Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Distance distributions and electron-density characteristics of protein models

P. H. Zwart and V. S. Lamzin*

EMBL Hamburg Outstation, c/o DESY, Notkestrasse 85, D-22603 Hamburg, Germany

Correspondence e-mail: victor@embl-hamburg.de

Received 7 July 2003
Accepted 20 August 2003

The analytical expression for the distribution of an interatomic distance resulting from a known error-free distance and a Gaussian perturbation of the atomic coordinates is presented. This is used to estimate the coordinate error on the basis of known geometric features of protein models *via* the nearest-neighbour or the radial distance distribution. A simple relation is presented that describes the dependence of the map correlation on the positional error of the protein model, the resolution of the X-ray data and the overall atomic displacement parameter. The distribution of geometrical features and the relation between the map correlation and the positional error can be used in assisting the decision-making process during automated model-building procedures.

1. Introduction

Estimation of the map quality is essential for automated model building. Decisions on the placement of a structural element based on density criteria ideally reflect both the prior information of the expected electron density for a particular structural element as well as a measure of how well the placement of such a fragment represents the observed density with a given error. An example of this principle are the building routines implemented in *RESOLVE* (Terwilliger, 2000, 2003), where available phase-probability distributions are utilized to compute the likelihood of a fragment at a certain location. The automated model-building routines implemented in *ARP*/*wARP* (Perrakis *et al.*, 1999) operate in real space and mechanisms are being developed to incorporate information on the positional errors of free atoms and the characteristics of the electron density in the decision-making process during model building. The effects of the errors of the positional parameters during C^{α}-backbone tracing have been addressed by Morris *et al.* (2002), who showed that error-specific scoring functions can be constructed. Furthermore, this error estimate can be used to give a reasonable approximation of the size of the search problem during automated model building (Morris, Perrakis *et al.*, 2003). Although a number of error and map-quality estimation procedures are available (Cox & Cruickshank, 1948; Luzzati, 1952; Read, 1986; Lunin & Skovodora, 1995; Pannu & Read, 1996; Murshudov *et al.*, 1997; Cruickshank, 1999; Colovos *et al.*, 2000), it is worthwhile investigating whether other straightforward estimates can be obtained or used in conjunction with existing methods. In this paper, real-space properties of protein models are utilized for the estimation of model and map quality. In a forthcoming paper, we will extend the methodology to reciprocal space (Zwart & Lamzin, 2003).

Table 1
Abbreviations.

R.m.s.d.	Root-mean-square displacement
p.d.f.	Probability density function
NCM	Non-central Maxwell
m.c.c.	Map correlation coefficient

Table 2
Notation and main symbols.

$\mathbb{E}[t(x)]_x$	Expectation value of $t(x)$ by integration over x
$\langle t(x) \rangle$	Average of $t(x)$
$\mathbf{x}_j$	Positional parameters of atom j
σ_m^2	The variance of a Gaussian error model applied on the positional parameters
$\mathbf{q}_j$	The perturbation term on atom j: a random vector from a spherically symmetric Gaussian centred on the origin
$d_{jk}^{\rm tar}$	The error-free (ideal) distance between atoms j and k
$d_{jk}^{\rm obs}$	The observed distance between atoms j and k
$F(d_{jk}^{\rm obs} \| d_{jk}^{\rm tar}, 2\sigma_m^2)$	The cumulative distribution function of a distance between atom j and k, given the target distance and a Gaussian error of the positional parameters
$f(d_{{\rm min},j}^{\rm obs} \| 2\sigma_m^2)$	The nearest-neighbour distribution for atom j in a protein structure with a given Gaussian error
$f(d_{\rm min}^{\rm obs} \| 2\sigma_m^2)$	The nearest-neighbour distribution for a protein with a given Gaussian error
$f_{\rm rad}(d^{\rm obs} \| 2\sigma_m^2)$	The radial distance distribution of a protein for a given Gaussian error
2ω	The atomic width given a Gaussian density model
α	The expected map correlation
$d_{\rm min}$	The nominal resolution of an X-ray data set
$B_{\rm Wil}$	The Wilson plot B value
$\|\mathbf{F_h}\|$	A structure-factor amplitude
h	Reciprocal-lattice spacing
$\mathbf{u}$	A Patterson vector

Table 3
The vth raw moments of $d^{\rm obs}$, distributed according to NCM$(d^{\rm obs}|d^{\rm tar}, \sigma^2)$.

Erf(z) denotes the error function, $\Gamma(\cdot)$ the Gamma function and $\Phi(a, b; z)$ the confluent hypergeometric function (Lebedev, 1972).

v	$\mathbb{E}[(d^{\rm obs})^v]$
0	1
1	$(2/\pi)^{1/2}\sigma \exp\left[-\frac{(d^{\rm tar})^2}{2\sigma^2}\right] + \frac{\sigma^2 + (d^{\rm tar})^2}{d^{\rm tar}} {\rm Erf}\left[\frac{d^{\rm tar}}{(2\sigma^2)^{1/2}}\right]$
2	$(d^{\rm tar})^2 + 3\sigma^2$
v	$\frac{\sigma^v 2^{[(2+v)/2]}}{\pi^{1/2}} \exp\left[-\frac{(d^{\rm tar})^2}{2\sigma^2}\right] \Gamma\left(\frac{3+v}{2}\right) \Phi\left[\frac{3+v}{2}, \frac{3}{2}; \frac{(d^{\rm tar})^2}{2\sigma^2}\right]$

Here, we present the distribution of an interatomic distance given a Gaussian error of the coordinates. This distribution is subsequently used to compute the distribution of nearest-neighbour distances in protein models as well as the radial distance distribution, both as a function of the coordinate error. These distributions can be used to assess the stereochemical quality of an atomic model prior to chemical interpretation. Linking the r.m.s.d. (root-mean-square displacement) to a quality estimate of the electron density provides a tool that can be used elsewhere for setting empirical decision boundaries for the acceptance of the placement of a structural element on the basis of its fit to the electron density. A simple relation is presented that connects the r.m.s.d. estimate to the expected map correlation of a map with errors to the final electron-density map. The relation is derived on the basis of a simplified real-space model of the electron density and involves an empirical atomic shape parameter which is related to the optical resolution as proposed by Vaguine (1999).

The abbreviations used are given in Table 1. The notation and main symbols used in this paper are given in Table 2. In the notation of probability and density functions, the usual subscript denoting the random variable has mostly been omitted for clarity. Structure factors are distinguished from distribution functions *via* the subscript h.

2. Methods

2.1. Distance distribution

Let a pair of atoms $(\mathbf{x}_i, \mathbf{x}_j)$ separated by a *target* distance $d_{jk}^{\rm tar}$ undergo a random Gaussian perturbation of the positional parameters. Let the variance of the displacement in the x, y and z directions to be equal to σ_j^2 for atom j and σ_k^2 for atom k. Assuming that the errors of the positional parameters are independent, it can be shown (Arfken & Weber, 1995; Abramovicz & Stegun, 1974) that the *observed* interatomic distance after the perturbation, $d_{jk}^{\rm obs}$, is distributed according to

$$f(d_{jk}^{\rm obs}) = \frac{1}{[2\pi(\sigma_j^2 + \sigma_k^2)]^{1/2}} \exp\left[-\frac{(d_{jk}^{\rm obs} - d_{jk}^{\rm tar})^2}{2(\sigma_j^2 + \sigma_k^2)}\right] \times \frac{d_{jk}^{\rm obs}}{d_{jk}^{\rm tar}} \left[1 - \exp\left(-\frac{2 d_{jk}^{\rm tar} d_{jk}^{\rm obs}}{\sigma_j^2 + \sigma_k^2}\right)\right]. \quad (1)$$

A plot of this distribution for $d_{jk}^{\rm tar}$ = 1.5 Å and $\sigma_j = \sigma_k = 0.4$ is shown in Fig. 1, as well as a Gaussian approximation and a distribution obtained *via* simulation. The raw moments of this probability density function (p.d.f.) are given in Table 3.

