–	0.20
1GP1	0.96	–	0.14
1HMQ	0.96	–	0.13
2PKA	1.37	0.20	0.14
2PFK	0.42	–	0.31
1FCB	0.74	–	0.23
4MDH	1.10	0.225	0.31
4ATC	1.17	–	0.36
1FC1	0.74	–	0.40
1HBS	0.90	0.40	0.55

Other influences on accuracy of atomic positions extracted from correlation analysis

Normalized error score

The predicted error curve, $\epsilon_\chi(B,\text{ATOM/REFL})$ contains contributions from atomic B factor and the resolution of the data. To evaluate further atomic attributes that might influence positional accuracy, a normalized error score was defined as the standard deviation of a Gaussian fit to the Z-scores of a selected subset of atoms divided by the standard deviation of a Gaussian fit to the Z-scores for all the carbon atoms (which constitute more than 64% of the atoms evaluated), e.g.,

$$\text{N.E.S.(subset)} = \frac{\sigma(Z\text{-score(subset)})}{\sigma(Z\text{-score(carbon)})}. \tag{11}$$

Thus, if $\epsilon_\chi(B)$ correctly predicts the accuracy of a subset of atoms, the normalized error score for that subset of atoms should be close to 1.0. If the $\epsilon_\chi(B)$ estimation is too large or too small, the N.E.S. will be less than or greater than 1.0, respectively.

A standard deviation for a normalized error score, σ_{NES}(subset), was calculated by first evaluating a separate normalized error score for each of the 18 structures separately, N.E.S.$_i$(subset), and then taking the standard deviation of these 18 values. This standard deviation indicates how consistent a particular normalized error score is over the 18 structures used in the study.

The N.E.S. should already account for errors associated with B factor and resolution. As shown in the left-hand panel of Figure 8, there is no variation in N.E.S. for atoms of different B factors, within the error given by σ_{NES}, which confirms that $\epsilon_\chi(B)$ has accounted for variations due to B factors, even in the bins below 14 Å^2 and above 31 Å^2, which include atoms that were not used in constructing $\epsilon_\chi(B)$.

