Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Percentile-based spread: a more accurate way to compare crystallographic models

Edwin Pozharski

University of Maryland, Baltimore, USA

Correspondence e-mail: epozhars@rx.umaryland.edu

Received 19 April 2010
Accepted 13 July 2010

The comparison of biomacromolecular crystal structures is traditionally based on the root-mean-square distance between corresponding atoms. This measure is sensitive to the presence of outliers, which inflate it disproportionately to their fraction. An alternative measure, the percentile-based spread (p.b.s.), is proposed and is shown to represent the average variation in atomic positions more adequately. It is discussed in the context of isomorphous crystal structures, conformational changes and model ensembles generated by repetitive automated rebuilding.

1. Introduction

Minimizing the root-mean-square difference between two sets of data is the most common tool of scientific data analysis. In structural biology, it is commonly used when comparing structural models. Proteins with the same fold, proteins crystallized in different forms and proteins undergoing conformational changes are some examples when similarity is assessed by root-mean-square (r.m.s.) difference. Two major reasons can be suggested as to why this particular measure is used almost exclusively.

Firstly, a fast and reliable algorithm exists for minimizing the r.m.s. difference between two sets of spatial points (Kabsch, 1976); it is at the core of any modern structural superposition protocol. Most of the effort in improving such methods has been directed at finding better ways to define the corresponding sets of atoms. Secondly, r.m.s. deviation is used in statistics to describe random variables and thus the r.m.s. difference between superimposed structures supposedly reflects the underlying variation between two sets of data. It is often compared in the crystal structure context with the estimated standard uncertainty of the structure determination itself. In this way, the r.m.s. distance between individual atoms is judged to reflect actual structural differences when it exceeds experimental error. For instance, when no significant structural differences are expected [as in the case of isomorphous crystals (Rashin *et al.*, 2009) or iterative refinement (Terwilliger *et al.*, 2007)], the r.m.s. distance is perceived to reflect the precision of the structural model.

This interpretation of the r.m.s. difference between two structural models is based on the implicit assumption that the underlying probability distribution is normal. Accordingly, it is expected that the r.m.s. difference between superimposed structures, ΔR, determines the corresponding probability of the distance, Δr, between particular atoms. For instance, if the overall $\Delta R = 0.2$ Å, one would normally assume that there is only a 5% chance of observing an atom shift by 0.4 Å or more. It is common in science to interpret the r.m.s. difference between two instances of a multi-parameter model as the

defining size of the element of parameter space to which a certain fraction of measurements are confined.

Such assumptions fail in the presence of outliers that violate the normal distribution and certain corrections are required to account for the three-dimensional nature of the structural models. Furthermore, outliers inflate the estimate of an r.m.s. difference disproportionately to their fraction in the ensemble of measurements. In application to macromolecular structures this means that calculated r.m.s. differences between superimposed structures are misleading and significantly overestimate the effective width of pairwise distance distributions. An alternative measure of structural variation based on percentile analysis is proposed and applied to model superposition.

2. Results and discussion

2.1. Expected distribution of pairwise distances

It is assumed that individual differences between atomic coordinates (x, y, z) after structures have been superimposed are distributed normally with a singular standard deviation $\sigma_x = \sigma_y = \sigma_z$. Derivation of the three-dimensional distribution is trivial and results in the Maxwell–Boltzmann distribution ($\sigma_r^2 = 3\sigma_x^2$; Chambers & Stroud, 1979)

$$p(r) = \frac{4\pi r^2}{(2\pi\sigma_r^2/3)^{3/2}} \exp\left(-\frac{3r^2}{2\sigma_r^2}\right) \quad (1)$$

and the corresponding cumulative distribution function is

$$P(r) = \mathrm{erf}\left(\frac{r}{\sigma_r} 1.5^{1/2}\right) - \left(\frac{6}{\pi}\right)^{1.2} \frac{r}{\sigma_r} \exp\left(-\frac{3r^2}{2\sigma_r^2}\right). \quad (2)$$

This distribution has a maximum density which is shifted from the origin (the most likely interatomic distance is $\sigma_r/1.5^{1/2}$). This is owing to the smaller volume available near the origin. The cumulative distribution function (2) is compared in Fig. 1 with the one-sided normal distribution with the same σ_r. It should be noted that the average interatomic distance is