Expression (1) becomes identical to the Maxwell distribution (Weisstein, 1999) for $d_{jk}^{\rm tar}$ = 0. For this reason, we denote

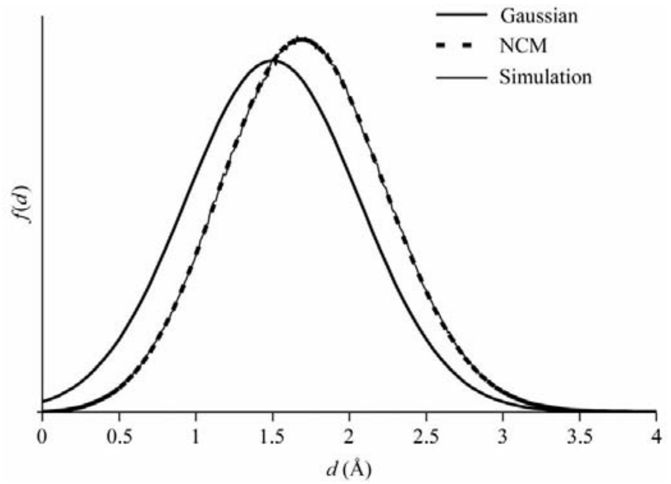


Figure 1
The non-central Maxwell distribution (dashed) and a Gaussian distribution (thick continuous line) for an ideal distance of 1.5 Å and an expected r.m.s.d. of 0.69 Å (σ_m = 0.4) on the positional parameters. The thin line represents a distribution obtained *via* simulation, exactly matching the theoretical curve.

(1) as the non-central Maxwell distribution, as it describes the distribution of a vector length of a spherical three-dimensional Gaussian centred on a vector with given length $d_{jk}^{\rm tar}$. A similar relation exists between the Rice and Wilson distributions (Read, 1990; Bricogne, 1997).

The following shorthand for (1) will be used:

$$\mathrm{NCM}(d_{jk}^{\rm obs}|d_{jk}^{\rm tar}, \sigma_j^2 + \sigma_k^2).$$

Note that the Rice, Rayleigh/Wilson, Maxwell and the NCM distribution can be seen as special cases of the generalized Rice distribution (Andersen & Kirsch, 1993). For completeness, the p.d.f. and moments of the generalized Rice distribution are given in Appendix *A*, as well as an outline of the derivation of the non-central Maxwell distribution.

The expected r.m.s.d. between an error-free and perturbed structure given a Gaussian error model with variances in each direction equal to σ_m^2 can be shown to be equal to the square root of the second raw moment of the Maxwell distribution:

$$\mathbb{E}(\mathrm{r.m.s.d.})_{d^{\rm obs}} = 3^{1/2}\sigma_m. \qquad (2)$$

(2) relates the expected r.m.s.d. to the width of the (non-central) Maxwell distribution. Precise knowledge of σ_m is thus needed to make accurate inferences on error-free bond lengths given an observed distance.

2.2. Nearest-neighbour distance distribution

Let $\mathcal{S}_j$ be the set of observed interatomic distances from atom j to all other atoms in a model:

$$\mathcal{S}_j = \{d_{j1}^{\rm obs}, \ldots, d_{j,j-1}^{\rm obs}, d_{j,j+1}^{\rm obs}, \ldots, d_{jN}^{\rm obs}\}. \qquad (3)$$

Assuming a Gaussian error with a variance σ_m^2 on the positional parameters of the model, the individual interatomic distances are distributed according to $\mathrm{NCM}(d_{jk}^{\rm obs}|d_{jk}^{\rm tar}, 2\sigma_m^2)$. Application of an error to the atomic positions results in a change of interatomic distances. A formerly large distance has a certain probability of becoming the shortest. Assuming that the elements are independently distributed, the smallest value in set $\mathcal{S}_j$ has the following (cumulative) distribution function:

$$F(d_{\min,j}^{\rm obs}|2\sigma_m^2) = 1 - \prod_{k=1;k\neq j}^{N}[1 - F(d_{jk}^{\rm obs}|d_{jk}^{\rm tar}, 2\sigma_m^2)]. \qquad (4)$$

Although the assumption of independence between the distributions of the distances is not strictly valid, the formulae are well applicable for true Gaussian errors, as shown in §3.

(4) gives the nearest-neighbour distance distribution for a single atom in a specific environment. Taking the derivative with respect to $d_{\min,j}^{\rm obs}$ to obtain the p.d.f. and averaging over all atoms results in the nearest-neighbour distance p.d.f. for a randomly selected protein atom,

$$f(d_{\min}^{\rm obs}|2\sigma_m^2) = \frac{1}{N}\sum_{j=1}^{N} f(d_{\min,j}^{\rm obs}|2\sigma_m^2). \qquad (5)$$

The summation in (5) can be seen as a marginalization: the atom names and corresponding different chemical environments are integrated out. The resulting p.d.f. thus describes the occurrence of nearest-neighbour distances given a protein model and the variance of a Gaussian disturbance. For values of $2\sigma_m^2$ larger than zero, distribution (5) is dominated by a set of short 1–2, 1–3 and 1–4 distances. All protein structures consist of the same set of basic elements and distribution (5) can thus be expected to be essentially the same for all proteins. The nearest-neighbour distance distributions for various values of σ_m can be pre-computed. When a protein model with a Gaussian coordinate error is available, the set of nearest-neighbour distances $\{d_{\min}^{\rm obs}\}$ in the model can be computed. This observed set of distances can then be used to estimate the coordinate error *via* the maximization of the likelihood of the observed distances as a function of σ^2,

$$\mathcal{L}_{nndd}(\sigma^2|\{d_{\min}^{\rm obs}\}) = \prod_{i=1}^{N} f(d_{\min}^{\rm obs} = d_{\min,i}^{\rm obs}|2\sigma^2). \qquad (6)$$

2.3. Radial distance distribution

Let us denote the radial distance distribution of a single protein molecule by $f_{\rm rad}(d^{\rm tar})$. Upon a Gaussian coordinate error, the interatomic distances will change and the distribution will be smeared out:

$$f_{\rm rad}(d^{\rm obs}|2\sigma_m^2) = \int_0^\infty f_{\rm rad}(d^{\rm tar})\mathrm{NCM}(d^{\rm obs}|d^{\rm tar}, 2\sigma_m^2)\,\mathrm{d}d^{\rm tar}. \qquad (7)$$

If the radial distance distribution up to a certain distance ($d_{\rm radmax}$) can be assumed to be the same for all proteins, the blurring described by (7) can be used to estimate the coordinate error. Maximizing the likelihood of the observed interatomic distances $\{d_{\rm rad}^{\rm obs}\}$ given the radial distance distribution for a proposed value of σ_m^2 will result in the maximum-likelihood estimate of the variance of the Gaussian error model,

$$\mathcal{L}_{\rm rad}(\sigma^2|\{d_{\rm rad}^{\rm obs}\}) = \prod_{i=1}^{M} f_{\rm rad}(d^{\rm obs} = d_{{\rm rad},i}^{\rm obs}|2\sigma^2), \qquad (8)$$

where M is the number of the observed interatomic distances that are smaller than $d_{\rm radmax}$.

2.4. Map correlation and its relation to the r.m.s.d.

Let us define the map correlation (m.c.c.) by the linear correlation over the whole unit cell of the map computed with phases from a model with a coordinate error to the map computed from phases of the final model. Both maps are constructed using ($F_{\rm obs}$, $\varphi_{\rm calc}$) coefficients.