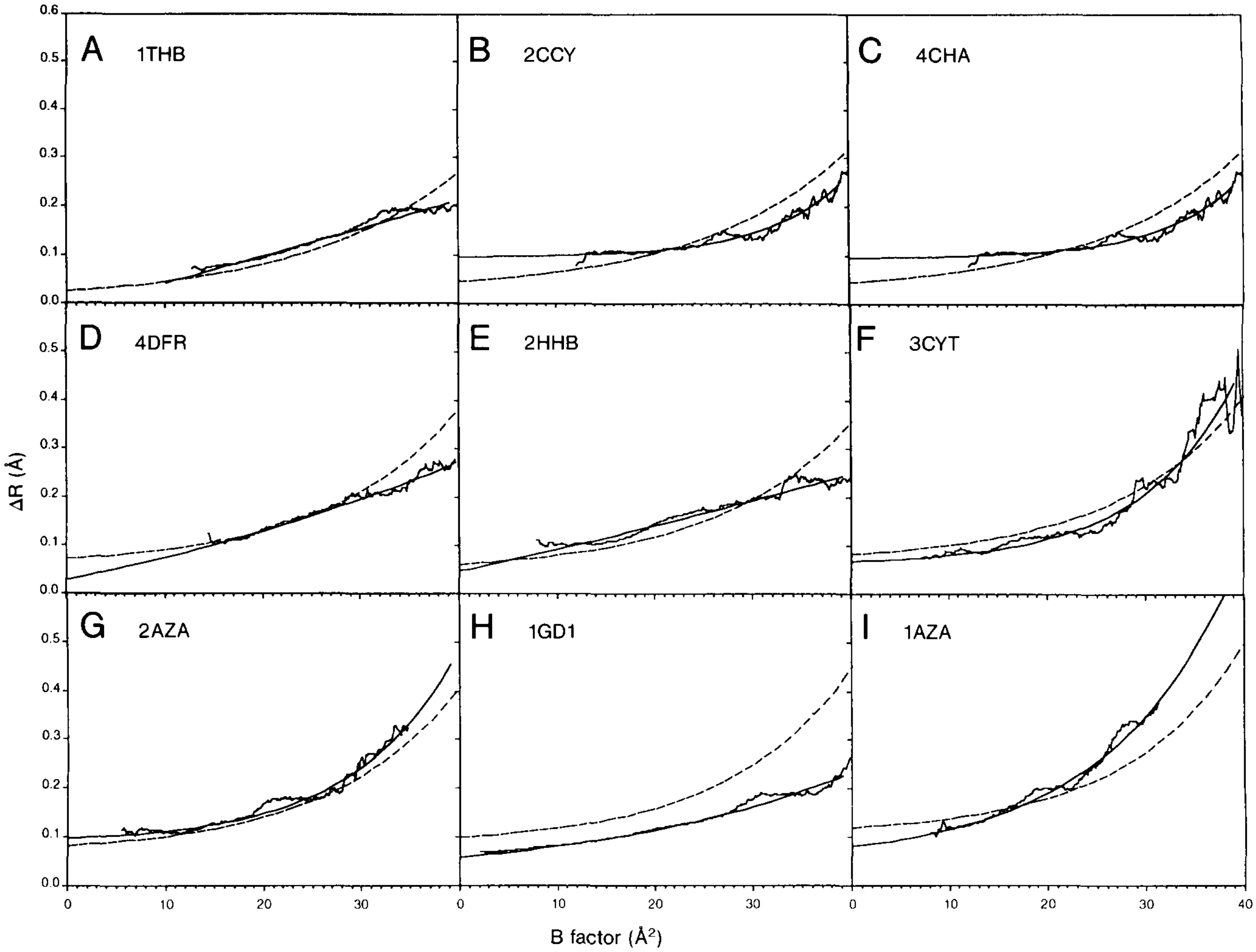

Fig. 7. Observed errors, $\sigma_\chi(B)$, and expected errors, $\epsilon_\chi(B)$, for each of the 18 structures used in the study. Because the $\sigma_\chi(B)$ curves were used to derive the empirical formula, the degree to which the two curves in each figure match indicates how well all the information from all the curves has been reduced to six parameters ($k1$-$k6$). The structures are displayed in order of resolution of the structure. In each figure, the choppy line is the standard deviation of $\Delta\chi$ in each B factor bin, the thick solid line is the three-parameter $\sigma_\chi(B)$ curve, and the dashed line is the curve calculated for that structure from $\epsilon_\chi(B,\text{ATOM/REFL})$. Note that ordinate of four figures (1MDH, 4ATC, 1FC1, and 1HBS) goes to 2.4 Å, whereas the ordinate of the other 14 goes to 0.6 Å. (*Continues on facing page.*)

No correlation with atomic number

The second panel of Figure 8 shows that there is also little or no difference in N.E.S. due to atomic number. That is, carbons, nitrogens, and oxygens are all positioned with equal accuracy on average. It is difficult to draw any conclusions about sulfurs, because there are so few in any given structure (between 5 and 14). This is reflected in the large error bar (σ_{NES}) for sulfur in the central panel of Figure 8. The N.E.S. for sulfur, however, indicates that its accuracy is close to that of the other atoms.

In contrast to the results presented here, however, the Cruickshank formula predicts that errors in position are inversely related to the number of electrons in the given atom type. This is because the "curvature" used in the Cruickshank equation will be greater for an atom with more electrons at a given B factor. The apparent lack of a dependence on atomic number here is probably due to the restraints and constraints applied in macromolecular crystallography, which ensures that the accuracy of one atom is highly related to the accuracy of its covalently bound neighbors (see below).

Correlation with connectivity of atoms

The connectivity of an atom is strongly correlated with accuracy of its position as demonstrated in the third panel of Figure 8. Atoms of the main chain (C, N, Cα, and O) have lower than expected errors (N.E.S. < 1.0). The main-chain atoms have lower positional errors than side-chain atoms with three non-hydrogen neighbors, which in turn have lower errors than atoms with two non-hydrogen neighbors. Side-chain atoms of only one non-hydrogen neighbor have the most uncertainty of all, with an N.E.S. 50% greater than that for main-chain atoms. It is especially noted that this is after B factor-correlated effects have been accounted for. Thus, on average, a side-chain atom

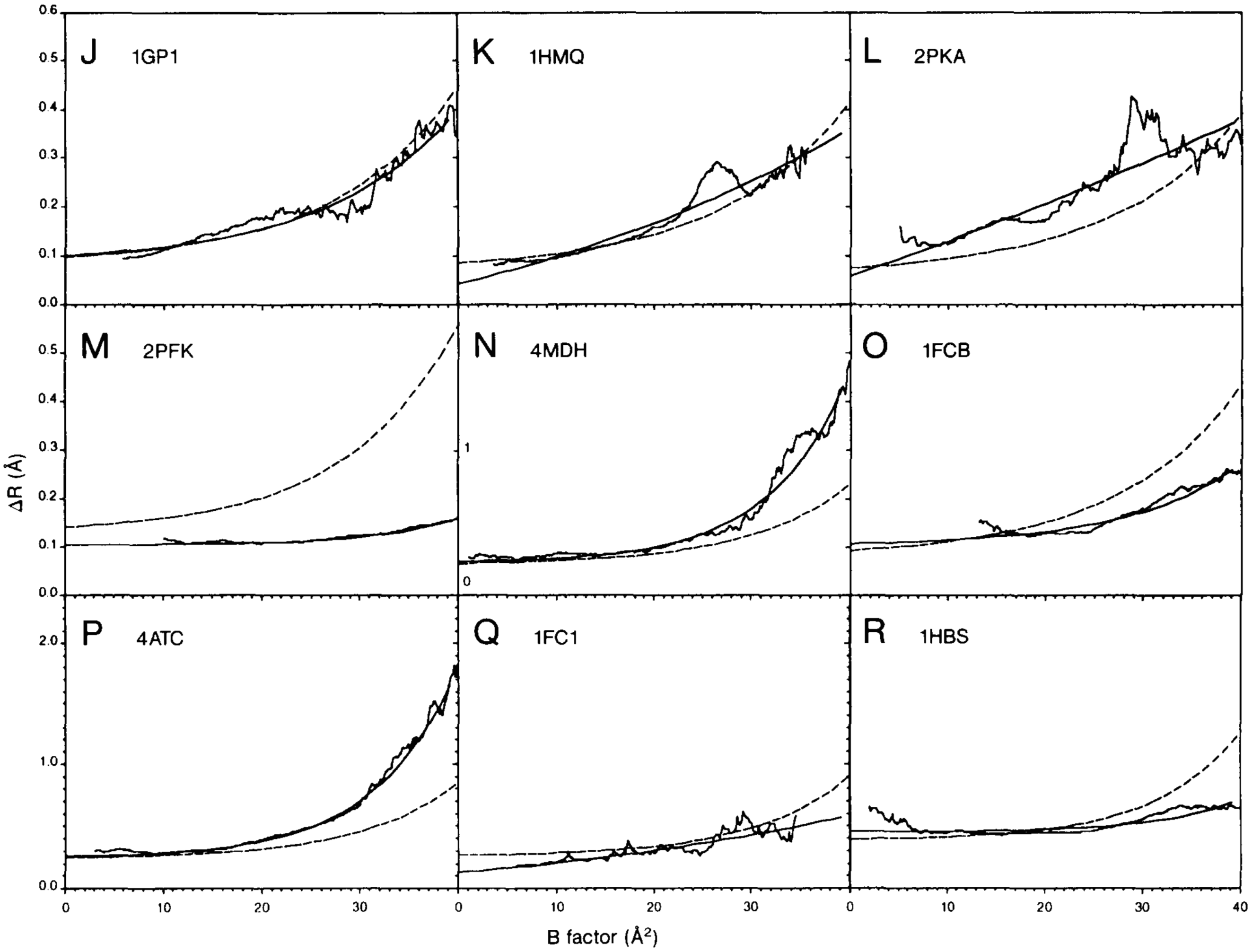

Fig. 7. *Continued.*

with a B factor of 15 Å^2 has a 50% greater positional uncertainty than a main-chain atom with a B factor of 15 Å^2.