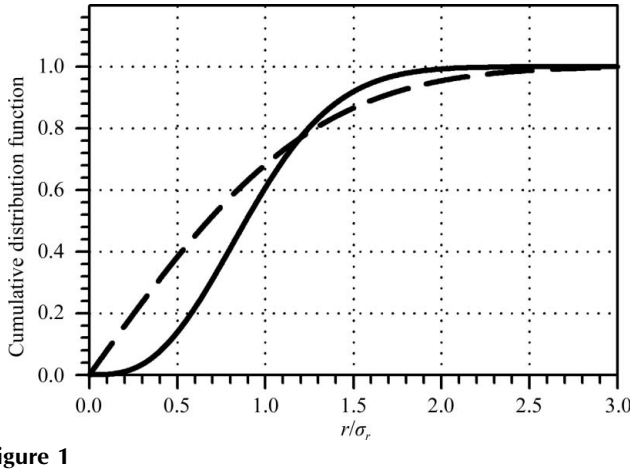


Figure 1
Cumulative distribution function according to (2). The normal distribution (dashed line) is shown for comparison.

$$\langle r \rangle = 4\sigma_r/(6\pi)^{1/2} \simeq 0.92\sigma_r > 0 \quad (3)$$

and the r.m.s. distance is simply equal to σ_r.

2.2. The r.m.s. deviation is sensitive to the presence of outliers

A small fraction of outliers can significantly inflate the apparent r.m.s. deviation owing to the quadratic nature of the measure. In a somewhat simplified form, let us assume that the series of measurements contain what can be described as two subgroups of random variables with the same mean of zero and two different variances, σ_1 and σ_2. The relative fraction of the second variable of higher variance is φ and the overall variance is

$$\sigma = [(1-\varphi)\sigma_1^2 + \varphi\sigma_2^2]^{1/2}. \quad (4)$$

The 'minority' variable represents a group of outliers, which in the context of structural comparison may be a part of the structure in which significant differences are observed that exceed those originating from experimental error. The amplitude of outliers is characterized by the Z score $Z = \sigma_2/\sigma_1$ and the variance inflation parameter is defined as $p = \sigma/\sigma_1$. Evidently,

$$p = [1 + \varphi(Z^2 - 1)]^{1/2}. \quad (5)$$

The inflation parameter approaches unity when $\varphi \ll 1/Z^2$. Fig. 2 shows how the variance is inflated with increasing fraction of outliers for various Z values. Values as high as $Z = 10$ are quite expected in structural comparison; for example, the variation driven by experimental error may be as low as 0.1 Å, whereas some elements of the structure may differ by as much as 1 Å. It is noteworthy that even a small fraction of such outliers can significantly inflate the overall r.m.s. distance.

If these results are placed in the context of macromolecular structure comparison it becomes clear that most models will have inflated r.m.s. differences. The overall statistical error of crystal structures is estimated by the Cruickshank DPI (Blow, 2002; Murshudov & Dodson, 1997) and in most cases is ~0.1 Å. With about 10% of residues deviating by an r.m.s. distance of 0.6 Å, the overall r.m.s. difference will double. As a result, the overall r.m.s. difference of 0.2 Å is used to characterize the observed differences between models and the 0.6 Å outliers are not considered as such. Variations as large as 1 Å are routinely included in the r.m.s. distances calculated by modern algorithms and only 3% of the model (as little as 11 residues per 40 kDa protein) deviating by such an amplitude will approximately double the calculated r.m.s. distance.

Outliers contribute disproportionately to the overall r.m.s. distance because of the quadratic nature of the measure. For example, if an atom shift is ten times that of an average atom, its contribution is equivalent to 100 'average' atoms.

2.3. Percentile-based spread

The variation in interatomic distances can be characterized by a direct least-squares fit of the histogram of distances to (1) to obtain σ_r. Appropriate algorithms are implemented