Let us assume that an electron density is modelled by a set of three-dimensional isotropic Gaussians $g(\mathbf{r}|\mathbf{r}_j, \omega^2)$ centred on $\mathbf{r}_j$ and with a width of 2ω,

$$\rho(\mathbf{r}) = \frac{1}{N}\sum_{j=1}^{N} g(\mathbf{r}|\mathbf{r}_i, \omega^2). \qquad (9)$$

Now assume that the electron density calculated from imperfect phases can be modelled using (9) with a Gaussian disturbance $\mathbf{q}$ (with variance σ_m^2) on the atomic centres,

$$\rho'(\mathbf{r}) = \frac{1}{N}\sum_{j=1}^{N} g(\mathbf{r}|\mathbf{r}_i + \mathbf{q}_i, \omega^2). \qquad (10)$$

The expression for the expected correlation α between these two electron-density functions is

$$\alpha = \frac{\mathbb{E}(\{\rho(\mathbf{r}) - \mathbb{E}[\rho(\mathbf{r})]_{\mathbf{r}}\}\{\rho'(\mathbf{r}) - \mathbb{E}[\rho'(\mathbf{r})]_{\mathbf{r}}\})_{\mathbf{r}}}{\left[\mathbb{E}\left(\{\rho(\mathbf{r}) - \mathbb{E}[\rho(\mathbf{r})]_{\mathbf{r}}\}^2\right)_{\mathbf{r}}\,\mathbb{E}\left(\{\rho'(\mathbf{r}) - \mathbb{E}[\rho'(\mathbf{r})]_{\mathbf{r}}\}^2\right)_{\mathbf{r}}\right]^{1/2}}. \qquad (11)$$

The expectation values are obtained by integration over $\mathbf{r}$, as denoted by the subscript. Ignoring effects arising from overlap of neighbouring atoms, it can be shown (Appendix *B*) that the correlation becomes equal to

$$\alpha = \frac{1}{N}\sum \exp\left(\frac{-q_i^2}{4\omega^2}\right) \simeq \int_0^\infty \exp\left(\frac{-q^2}{4\omega^2}\right)\mathrm{NCM}(q|0, \sigma_m^2)\;\mathrm{d}q. \qquad (12)$$

Working out (12) results in

$$\alpha = \left(1 + \frac{\sigma_m^2}{2\omega^2}\right)^{-3/2}. \qquad (13)$$

(12) relates the expected coordinate error, which is a function of σ_m^2, to the expected map correlation *via* the atomic shape parameter ω. (12) is in principle only valid for an isolated atom. Effects arising from atomic overlap can be taken into account in the following empirical way. An electron density for a reasonably large protein structure is constructed according to (9) with a given value of ω. The same protein structure is perturbed with Gaussian noise with a variance equal to σ_m^2 on the positional parameters and an electron density is created as performed for the unperturbed structure. If this procedure is carried out for a reasonable range of known ω and σ_m^2 values, the correlation α between the perturbed and unperturbed densities can be used to estimate an empirical relation between α, ω and σ_m^2. Using a wide range of ω and σ_m^2 values, the following dependency has been obtained:

$$\alpha = \left(1 + \frac{\sigma_m^2}{2.56\omega^{2.8}}\right)^{-1}. \qquad (14)$$

Thus, if an atomic shape parameter ω is known, (14) provides simple means of relating a coordinate error estimate of a model to a map correlation estimate for a map calculated with phases from that model. The procedure for the estimation of the parameter ω is described below.

2.5. Estimation of ω

From (2) and (14), it can be seen that the slope of the least-squares line fitted through a set of points {r.m.s.d.^{2}, $7.68(\alpha^{-1} - 1)$} is equal to $\omega^{2.8}$. Given the electron-density map of the final model and a number of electron-density maps computed with phases originating from protein models with known Gaussian errors, ω can be obtained *via* least-squares methods.

2ω thus models the width of the electron density of an atom with a Gaussian shape, without the contribution of neighbouring atoms. 2ω is likely to be influenced by the resolution of the X-ray data and the Wilson plot B value (Wilson, 1942,

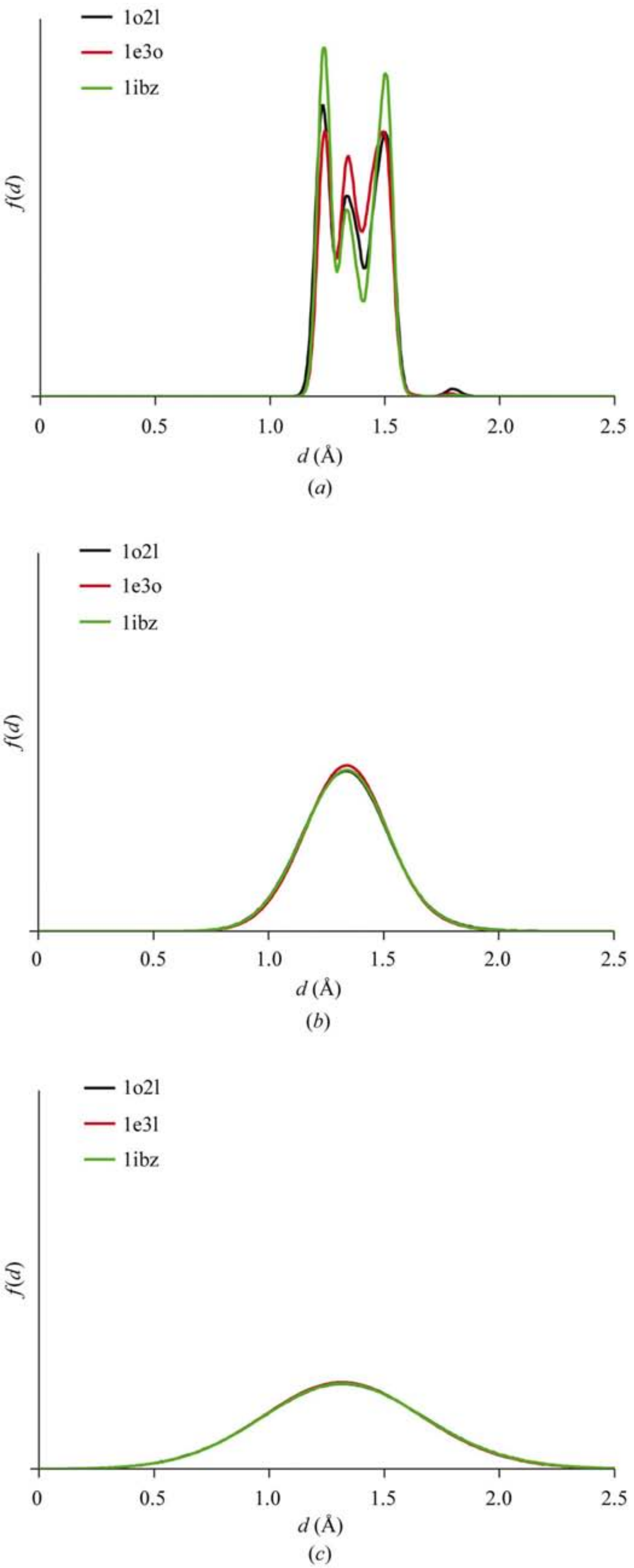


Figure 2
Nearest-neighbour distance distributions for three protein models (1o2l, 1e3o and 1ibz) for r.m.s.d. values equal to 0.02 (*a*), 0.2 (*b*) and 0.50 Å (*c*). For errors larger than 0.20 Å, the distributions for different models are essentially identical.