The more non-hydrogen neighbors a given atom has, the lower its error, independent of the B factor of the atom. This can be seen as an extension of relationship between observations/parameters and overall accuracy. The positions of neighboring atoms can be seen as additional observations affecting the given atom position. Likewise, the more neighbors, the fewer degrees of freedom, or parameters, are available for positioning the given atom.

Discussion

Comparison with the "Luzzati" formula

A widely used measure of the positional accuracy of crystal structures is the relationship of Luzzati (1952). However, constructing a "Luzzati plot" requires access to the original structure factors (F_o's) and the Luzzati method assumes all atoms have the same B factor. The Luzzati method produces a single overall value for the accuracy of a structure, $\langle \Delta\mathcal{R} \rangle$, which is portrayed as the average atomic displacement from the "true" structure. To the extent that atoms have a range of B factors, the Luzzati plot, because it emphasizes the high-resolution data, represents the expected errors of only the atoms with the lowest B factors in the structure.

To calculate an overall $\langle \Delta\mathcal{R} \rangle$ for a structure from $\epsilon_\chi(B)$, individual atomic $\Delta\mathcal{R}$'s were calculated from the relationship:

$$\Delta\mathcal{R}_{mp} = \epsilon_\chi(B, \text{ATOM/REFL}), \qquad (12)$$

where $\Delta\mathcal{R}_{mp}$ is the most probable value for $\Delta\mathcal{R}$ based on the atom's B factor and the value of ATOM/REFL for the structure (Appendix 4).

Ten of the structures used in this study had Luzzati values reported for them. For comparison, in Table 2, a value for $\langle \Delta\mathcal{R} \rangle$ has been calculated from $\epsilon_\chi(B)$ for those 10 structures using all atoms with B factors less than 40 Å^2, by our method (Equation 9). There is a rough correspondence between the values, with a correlation coefficient of 0.86 (which is statistically significant at the 99.9% level). However, most of this correlation is due to 1HBS, the lowest-resolution structure in the study, because, if this point is omitted, the remaining nine structures yield a correlation coefficient of only 0.15.

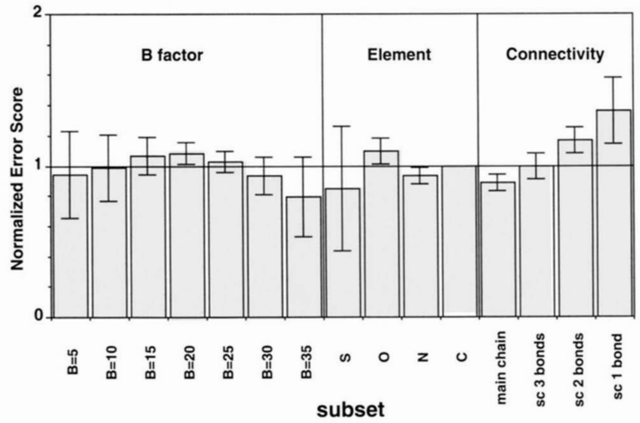

Fig. 8. Normalized error score (N.E.S.) indicates how well $\epsilon_\chi(B)$ accounts for different levels of errors in different subgroups of atoms. An N.E.S. value below 1.0 implies $\epsilon_\chi(B)$ overestimates the errors in that subgroup; a value above 1.0 implies $\epsilon_\chi(B)$ underestimates the errors for that subgroup.

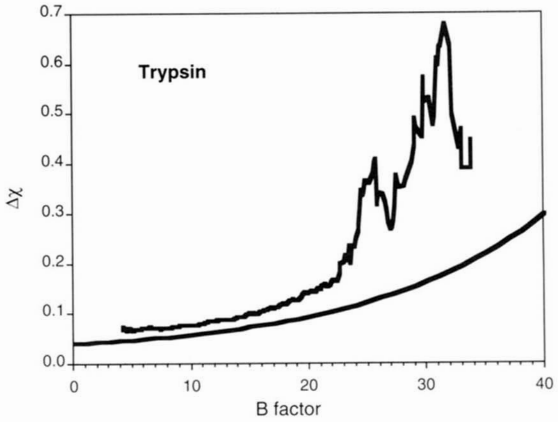

Fig. 9. Application of the $\epsilon_\chi(B)$ curve. The smooth curve represents the predicted one-dimensional standard deviation of the positional differences between two independently solved trypsin structures. The jagged curve shows the observed one-dimensional standard deviation of the positional differences.

Because the Luzzati method uses the observed structure factors, it is useful for evaluating the progress of refinement, which $\epsilon_\chi(B)$ is not. However, the Luzzati method cannot assign errors to individual atoms, as $\epsilon_\chi(B)$ does. In addition, calculation of $\epsilon_\chi(B)$ for a structure requires only the number of atoms and the number of reflections used in refinement, which should be provided in any published report of a crystal structure, and our goal is to derive an error estimate that is readily accessible from the commonly reported data or data recorded within the Protein Data Bank files.