1949). Because 2ω conceptually models the atomic width, we suspect that it is related to the width of the Patterson origin peak, which was estimated by a procedure outlined in Appendix *C*. The width of the Patterson origin peak is furthermore linked to the optical resolution as defined by Vaguine (1999) and has been argued to be an objective measure of the expected level of detail in electron-density maps (Weiss, 2001). An empirical relation between the observed 2ω and the width of the Patterson origin peak can be used to estimate 2ω from the available diffraction data.

Another approach is to construct an empirical relation that would give an estimation of the value of ω^2 on the basis of the characteristics of the X-ray data set. We have chosen a simple polynomial function with the nominal resolution $d_{\rm min}$ and the Wilson plot B value $B_{\rm Wil}$ as variables,

$$\omega^2 = (a_1 d_{\rm min} + a_2 B_{\rm Wil}^{1/2} + a_3)^2. \qquad (15)$$

The coefficients a_1 and a_2 are the weights to the contributions of the nominal resolution and the average atomic displacement factor on the blurring of the electron density. Coefficient

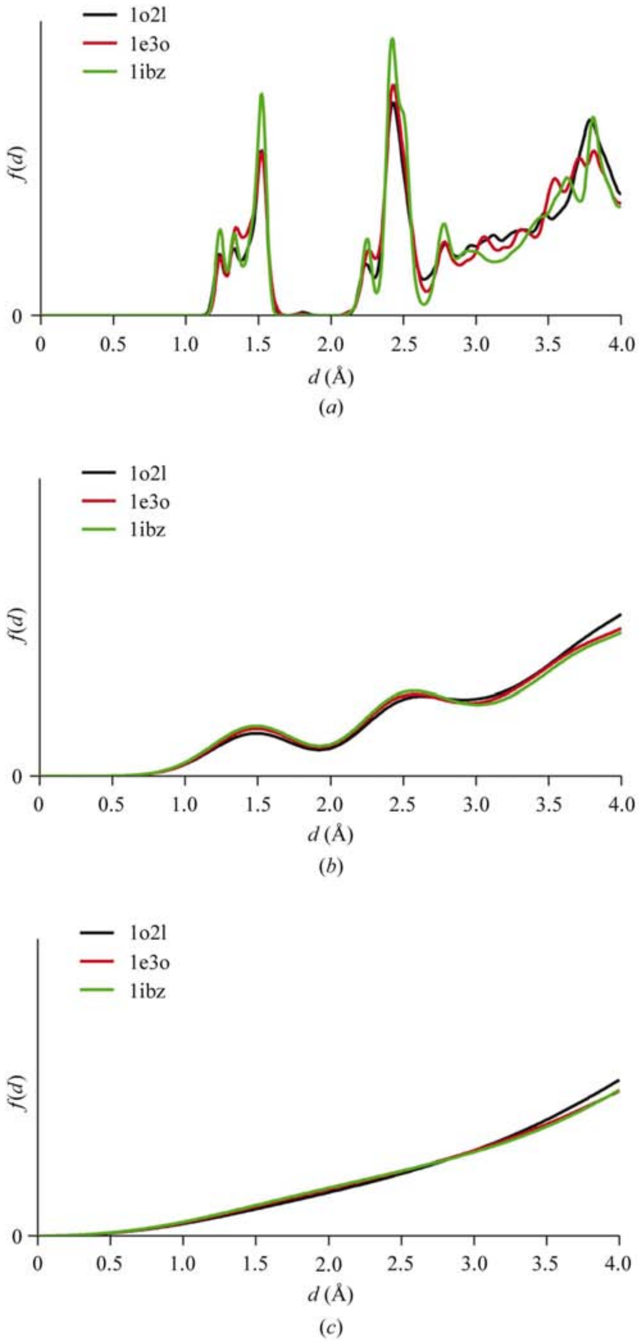


Figure 3
Radial distance distribution for three protein models (1o2l, 1e3o and 1ibz) for r.m.s.d. values equal to (*a*) 0.02, (*b*) 0.20 and (*c*) 0.50 Å. As for the nearest-neighbour distributions, the radial distance distributions are essentially identical for different protein models for coordinate errors larger than 0.20 Å.

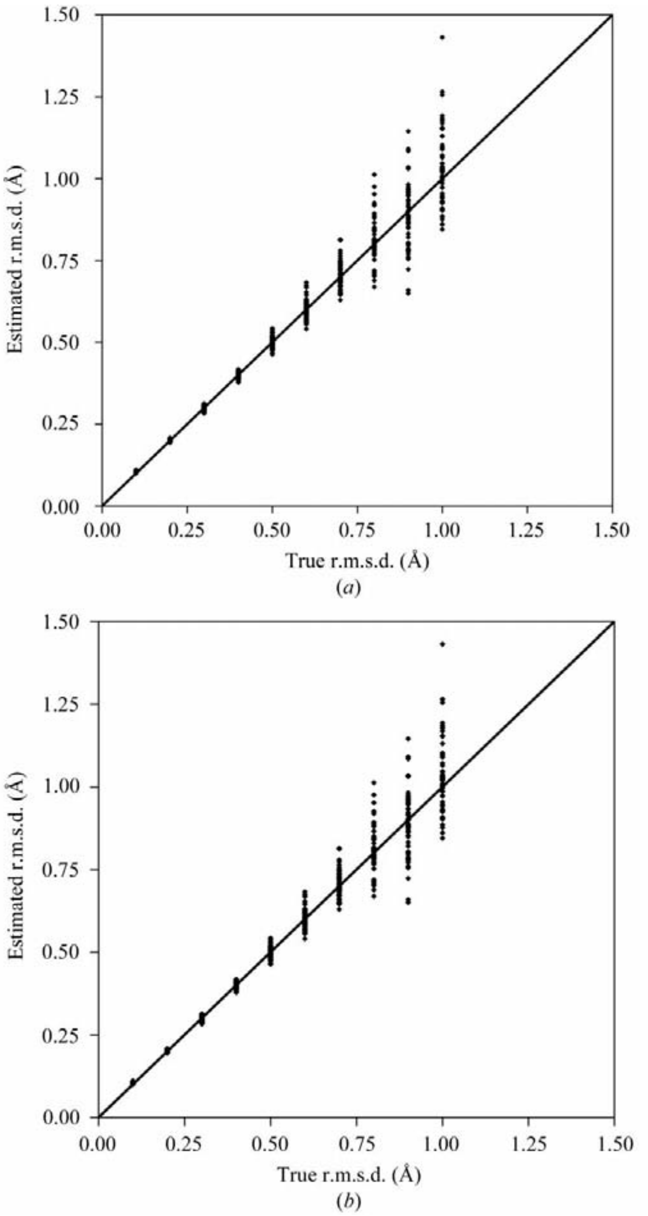


Figure 4
The dependence of the estimated r.m.s.d. values *via* nearest-neighbour (*a*) and radial distance distributions (*b*) to the true r.m.s.d. (simulated Gaussian errors, 50 independent randomizations per r.m.s.d. bin). A line with a slope of 1 and intercept equal to 0 is shown for comparison.

a_3 can be seen as modelling the average width of an atom at rest at infinite resolution.

3. Results

3.1. Nearest-neighbour and radial distance distributions

Nearest-neighbour distance distributions have been obtained numerically for three different protein structures (PDB codes 1o2l, 1e3o and 1ibz) *via* expression (5) and are shown in Fig. 2 for three different r.m.s.d. values. The distributions are very similar at moderate to high r.m.s.d. values, but differ for lower errors. This observation has been confirmed by calculating the Kolmogorov–Smirnov statistic (Dudewicz & Mishra, 1988) between the different distributions at a number of r.m.s.d. values, indicating that the distributions of the three proteins can be regarded as equal for coordinate errors larger than 0.20 Å.

Error-dependent radial distance distributions up to 4 Å were obtained numerically for the same protein structures *via* expression (7). A number of these distributions is shown in Fig. 3.

To test the use of the error-dependent nearest-neighbour distributions, the atomic model of crambin (PDB code 1ab1) was randomized with different r.m.s.d. values. The nearest-neighbour distances of the resulting models were computed and were used in the likelihood-maximization procedure described by (6). The results are plotted in Fig. 4. Estimates of the coordinate error using the radial distance distribution *via* expression (8) are also shown.

3.2. Determination of ω

The atomic shape parameter ω has been estimated by randomizing the atoms of a well refined structure and using that coordinate set to calculate phases and a ‘scrambled’ electron-density map. This has been carried out for a number of different r.m.s.d. values and the resulting map correlations to the original map have been calculated (Fig. 5). This procedure has been carried out for 69 structures with experimental X-ray data (see Table 4 for a summary of some statistics) downloaded from the PDB (Berman *et al.*, 2000; Bernstein *et al.*, 1977). For each structure, ω has been determined (hereafter denoted as observed ω) using expression (14). A typical plot of observed and fitted map correlations as a function of r.m.s.d. is shown in Fig. 5.

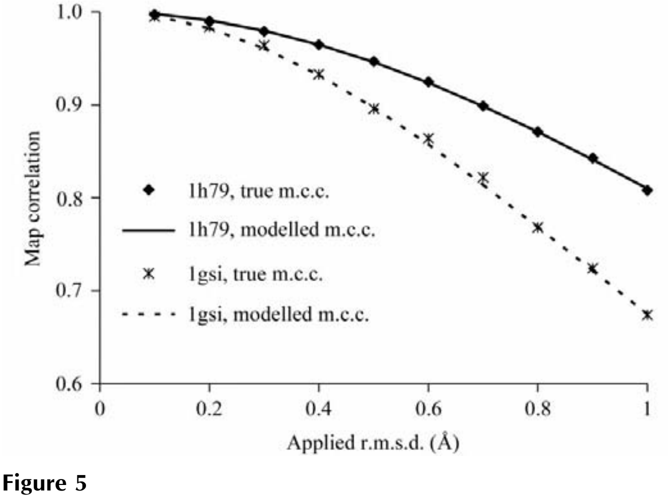


Figure 5
Map correlation as a function of r.m.s.d. for two data sets: 1h79 ($d_{\min}$ = 2.9 Å, $B_{\rm Wil}$ = 51.3 Å^2) and 1gsi ($d_{\min}$ = 1.6 Å, $B_{\rm Wil}$ = 15.0 Å^2).

Table 4
Summary statistics of the 69 data sets used in the estimation of ω.

	Average	Minimum	Maximum
Resolution $d_{\min}$ (Å)	2.05	1.35	2.91
Wilson plot B value $B_{\rm Wil}$ (Å^2)	24.5	7.24	74
Overall completeness (%)	95.1	79.5	99.9

The estimated values of 2ω from the isolated atom approximation (13) and those obtained from the overlapping atom approximation (14), are plotted in Fig. 6 against the width of the Patterson origin peak. The width of the Patterson origin peak is affected by atomic electron-density overlap, resulting in a broader width than expected from the width of a single atom. Therefore, we expect the width of the Patterson origin peak to be larger than or equal to 2ω. The values of 2ω obtained *via* expression (14) reflect this expecation surprisingly well. For this reason, any further reference to 2ω is thus based on the estimates obtained *via* expression (14). Least-squares fitting of the parameters a_1, a_2, a_3 in expression (15), given the nominal resolution quoted in the 69 PDB entries and the Wilson plot B value as determined by *ARP/wARP* (Morris, Zwart *et al.*, 2003), resulted in

$$\omega^2 = (0.078 d_{\min} + 0.043 B_{\rm Wil}^{1/2} + 0.322)^2 \qquad (16)$$

with estimated standard deviations of 0.014, 0.004 and 0.021, respectively. A plot of the observed ω^2 values *versus* the values

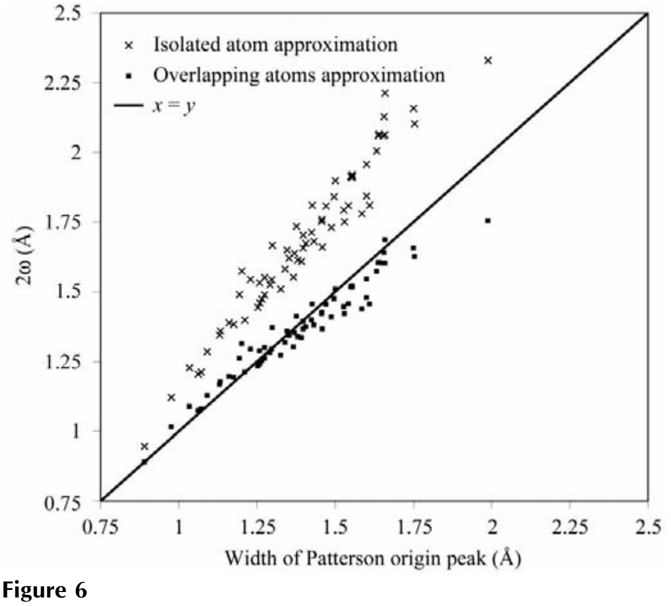


Figure 6
The dependence of the 2ω estimates based on the isolated and overlapping atoms approximations on the width of the Patterson origin peak. A line with a slope equal to 1 is shown for comparison.

predicted *via* expression (16) is shown in Fig. 7.

In a further simplification where we use the observed dependency of $B_{\rm Wil}^{1/2}$ on $d_{\min}$ for 69 PDB entries, (16) reduces to

$$\langle 2\omega\rangle = 0.34d_{\min} + 0.64 \qquad (17)$$

and has a correlation of 0.82 against the observed 2ω. A plot of the nominal resolution *versus* $\langle 2\omega\rangle$, the resolution measure according to James (1948), Stenkamp & Jensen (1984) and the width of the Patterson origin peak, is shown in Fig. 8.

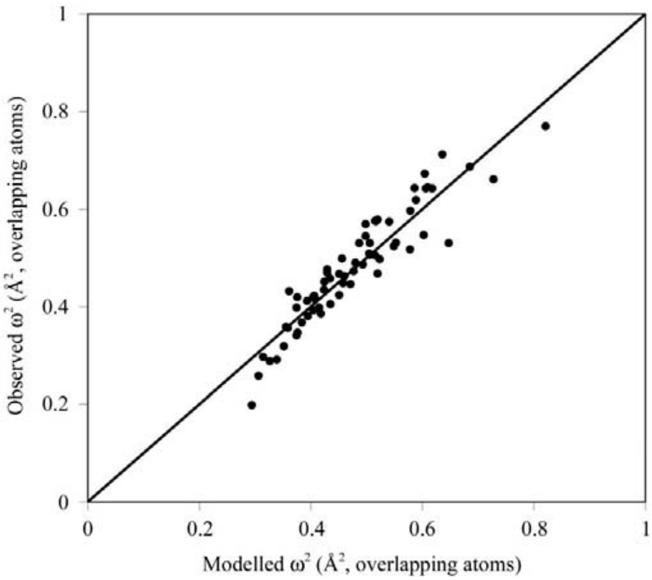


Figure 7
Observed ω^2 *versus* modelled ω^2 using the overlapping atom approximation, fitted on the basis of the Wilson plot *B* value and the resolution of the data set. The correlation coefficient is 0.93.