Use of $\epsilon_\chi(B)$

The function $\epsilon_\chi(B)$ can be used to estimate the errors in refined macromolecular crystal structures. Because pairs of structures were used to derive $\epsilon_\chi(B)$, the function represents the expected (one-dimensional) differences between two structures. The expected errors in any one structure are then $\epsilon_\chi(B)/\sqrt{(2)}$. The expected random differences between two structures will then be:

$$\epsilon_{total}^2 = \frac{\epsilon_{s1}^2 + \epsilon_{s2}^2}{2}, \tag{13}$$

where ϵ_{s1} is $\epsilon_\chi(B)$ for the first structure and ϵ_{s2} is $\epsilon_\chi(B)$ for the second structure. ϵ_{s1} and ϵ_{s2} will be different if the ratios, ATOM/REFL, are different for the two structures.

The results of such an analysis are presented in Figure 9, for the comparison of two independently solved structures of bovine trypsin (Chambers & Stroud, 1979). The expected errors are, in general, close to the observed differences between the structures, especially for B factors below 20 Å^2. For B factors above 20 Å^2, the observed differences exceed the predicted errors. This probably indicates that $\epsilon_\chi(B)$ underestimates the true uncertainty in a crystal structure, because it was derived from pairs of molecules that were solved simultaneously and under identical conditions.

One important exception occurs when difference $F_{o_1} - F_{o_2}$ Fourier maps are used to determine shifts in position (Chambers & Stroud, 1977). These differences can be extracted with potentially much greater accuracy than our formula suggests for either structure alone. Difference in position for atoms may approach ~1/10–1/30 of the resolution in Å (see, for example, Krieger Kay & Stroud, 1974).

A word about B factors

Two assumptions about B values that are inherent to our analysis are worth further consideration: (1) that B factors are accurate, and (2) that they are refined in a consistent manner by all crystallographers. In the case of macromolecular crystallography, these contentions are clearly debatable. For example, B factors, far more than the positional parameters (x, y, and z), are extremely sensitive to how the observed amplitudes (F_o's) are scaled and to the resolution range that was used in refinement. In addition, whereas atomic positions are restrained by known stereochemistry and van der Waals interactions, atomic B factors are typically restrained only minimally, for example, through a standard deviation linking the B factors of bonded atoms. The use of minimal restraints on B factors of bonded atoms is well justified by highly refined structures where B factors were determined for each atom completely independently (Chambers & Stroud, 1977), in which it is observed that B factors of neighboring atoms are coupled. Rigid body motions such as those of rings that may oscillate are reflected in the variations of individually refined B factors around the rings, such that a predominant zone of anisotropic rigid body motion of the ring is suggested by trends of B factors around the ring. This kind of evidence validates the use of some restraints in coupling the B factors of neighboring atoms in the structure.

If the observed F_o data are "scaled up" at some point in the refinement to correspond to a different fall off in intensity by resolution, the apparent B factors will all be decreased (by the same ΔB amount if the correction is isotropic), and the apparent error estimates calculated by the procedures outlined here will perhaps be falsely low. Such scaling changes should be apparent from the original publication of procedures. They can also be suspected and queried if the limiting resolution of recorded data is not commensurate with the average B factors listed. Because it is usual to record, and include in refinement, all data to the highest observable resolution, the high-resolution limit reflects levels of static and dynamic disorder that are evidenced in the crystal and in the protein structure. These effects will also be reflected in the *B* factors, higher *B* factors when the disorder is greater. Thus, the *B* factors in protein structures do incorporate many factors besides the intended harmonic motion; however, each of these factors also relates to the accuracy of the final reported structure. The approach we have taken here incorporates the limits of resolution as a correlate with differences in structure, and so determines the best overall aggregate of parameters that best account for observed differences in structure, including this limit of resolution attainable. However, the objective here is to define expected differences between independent determinations of a protein structure that should otherwise be identical, in the present-day environment where such adjustments to the data will be made if they are deemed to improve the resulting structure, at each particular resolution: the goal is always the best structure possible with the best present-day presumptions. The expression we derived, by the way in which it is derived from protein structures that are chemically identical molecules in identical solvent conditions, presumes a best case scenario. It provides a "best case" estimate. It can therefore be useful as the expected standard deviation in position, and hence in defining what is a significant alteration in structure.

To the degree that restraints applied during refinement are based on previous structures that were determined without constraints or restraints being applied, they serve to restrain those aspects of the new structure toward the consensus values, or more realistically, toward the distribution of presently validated values. An example is the restraints applied to dihedral angles, that reflect values observed from protein structures that were unrestrained in this or related parameters. Applied restraints may also be based on other knowledge or expectation, known more accurately than could be determined simply by consensus values from similar X-ray structures, such as bond lengths and angles in amino acids or other groups from small molecules, or spectroscopically determined distances. Thus, any use of restraints and constraints that are applied during crystallographic refinement will usually be applied only if the overall structure will improve as a result, and so they serve to increase the accuracy of the structure. Resulting atom positions should be closer to their "true" median positions than they would be if those stereochemical restraints had not been applied. They provide an improved structure that is most consistent with the restraints from data observable only to limited resolution, or where the number of observations per parameter is limited. More restraints are likely to be applied where the data are fewer. Because what we seek to define here is the level of accuracy that can be expected with the application of the level of restraints most appropriate to the particular limitations of resolution in the crystallographic data, we extracted the most highly correlated parameters, taking account of the crystallographers having done the best reasonable refinement consistent with the observed data. We extracted the parameters using an "artificial intelligence" approach to extracting those interrelationships that provide the most important correlates with observed differences.