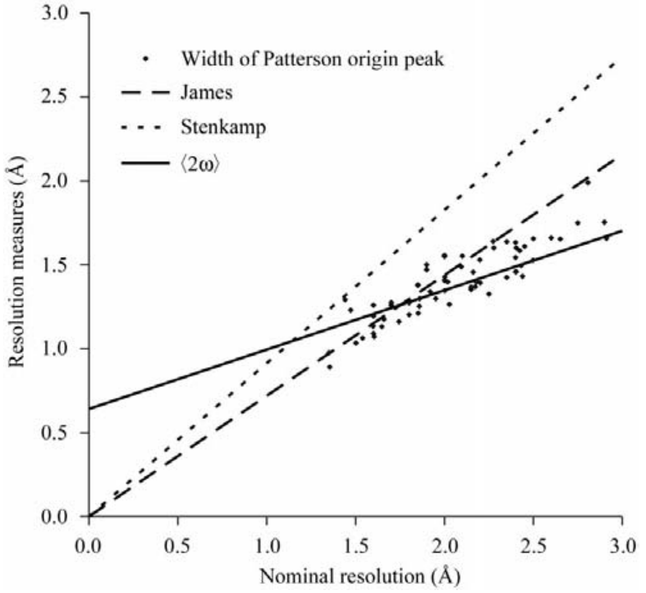


Figure 8
The nominal resolution as quoted in the 69 PDB entries is plotted against a number of different resolution measures. See main text for details.

Table 5
R.m.s.d. and map correlation estimates of scrambled and subsequently refined models.

(*a*) PSP data set. The resolution range was 20–2.0 Å. The Wilson plot *B* value was 18.0 Å^2, which corresponds to a value of 1.32 Å for 2ω.

True r.m.s.d.	0	0.19	0.37	0.56	0.77
Estimated r.m.s.d.	0.03	0.17	0.32	0.52	0.76
True m.c.c.	1.00	0.99	0.96	0.90	0.79
Estimated m.c.c.	1.00	0.99	0.95	0.90	0.80

(*b*) Adaptin data set. The resolution range was 20–1.45 Å, the Wilson plot *B* value was equal to 10.5 Å^2 and 2ω is thus equal to 1.15 Å.

True r.m.s.d.	0	0.15	0.33	0.56	0.80
Estimated r.m.s.d.	0.20	0.22	0.33	0.76	1.03
True m.c.c.	1.0	0.99	0.94	0.84	0.69
Estimated m.c.c.	0.98	0.97	0.94	0.74	0.60

Table 6
R.m.s.d. and map correlation estimates in the final *ARP* cycle.

The r.m.s.d. was estimated using the nearest-neighbour distances and converted to a map correlation estimate as described in the text. 'True r.m.s.d.' stands for an r.m.s.d. estimate obtained using the known map correlation and the inverse of expression (14).

	PSP	Adaptin
True r.m.s.d.	0.43	0.40
Estimated r.m.s.d.	0.45	0.33
True m.c.c.	0.93	0.91
*REFMAC*5 m.c.c.	0.98	0.95
Estimated m.c.c.	0.92	0.94

3.3. Map correlation estimates *via* an r.m.s.d. estimate of unrestrained refined atoms

A number of unrestrained refinements were carried out on scrambled models of leishmanolysin (PSP; courtesy of P. Metcalf) and the γ-adaptin appendage domain (adaptin; courtesy of S. Panjikar and H. M. Kent). A Gaussian error was applied to the positional parameters and the model has been subsequently refined without restraints using *REFMAC*5 (Murshudov *et al.*, 1997). The r.m.s.d. of the coordinate set has been estimated using the nearest-neighbour distance distributions and the corresponding map correlations *via* expressions (16) and (14). The results of these experiments are listed in Table 5.

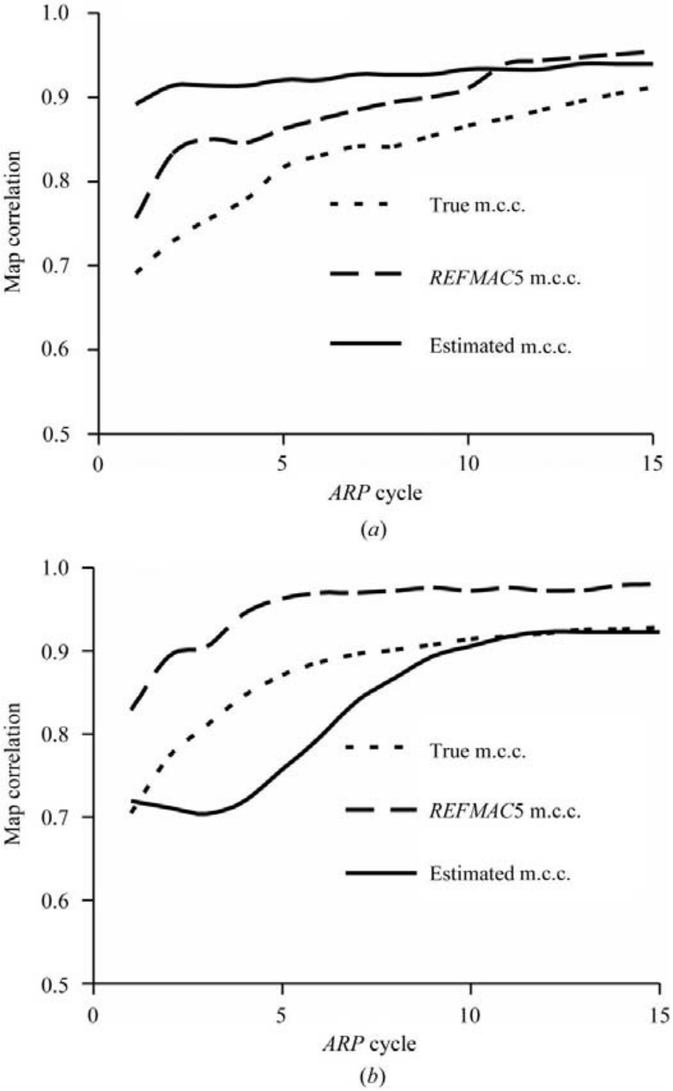
Furthermore, two free-atom modelling experiments have been carried out. The intermediate models of these runs have been used to monitor the progress of phase improvement. The r.m.s.d. and the map correlation were estimated as performed for the scrambled models. Figure-of-merit estimates from *REFMAC*5 were used to estimate the map correlation, with the aid of an expression derived by Lunin & Woolfson (1993), albeit in a modified form:

$$\alpha = \frac{\sum |\mathbf{F}_h|^2 {\rm fom}_{\mathbf{h}}}{\sum |\mathbf{F}_h|^2}. \qquad (18)$$

The results are shown in Fig. 9 and summarized for the last free-atom modelling cycle in Table 6.

4. Discussion and conclusions

The presented analytical distribution of an interatomic distance given the target distance and a Gaussian error model serves as an essential component in modelling of distance distributions in proteins. The non-central Maxwell distribution can be described reasonably well by a Gaussian when the variance is small compared with the error-free distance, as indicated by the mean and second raw moment in Table 3. For errors that are large compared with the error-free distance, the Gaussian approximation becomes inadequate (Fig. 1). Similar observations have also been made when comparing the Rice and Gaussian distribution (Pannu & Read, 1996; Bricogne, 1997).



(*a*)

(*b*)

Figure 9
Map correlation estimates *via* an r.m.s.d. estimate (estimated m.c.c.), *via* the figures of merit of *REFMAC*5 (*REFMAC*5 m.c.c.) and true map correlation (true m.c.c.) as a function of the *ARP* cycle. (*a*) Adaptin, (*b*) PSP.