To the extent that $\epsilon_x(B)$ could be parameterized in a way that depends on *B* factors, the assumptions 1 and 2 (above) about *B* factors are justified. The remaining discrepancy between observed and predicted error levels exhibited in Figure 7, however, could be due to the breakdown of the above assumptions. For example, for 2PFK (Fig. 7M), $\sigma_x(B)$ falls far below $\epsilon_x(B)$. However, the *B* factors in 2PFK extend up to 100 $Å^2$ and have a mean of 40 $Å^2$ – indicating a struggle with a fairly low-resolution structure, and $\sigma_x(B)$ extends far to the right of that displayed in Figure 7M, resembling $\epsilon_x(B)$ with *B* replaced by $B/2$. Thus, our derived expression is probably best suited to those structures where the *B* factor distribution resembles that observed for the majority of the test cases, that is, with most atomic *B* factors falling between 10 and 40 $Å^2$. The 2PFK structure that is aberrant could now be identified beforehand as atypical because its mean and maximum *B* factors differ greatly from the other structures. As another example, the errors in 1HBS (Fig. 7R) do not seem to correlate significantly with *B* factor. However, the *B* factors in this 3-Å (low) resolution structure analysis show little consistency from molecule to molecule in the asymmetric unit, questioning the validity of refining atomic *B* factors at this resolution. The recently introduced concept of the free *R* factor as a test (Brünger, 1992) is useful in determining when atomic *B* factors can be refined safely. Hence, the expression we derived will work optimally for structures whose *B* factor distributions resemble those in the majority of the test cases – namely, the majority of atoms should have *B* factors between about 10 and 40 $Å^2$.

More complicated expressions could clearly be derived, however, we sought the most accurate overall expression that could readily be applied to structures as they are recorded in the Protein Data Bank, with the data recorded in the PDB, and allowing for aspects of the analysis that may be obtained from the original publication. In the future it may become desirable to record enough information about the structure factors and phases that more direct assessment of the expected variations in position of each atom position are recorded or can be extracted from the PDB files. But even this approach is potentially flawed because the motions of atoms within a highly coupled system such as a protein molecule are not isotropic – as represented by isotropic *B* factors – as they are almost uniformly treated. Indeed, they are not even harmonic in anisotropic motions, as suggested for example by the trajectories seen in a molecular mechanics simulation (Schiffer et al., 1993). Thus, the treatment as isotropic *B* factors is often necessary to keep the number of refinable parameters within bounds of reasonableness, and to keep the ratio of the number of observations per parameter reasonable (above 3–10). The cost of this approximation is probably largely seen in the *R* factors representing the best agreement between observed and calculated structure factors, typically 13–21% between different structures, which is well above the accuracy with which the data can be recorded.

Use of restraints that serve to reduce the number of parameters is especially justified when the ratio of the number of observations to parameters is small. All other things being equal, the ratio of number of observations per parameter is independent of size of the protein molecule in protein crystallography. The number of parameters correlates one-to-one with the number of amino acids in one asymmetric unit of the structure. The number of observations at a given resolution goes up linearly with the unit cell volume, and hence approximately as the volume of, or number of, amino acids in the asymmetric unit of structure. Thus, the ratio of observations to parameters does not inherently change as a consequence of protein size per se, crystal symmetry, or space group. However, it does change with the observed resolution of diffraction. The number of observations depends on the inverse third power of resolution attainable. If resolutions of ~2.5–2.0 Å are accessible, the parameters to observations ratio becomes more acceptable. Increasing the restraints, or averaging of multiple copies if they exist in one asymmetric unit, is a means of decreasing the effective number of parameters to maintain a reasonable ratio when the attainable resolution is worse. Thus, much useful biological information can be extracted from structures determined from data that is only observable to limited resolution, by applying restraints.

As an example, typically, for a 25-kDa protein structure of ~225 amino acids, ~2,000 heavy atoms (not including hydrogen atoms), the number of parameters used to describe the structure is about 2,000 B factors (one for each atom, and fewer if restrained) and, because of stereochemical restraints, approximately four to five positional parameters on average per amino acid. This can be thought of as ϕ,ψ angles for each backbone peptide, and a value for χ_1, χ_2 of each side chain. The approximate total is ~2,900 parameters, depending somewhat on the restraints applied during refinement. This is far short of the 8,000 parameters that would be calculated if each x, y, z, and B were independent. Presuming that the protein occupies p% of the unit cell – a typical value being ~60% of a unit cell that is ~40% solvent (typical values for protein crystals, that have a spread of ~±20%) – the number of theoretically observed observations, typically the number of recorded independent intensities I_{hkl}, depends on the inverse cube of the resolution (generally recorded in Å). Thus, the ratio of observations to parameters becomes $\pi(*a*b*c)/6*2{,}900\ (\text{Resolution})^3$, ~0.18 at 6 Å resolution, 1.45 at 3 Å resolution, 2.5 at 2.5 Å, 4.9 at 2 Å and ~11.6 at 1.5 Å resolution. Thus, restraints that serve to increase this ratio above 1, which is unreliable, toward ~10 can increase the accuracy of the structure.

Conclusion

By analysis of 18 structures with multiple molecules in the asymmetric unit, a function has been derived that reproduces the positional differences observed between equivalent atoms in the chemically identical structures. We believe this function, $\epsilon_\chi(B)$, truly represents the accuracy of a macromolecular structure because: (1) systemic differences (crystal contacts) were removed by using only normally distributed positional differences; (2) the overall level of error predicted for a structure based on appropriately averaging $\epsilon_\chi(B)$ values generates an overall error estimate similar to that by obtained by the Luzzati or other methods; and (3) the predicted errors decrease with increasing resolution of the analysis.

We show here that $\epsilon_\chi(B)$ can be used to predict the expected level of errors in other macromolecular crystal structures from usually obtainable information presented at the end of an analysis, and thus can be used in evaluating the reliability of crystallographic coordinates atom by atom. Because of the empirical nature of this study, our expressions for $\epsilon_\chi(B)$ are validated more when interpolating, rather than extrapolating to conditions beyond which our database extended. This covers most usual cases, but means that $\epsilon_\chi(B)$ is best validated when applied to structure analyses between resolution 1.5 Å and 3.0 Å, and for those atoms with B factors less than 40 Å^2. It is clear that atoms with B factors above this will have larger errors, but the exponential form used here may not be as appropriate in that range. This function therefore provides an extremely good and readily accessible way of evaluating the significance of any shift or movement of structure, for example, in response to the binding of ligands or seen in mutational analysis of any protein structure, in comparison with expected errors in position for any particular atoms. A computer program extracting these error estimates is obtainable from R.M.S.

Materials and methods

Structures

Eighteen structures that contain multiple molecules in the asymmetric unit were found by searching the Brookhaven Protein Data Bank (Bernstein et al., 1977) (Table 1). They were selected such that in no case was the equivalence of structure between noncrystallographically related molecules imposed during refinement. However, in individual cases, some predisposition toward equivalent structure had been imposed, usually in early stages. However, this is not a weakness, but rather a strength in our analysis because it is weighted against incorporating errors of interpretation, or standard errors, in the consideration of what we seek, namely the variations due to random errors. Thus, in some cases one or more of the following constraints were applied: the same starting structure was used for independent molecules; noncrystallographic symmetry was applied at the earliest stages of refinement; dihedral angles were averaged over the independent molecules early in refinement; the atomic positions were averaged early in refinement; changes to be made to one molecule were checked against the electron density maps for the other independent molecule(s). These techniques will all tend to reduce systematic errors in observed positional differences, and so focus concern on the effective amplitudes of vibration of atoms in the structure within their own potential wells. This expectation bears directly on extraction of true positional shifts between different regions of protein structures induced, for example, by binding different ligands or with different mutations.

In preparing the structures, atoms were excluded if they had no refined B factor or a low occupancy. Eighty-seven atoms were eliminated because they or their analogous atom in the other molecule of the asymmetric unit had a B factor of zero. Also, for atoms with multiple occupancies, only the greater occupied site was used. These criteria eliminated 45 atoms. In all, 70,120 atoms were included in the analysis representing 28,720 "multiply observed" atoms. For those structures that contain four independent molecules in the asymmetric unit (1HBS, 1HMQ, 1GD1, 2PFK), six separate comparisons were performed and then pooled and averaged to represent a single data point for the structure of that protein.

Superposition of equivalent protein structures for comparison of differences

Independent molecules in each asymmetric unit were overlapped by minimizing the RMS deviation of a set of core Cα's. The core of Cα's was selected by analysis of a difference distance matrix to identify the largest set of Cα's that simultaneously fulfills the two criteria that (1) every Cα in the core moves less than 0.5 Å relative to every other Cα in the core, and (2) the set is connected in a structural sense, with each Cα in the core being within 10 Å of at least one other Cα in the core. The core so selected varies from 15 to 70% of all Cα's in the protein, depending on the overall level of difference between the proteins being compared. This algorithm for identifying a common core is coded in the program NEWDOME (Montfort et al., 1990; Perry et al., 1990).

Correcting rotamers of twofold symmetric side chains

The side chains of Phe, Tyr, Asp, Glu, and Arg are twofold symmetric about the last free dihedral angle. Thus, labeling of certain atoms in these side chains is arbitrary with respect to the chemistry of the amino acids.[2] Therefore, these side chains were overlapped and the matching rotamers adjusted to minimize the RMS deviation between structures. Side chains of His, Asn, and Gln were similarly corrected because, crystallographically, these side chains can be twofold symmetric unless hydrogen bonding geometry is clear. This affected 120 residues of a total of 4,500 in the analysis. The overall RMS deviation between pairs of structures was decreased by as much as 0.5% by the adjustment.

Acknowledgment

This work was supported by National Institutes of Health grant CA63081.

References

Baker E. 1988. Structure of azurin from *Alcigenes denitrificans*. Refinement at 1.8 Å and comparison of the two crystallographically independent molecules. *J Mol Biol 203*:1071–1095.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Birktoft J, Rhodes G, Banaszak L. 1989. Refined crystal structure of cytoplasmic malate dehydrogenase at 2.5 Å resolution. *Biochemistry 28*: 6065–6081.

Bode W, Chen Z, Bartels K. 1983. Crystallization, structure determination, crystallographic refinement, structure and its comparison with bovine trypsin. *J Mol Biol 164*:237–282.

Bolin J, Filman D, Matthews D, Hamlin R, Kraut J. 1982. Crystal structures of *Escherichia coli* and *Lactobacillus casei* dihydrofolate reductase refined at 1.7 Å resolution. *J Biol Chem 257*:13650–13662.

Bott R, Frane J. 1990. Incorporation of crystallographic temperature factors in the statistical analysis of protein tertiary structures. *Protein Eng 3*:649–657.

Brünger AT. 1992. Free *R*-value—A novel statistical quantity for assessing the accuracy of crystal structures. *Nature 335*:472–475.

Chambers JL, Stroud RM. 1977. Difference Fourier refinement of the structure of DIP-trypsin at 1.5 Å using a minicomputer technique. *Acta Crystallogr B 33*:1824–1837.

Chambers JL, Stroud RM. 1979. The accuracy of refined protein structures: Comparison of two independently refined models of bovine trypsin. *Acta Crystallogr B 35*:1861–1874.

Chothia C, Lesk AM. 1986. The relationship between the divergence of sequence and structure in proteins. *EMBO J 5*:823–826.

Clore GM, Gronenborn AM. 1991. Comparison of the solution nuclear magnetic resonance and X-ray crystal structures of human recombinant interleukin-1beta. *J Mol Biol 221*:47–53.

Cruickshank DWJ. 1949. The accuracy of electron-density maps in X-ray analysis with special reference to dibenzyl. *Acta Crystallogr 2*:65–82.

Deisenhofer J. 1981. Crystallographic refinement and atomic models of a human Fc fragment and its complex with fragment B of protein A from *Staphylococcus aureus* at 2.9- and 2.8-Å resolution. *Biochemistry 20*:2361–2370.