Both the derived theoretical nearest-neighbour and radial distance distributions as a function of the coordinate error model the true Gaussian errors of the positional parameters extremely well (Fig. 4) in spite of the approximation of the independence of distances, which is not fully justified. The average relative difference between predicted and true r.m.s.d. is below 12% for errors up to 1.0 Å. The error estimates of the refined randomized models also lie close to the correct values (Table 5), indicating that the errors in the positional parameters follow an (approximate) Gaussian distribution. The error estimates on the free-atom models, both the r.m.s.d. and map correlation, are not quite correct in the early stages of the iterative model update and refinement procedure (Fig. 9). However, the results improve as the free-atom modelling procedure comes close to convergence (Table 6). The reason for this is ascribed to specifics of *ARP*/*wARP*: the presence of noise atoms and the distance limits used while placing free atoms in the electron density with underlying errors that are not accounted for. Clearly, if the initial free-atom model is built on the grid points on which the electron density is calculated, the distribution of the nearest neighbours is primarily governed by the grid rather than by the quality of the phases. Use of the statistical characteristics of these *lattice distributions* (Abramovicz & Stegun, 1974) in automated model-building procedures will be presented elsewhere.

Addition of tight stereochemical restraints on the model will of course bias the set of distances towards the geometrical targets. This will result in an underestimation of the coordinate error using either the nearest-neighbour or radial distance distributions. Nevertheless, some direct-method approaches based on real/reciprocal-space recycling (*e.g.* Miller *et al.*, 1993; Usón & Sheldrick, 1999) or conditional optimization approaches (Scheres & Gros, 2001) might benefit from the error analyses presented here.

The expression derived for the dependence of the map correlation on the r.m.s.d. is a crude approximation to the reality. The electron-density fall-off at truncated resolutions shows a sharper drop as a function of distance from the atomic centre than can be modelled by a Gaussian function (Chapman, 1995). Furthermore, an electron-density map with errors in the phases cannot be appropriately described by a simple shift of the atomic centres while keeping the atomic shape constant. Surprisingly, however, the (empirical) functional forms obtained are good enough to obtain a workable model that allows the prediction of the map correlation within reasonable accuracy.

As expected, 2ω does correlate with the width of the Patterson origin peak (Fig. 6). A relation with the optical resolution as computed by *SFCHECK* is also present, but has a lower correlation (0.93 instead of 0.97), most likely owing to the fact that the optical resolution has an added nominal resolution-dependent term. We regard 2ω as a model of the shape of an average protein atom that is linked *via* a simple expression (14) to the sensitivity of the map correlation to Gaussian positional errors. Expression (17) is simple but rather approximate, as it does not take the Wilson plot B value directly into account but rather relies on the correlation between the Wilson plot B value and the nominal resolution in the data present in our test set. Estimating the value of 2ω should rather be carried out by use of expression (16) or *via* its relation to the width of the Patterson origin peak.

When a reasonably accurate estimate of the map correlation is thus available, this can be converted to an estimate of the r.m.s.d. of the free atoms using the inverse of expression (14), as can be seen from Table 6. The r.m.s.d. estimate defines the geometrical part of the error model which can be used in the scoring function in main-chain tracing or side-chain building. The use of the nearest-neighbour distance distribution in estimating the quality of a free-atom model in the early stages of model construction is limited, but can possibly be used in assisting atom update and removal or as a source of prior information in non-grid-based free-atom model construction. Using the empirically derived quantity 2ω as a classifier for an X-ray data set to choose appropriate density templates during chain tracing and the construction of tailor-made decision boundaries as a function of map quality and data-set characteristics is currently being implemented in the latest version of *ARP*/*wARP* and awaits thorough testing to validate the results.

The 69 X-ray data sets used to obtain the coefficients for (16) and (17) can certainly be used for more elaborate analyses of characteristics of the electron density. Impressive work along these lines has already been carried out by Kleywegt *et al.* (2003), who constructed a database of electron-density maps and quality indicators for local density fits given atomic models and corresponding X-ray data submitted to the PDB.

APPENDIX *A*
The non-central Maxwell distribution

Let $f(x, y, z)$ be the p.d.f. of a three-dimensional spherical Gaussian in an orthonormal coordinate framework centred on an arbitrary vector (x_{jk}, y_{jk}, z_{jk}),

$$f(x,y,z) = \frac{1}{2\pi\sigma^3(2\pi)^{1/2}} \times \exp\left[-\frac{(x-x_{jk})^2+(y-y_{jk})^2+(z-z_{jk})^2}{2\sigma^2}\right]. \tag{19}$$

Transforming this p.d.f. to polar coordinates results in

$$f(r,\theta,\psi) = \frac{r^2\sin(\theta)}{2\pi\sigma^3(2\pi)^{1/2}}\exp\left[-\frac{r^2+r_{jk}-2r_{jk}r\cos(\theta)}{2\sigma^2}\right]. \tag{20}$$

Integrating out the angles [note that ψ is uniformly distributed over $(0, 2\pi)$] results in the distribution of the length of a vector drawn from a three-dimensional Gaussian centred on x_{jk}, y_{jk}, z_{jk},

$$f(r|r_{jk},\sigma^2) = \frac{1}{\sigma(2\pi)^{1/2}}\exp\left[-\frac{(r-r_{jk})^2}{2\sigma^2}\right]\frac{r}{r_{jk}}\left[1-\exp\left(-\frac{2r_{jk}r}{\sigma^2}\right)\right]. \tag{21}$$

Subsequently setting $\sigma^2 = \sigma_j^2 + \sigma_k^2$ by assumed independence in the errors of the atoms results in expression (1). The (cumulative) distribution function of the non-central Maxwell distribution (21) is given as

$$F(r) = \frac{\sigma}{(2\pi)^{1/2}r_{jk}}\left\{\exp\left[-\frac{(r_{jk}+r)^2}{2\sigma^2}\right]-\exp\left[-\frac{(r_{jk}-r)^2}{2\sigma^2}\right]\right\} + \frac{1}{2}\left[\mathrm{Erf}\left(\frac{r-r_{jk}}{2^{1/2}\sigma}\right)+\mathrm{Erf}\left(\frac{r+r_{jk}}{2^{1/2}\sigma}\right)\right]. \tag{22}$$

The K-dimensional generalization of (21) is known as the generalized Rice distribution and has the following p.d.f.:

$$f_{r_{\rm obs}}(r|\sigma^2,r_{jk}) = \frac{r}{2\sigma^2}\left(\frac{r}{r_{jk}}\right)^{\frac{K-2}{4}}\exp\left(\frac{r^2+r_{jk}^2}{2\sigma^2}\right)I_{\frac{K}{2}-1}\left(\frac{r_{jk}r}{\sigma^2}\right), \tag{23}$$

where $I_{\frac{K}{2}-1}$ denotes a modified Bessel function of the first kind. Depending on whether the dimension of the system is odd or even, it is of fractional or integer order, respectively. Setting $K = 3$ results, after some rearranging, in expression (21). The raw moments of the generalized Rice distribution are given for completeness:

$$E(r^\nu|K)_r = (2\sigma^2)^{\nu/2}\frac{\Gamma[(K+\nu)/2]}{\Gamma(K/2)}\Phi\left(-\frac{\nu}{2},\frac{K}{2};-\frac{r_{jk}^2}{2\sigma^2}\right). \tag{24}$$

It can be shown that when $K = 2$, (24) reduces to the moments of the Rice distribution given by Pannu & Read (1996). For $K = 3$, (24) is equivalent to the expression for the raw moments given in Table 3.