Epp O, Ladenstein R, Wender A. 1983. The refined structure of the selenoenzyme glutathione peroxidase at 0.2-nm resolution. *Eur J Biochem 133*:51–69.

Fermi G, Perutz M, Shaanan B. 1984. The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution. *J Mol Biol 175*:159–174.

Finzel B, Weber P, Hardman K, Salemme F. 1985. Structure of ferricytochrome *c*′ from *Rhodospirillum* at 1.67 Å resolution. *J Mol Biol 186*: 627–643.

Ke HM, Hozatko R, Lipscomb W. 1984. Structure of unligated aspartate carbamoyltransferase of *Escherichia coli* at 2.6 Å resolution. *Proc Natl Acad Sci USA 81*:4037–4040.

Kossiakoff AA, Randal M, Guenot J, Eigenbrot C. 1992. Variability of conformations at crystal contacts in BPTI represent true low-energy structures—Correspondence among lattice packing and molecular dynamics structures. *Proteins Struct Funct Genet 14*:65–74.

Krieger M, Kay LM, Skiond, RM. 1974. *J Mol Biol 83*:200–230.

Luzzati V. 1952. The statistical treatment of errors in crystal structures. *Acta Crystallogr 5*:802–810.

Montfort WR, Perry KM, Fauman EB, Finer-Moore JS, Maley GF, Hardy L, Maley F, Stroud RM. 1990. Structure, multiple site binding, and segmental accommodation in thymidylate synthase on binding dUMP and an anti-folate. *Biochemistry 29*:6964–6977.

Morris AL, MacArthur MW, Hutchinson EG, Thornton JM. 1992. Stereochemical quality of protein structure coordinates. *Proteins Struct Funct Genet 12*:345–364.

Norris G, Anderson B, Baker E. 1983. Structure of azurin from *Alcigenes denitrificans* at 2.5 Å resolution. *J Mol Biol 165*:501–521.

Padlan E, Love W. 1985. Refined crystal structure of deoxy-hemoglobin S I. Restrained least squares refinement at 3.0 Å resolution. *J Biol Chem 260*:8272–8279.

Perry KM, Fauman EB, Finer-Moore JS, Montfort WR, Maley GF, Maley F, Stroud RM. 1990. Plastic adaptation toward mutation in proteins: Structural comparison of thymidylate synthases. *Proteins Struct Funct Genet 8*:315–333.

Rypniewski W, Evans P. 1989. Crystal structure of unliganded phosphofructokinase from *Escherichia coli*. *J Mol Biol 207*:805–821.

Schiffer CA, Caldwell JW, Kollman PA, Stroud RM. 1993. Protein structure prediction with a combined solvation free energy-molecular mechanics force field. *Mol Simulation 10*:121–149.

Skarzynski T, Moody P, Wonacott A. 1987. Structure of hologlyceraldehyde-3-phosphate dehydrogenase from *Bacillus stearothermophilus* at 1.8 Å resolution. *J Mol Biol 193*:171–187.

Stenkamp R, Sieker L, Jensen L. 1982. Restrained least-squares refinement of *Themiste dyscritum* methydroxohemerythrin at 2.0 Å resolution. *Acta Crystallogr B 38*:784–792.

Takano T, Dickerson R. 1980. Redox conformation changes in refined tuna cytochrome. *Proc Natl Acad Sci USA 77*:6371–6375.

Tsukada H, Blow D. 1985. Structure of α-chymotrypsin refined at 1.68 Å resolution. *J Mol Biol 184*:703–711.

Waller DA, Liddington RC. 1990. Refinement of a partially oxygenated T state haemoglobin at 1.5 Å resolution. *Acta Crystallogr B 46*:409–418.

Xia Z, Mathews FS. 1990. Molecular structure of flavocytochrome B2 at 2.4 Å resolution. *J Mol Biol 212*:837–863.

Appendix 1

Proof that P(X) = 1/R for 0 < X < R

The exact form used here assumes only positive values of X.

This result says that all values of the X component from a random three-dimensional vector of length R are equally likely. Because this result is rather counterintuitive, the derivation is presented below.

The probability of a given event X, $P(X)$, is defined as the number of outcomes with a value between X and $X + \delta x$ divided by the total number of outcomes.

Consider a sphere of radius R. $P(X)dx$ is then that surface area of the sphere generated with an X component between X and $X + dx$, divided by the total surface area of the hemisphere (because we are only interested in positive values of X).

The desired surface area can be calculated from the following integral:

$$\int_{\theta=\cos^{-1}(X/R)}^{\theta=\cos^{-1}[(X+dX)/R} \int_{\varphi=0}^{\varphi=2\pi} R^2 \sin\theta \, d\theta \, d\varphi, \tag{1.1}$$

[2] There is a convention for unambiguously labeling these residues; the atoms should be named so as to give the lower dihedral angle value for the appropriate atoms. However, as noted by Morris et al. (1992), this convention is rarely if ever used by macromolecular crystallographers.

where θ is the angle between the R vector and the x axis, φ is the angle between the projection of R into the y-z plane, and the z axis, and $R^2 \sin\theta d\theta d\varphi$ is the surface area element in spherical coordinates.

Evaluating this double integral yields the value $2\pi RdX$. The surface area of a hemisphere is $2\pi R^2$, so that

$$P(X)dX = \frac{2\pi RdX}{2\pi R^2} = \frac{dX}{R}, \tag{1.2}$$

or, equivalently

$$P(X) = \frac{1}{R}. \tag{1.3}$$

Appendix 2

ATOM/REFL

The ratio ATOM/REFL is related to the maximum resolution and the protein fraction in the unit cell. The following symbols are used:

ρ = density of protein = 1.3 Å^3/Da
A = average molecular mass per non-hydrogen atom = 14 Da
V = volume of unit cell
a = unit cell length, defined as the cube root of the cell volume
ATOM = non-hydrogen atoms per asymmetric unit
m = asymmetric units per unit cell
F_p = protein fraction in the unit cell
a^* = reciprocal unit cell length = $1/a$
d = maximum resolution
d^* = maximum $|s| = 1/d$
$N(hkl)$ = number of observations
REFL = number of independent reflections

The number of possible observations is related to the volume of the sphere in reciprocal space:

$$obs = \frac{4}{3}\pi\left(\frac{d^*}{a^*}\right)^3. \tag{2.1}$$

To get the number of unique reflections, divide by $2m$. Also, we can replace d^* and a^* by $1/d$ and $1/a$, respectively, to get:

$$\text{REFL} = \frac{4}{3}\pi\left(\frac{a}{d}\right)^3\Big/2m. \tag{2.2}$$

The volume of the unit cell is given by:

$$V = (m\rho A)\text{ATOM}/F_p. \tag{2.3}$$

Assuming a cubic lattice, $V = a^3$. Thus, we can rewrite the ratio

$$\text{ATOM/REFL} = (\text{ATOM})(2m)\frac{3}{4\pi}\left(\frac{d}{a}\right)^3 \tag{2.4}$$

as

$$\text{ATOM/REFL} = \frac{(\text{ATOM})(2m)3d^3F_p}{4\pi\rho(\text{ATOM})mA} = \frac{3}{2\pi\rho A}d^3F_p. \tag{2.5}$$

Taking $\rho = 1.3$ Å^3/Da and $A = 14$ Da, this simplifies to:

$$\text{ATOM/REFL} = \frac{F_p}{38\ \text{Å}^3}d^3. \tag{2.6}$$

Taking an average value of 0.5 for F_p (50% solvent content) we can write:

$$\text{ATOM/REFL} = \frac{d^3}{76\ \text{Å}^3}. \tag{2.7}$$

Appendix 3

Constructing an exponential curve expression for $\epsilon_\chi(B)$

$\epsilon_\chi(B,\text{ATOM/REFL})$ is defined by the following equation:

$$\epsilon_\chi(B,\text{ATOM/REFL}) = \text{Intercept}(B) + \text{Slope}(B)*e^{(-2*\text{ATOM/REFL})}, \tag{8}$$

where Intercept(B) and Slope(B) are defined by Equations 6 and 7.

At one specific value of ATOM/REFL, $\epsilon_\chi(B)$ is an exponential function of the B factor, as are Intercept(B) and Slope(B). This dependence can be made explicit by recasting $\epsilon_\chi(B)$ as:

$$\epsilon_\chi(B) = p1 + p2*e^{(B/p3)}. \tag{9}$$

Any three points can be fit exactly by a three-parameter exponential. Thus, the values of $\epsilon_\chi(B)$ for $B = 10$ Å^2, 20 Å^2, and 30 Å^2 calculated from Equation 7 uniquely determine the three parameters, $p1$, $p2$, and $p3$, for any given value of ATOM/REFL. Namely:

$$p3 = 10\frac{\epsilon_\chi(20,\text{ATOM/REFL}) - \epsilon_\chi(10,\text{ATOM/REFL})}{\epsilon_\chi(30,\text{ATOM/REFL}) - \epsilon_\chi(20,\text{ATOM/REFL})}, \tag{3.1}$$

$$p2 = \frac{\epsilon_\chi(20,\text{ATOM/REFL}) - \epsilon_\chi(10,\text{ATOM/REFL})}{e^{(20/p3)} - e^{(10/p3)}}, \tag{3.2}$$

$$p1 = \epsilon_\chi(20,\text{ATOM/REFL}) - p2*e^{(20/p3)}. \tag{3.3}$$

Appendix 4

The relationship between $\Delta\Re_{mp}$ and ϵ_χ

The following symbols are used:

$\Delta\Re$ = distance between the observed position for an atom and the "true" position for that atom;
$\Delta\Re_{mp}$ = most probable value for $\Delta\Re$, given a Maxwellian distribution;
ϵ_χ = expected one-dimensional standard deviation of the differences in atomic positions between two observations of the same "true" structure;
ϵ_r = expected three-dimensional standard deviation of the differences in atomic positions between two observations of the same "true" structure;
$\sigma_\Re$ = three-dimensional standard deviation of the differences in atomic position between the "true" structure and an observed structure.

By propagation of error,

$$\epsilon_r^2 = \epsilon_\chi^2 + \epsilon_\chi^2 + \epsilon_\chi^2, \tag{4.1}$$

therefore

$$\epsilon_r = \sqrt{3}\epsilon_\chi. \tag{4.2}$$

Likewise, by propagation of errors again,

$$\epsilon_r^2 = \sigma_\Re^2 + \sigma_\Re^2, \tag{4.3}$$

so

$$\sigma_\Re = \sqrt{\tfrac{3}{2}}\,\epsilon_\chi. \tag{4.4}$$

This three-dimensional standard deviation corresponds to the Maxwellian distribution:

$$P(\Delta\Re) = \frac{\Delta\Re^2}{\sigma_\Re^3}\sqrt{\frac{54}{\pi}}\exp\left(-\frac{3}{2}\frac{\Delta\Re^2}{\sigma_\Re^2}\right). \tag{4.5}$$

The maximum value of this function yields $\Delta\Re_{mp}$, which is

$$\Delta\Re_{mp} = \sqrt{\tfrac{2}{3}}\,\sigma_\Re. \tag{4.6}$$

Therefore,

$$\Delta\Re_{mp} = \sqrt{\tfrac{2}{3}}\,\sigma_\Re = \sqrt{\tfrac{2}{3}}\sqrt{\tfrac{3}{2}}\,\epsilon_\chi = \epsilon_\chi. \tag{4.7}$$