APPENDIX *B*
On the derivation of expression (12)

Let us describe an atomic density by a Gaussian $g(\mathbf{r}|\mathbf{q}, \omega^2)$ placed on the atomic centre $\mathbf{q}$ and expand it along the x, y and z directions:

$$g_q(\mathbf{r}|\mathbf{q},\omega^2) = g_{x,q}(x|q_x,\omega^2)g_{y,q}(y|q_y,\omega^2)g_{z,q}(z|q_y,\omega^2). \tag{25}$$

The following shorthand will be used:

$$g_{x,q_x}(x|q_x,\omega^2) = g_{x,q_x}. \tag{26}$$

Let x, y and z be distributed uniformly over $(-a, a)$. If a is large, then

$$\mathbb{E}(g_{x,q_x})_x = \int_{-a}^{+a}\frac{1}{2a}g_{x,q_x}\,\mathrm{d}x \simeq \frac{1}{2a}, \tag{27}$$

$$\mathbb{E}[(g_{x,q_x})^2]_x = \int_{-a}^{+a}\frac{1}{2a}(g_{x,q_x})^2\,\mathrm{d}x \simeq \frac{1}{4a\pi^{1/2}\omega} \tag{28}$$

and

$$\mathbb{E}[(g_{x,q_x})(g_{x,0})]_x = \int_{-a}^{+a}\frac{1}{2a}g_{x,q_x}g_{x,0} \simeq \frac{1}{4a\pi^{1/2}\omega}\exp\left(-\frac{q_x^2}{4\omega^2}\right). \tag{29}$$

Expanding (11) results in

$$\alpha = \frac{\mathbb{E}[\rho(\mathbf{r})\rho'(\mathbf{r})]_{\mathbf{r}} - \mathbb{E}[\rho(\mathbf{r})]_{\mathbf{r}}\mathbb{E}[\rho'(\mathbf{r})]_{\mathbf{r}}}{\left(\{\mathbb{E}[\rho(\mathbf{r})^2]_{\mathbf{r}} - \mathbb{E}[\rho(\mathbf{r})]_{\mathbf{r}}^2\}\{\mathbb{E}[\rho'(\mathbf{r})^2]_{\mathbf{r}} - \mathbb{E}[\rho'(\mathbf{r})]_{\mathbf{r}}^2\}\right)^{1/2}}. \quad (30)$$

Writing out the first part of the covariance term in (30) explicitly for a single -atom molecule, we obtain

$$\mathbb{E}[\rho(\mathbf{r})\rho'(\mathbf{r})]_r = \mathbb{E}(g_{x,0}g_{x,q_x})_x\mathbb{E}(g_{y,0}g_{y,q_y})_y\mathbb{E}(g_{z,0}g_{z,q_y})_z. \quad (31)$$

Working out the other moments and using (27) and (28) results in

$$\alpha = \frac{[1/(\pi^{1/2}\omega)^3]\exp[-(q_x^2+q_y^2+q_z^2)/4\omega^2] - a^{-2}}{[1/(\pi^{1/2}\omega)^3] - a^{-2}}. \quad (32)$$

Neglecting the a^{-2} terms (a is large) results in

$$\alpha = \exp\left(-\frac{q_x^2+q_y^2+q_z^2}{4\omega^2}\right) = \exp\left(-\frac{|\mathbf{q}|^2}{4\omega^2}\right). \quad (33)$$

The validity of the previous expression has been checked numerically. The larger a, the more justified the approximations made. For a multi-atom molecule a similar derivation can be made, resulting in expression (12), assuming that atoms do not overlap.

APPENDIX *C*
The width of the Patterson origin peak

Consider the Patterson synthesis:

$$P(\mathbf{u}) = \sum_{\mathbf{h}}|\mathbf{F}_{\mathbf{h}}|^2\exp(-2\pi i\mathbf{hu}). \quad (34)$$

Averaging over the sphere with radius u results in

$$\langle P(u)\rangle = \sum_h|\mathbf{F}_{\mathbf{h}}|^2\frac{\sin(2\pi hu)}{2\pi hu}. \quad (35)$$

This can be approximated further by summing over resolution bins h_j, rather than individual reflections,

$$\langle P(u)\rangle = \sum_j N_j\langle|\mathbf{F}_{\mathbf{h}}|^2\rangle_j\frac{\sin(2\pi h_ju)}{2\pi h_ju}. \quad (36)$$

Modelling the Patterson origin peak by a Gaussian function leaves us with fitting the following dependency around the origin,

$$\ln[\langle P(u)\rangle] - \ln[\langle P(0)\rangle] = p_1u^2. \quad (37)$$

$(-1/p_1)^{1/2}$ is than equal to the width of an average Gaussian-shaped atom. This number is quoted in the text as the width of the Patterson origin peak.

The authors would like to thank A. Schmidt, R. J. Morris, G. Murshudov and the referees for useful comments on this manuscript.

References

Abramovicz, M. & Stegun, I. A. (1974). *Handbook of Mathematical Functions.* New York: Dover.

Arfken, G. B. & Weber, H. J. (1995). *Mathematical Methods for Physicists.* San Diego: Academic Press.

Andersen, A. H. & Kirsch, J. E. (1993). *Med. Phys.* **23**, 857–869.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

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.

Bricogne, G. (1997). *Methods Enzymol.* **276**, 361–423.

Colovos, C., Toth, E. A. & Yeates, T. O. (2000). *Acta Cryst.* D**56**, 1421–1429.

Cowtan, K. (2002). *J. Appl. Cryst.* **35**, 655–663.

Cox, E. G. & Cruickshank, D. W. J. (1948). *Acta Cryst.* **1**, 92–93.

Cruickshank, D. W. J. (1999). *Acta Cryst.* D**55**, 583–601.

Chapman, M. S. (1995). *Acta Cryst.* A**51**, 69–80.

Dudewicz, E. J. & Mishra, S. N. (1988). *Modern Mathematical Statistics.* New York: Wiley.

James, R. W. (1948). *Acta Cryst.* **1**, 132–134.

Kleywegt, G. J., Harris, M. R., Zou, J. Y., Taylor, T. C., Wählby, A. & Jones, T. A. (2003). *Uppsala Electron Density Server*, http://fsrv1.bmc.uu.se/eds/.

Lebedev, N. N. (1972). *Special Functions and Their Applications*, translated and edited by R. A. Silverman. New York: Dover.

Lunin, V. Yu. & Skovodora, T. P. (1995). *Acta Cryst.* A**51**, 880–887.

Lunin, V. Yu. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 530–533.

Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.

Miller, R., DeTitta, G. T., Jones, R., Langs, D. A., Weeks, C. M. & Hauptman, H. A. (1993). *Science*, **259**, 1430–1433.

Morris, R. J., Perrakis, A. & Lamzin, V. S. (2002). *Acta Cryst.* D**58**, 968–975.

Morris, R. J., Perrakis, A. & Lamzin, V. S. (2003). In the press.

Morris, R. J., Zwart, P. H., Cohen, S., Fernandez, F. J., Kakaris, M., Kirillova, O., Vonrhein, C., Perrakis, A. & Lamzin, V. S. (2003). Submitted.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.

Pannu, N. S. & Read, R. J. (1996). *Acta Cryst.* A**52**, 659–668.

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

Read, R. J. (1990). *Acta Cryst.* A**46**, 900–912.

Scheres, H. W. S. & Gros, P. (2001). *Acta Cryst.* D**57**, 1820–1828.

Stenkamp, R. E. & Jensen, L. H. (1984). *Acta Cryst.* A**40**, 251–254.

Terwilliger, T. (2000). *Acta Cryst.* D**57**, 1755–1762.

Terwilliger, T. (2003) *Acta Cryst.* D**59**, 38–44.

Usón, I. & Sheldrick, G. M. (1999). *Curr. Opin. Struct. Biol.* **9**, 643–648.

Vaguine, A. (1999). *Acta Cryst.* D**55**, 191–205.

Weiss, M. (2001). *J. Appl. Cryst.* **34**, 130–135.

Weisstein, E. (1999). *CRC Concise Encyclopedia of Mathematics.* New York: Chapman & Hall.

Wilson, A. J. C. (1942). *Nature (London)*, **150**, 152.

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

Zwart, P. H. & Lamzin, V. S. (2003). *Acta Cryst.* D. In the press.