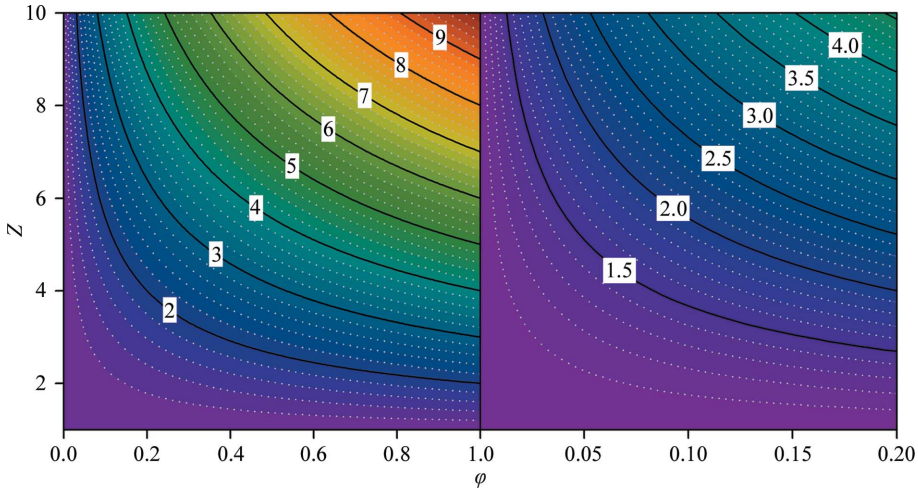


Figure 2
Relative inflation of the r.m.s.d. *versus* the fraction of outliers and the *Z* score. The second panel corresponds to the minor (less than 20%) presence of outliers.

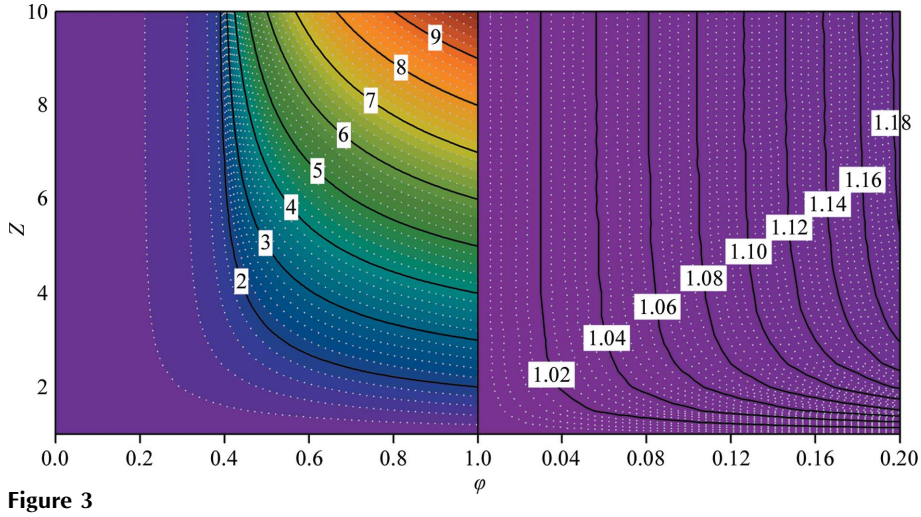


Figure 3
Relative inflation of the p.b.s. *versus* the fraction of outliers and the *Z* score. The second panel corresponds to the minor (less than 20%) presence of outliers.

and the software is available as part of *ShakErr* (http://shakerr.sourceforge.net). To characterize more complicated cases, a multivariate analysis is also available in which the distribution of interatomic distances is approximated by multiple components each conforming to (1).

In addition, an alternative singular measure (referred to in the following as percentile-based spread or p.b.s.) is proposed based on the following consideration. Assuming that the interatomic distance variation follows (1) with single σ_r, the latter will correspond to the ∼60.8% percentile. Hence, the overall data spread can be quickly estimated. This approach is similar to the interquartile analysis used in descripive statistics.

The presence of other components in the distance distribution skews the p.b.s. to higher values. Nevertheless, it remains a very good estimate of the variation of the principal component unless the minor component(s) contribute more than ∼40% to the overall distribution (see Fig. 3). The inflation of the p.b.s. for increasing variation of the secondary component is approximately equal to the fraction of this component in relative terms. The stability in the presence of outliers is an important advantage of the p.b.s. measure compared with the r.m.s. difference.

The breakdown point of the p.b.s. as the σ_r estimator can be further extended by using a lower percentile point (*e.g.* 50%) with an appropriate correction factor corresponding to the percentile of the Maxwell–Boltzmann distribution. It must be noted, however, that instances of outliers comprising over 40% of the sample are better addressed by multimodal distribution analysis.

Interestingly, minimization of the p.b.s. instead of the r.m.s. distance does not produce significant changes in either the transformation required to superpose two structures or the p.b.s. or r.m.s. distance values. This indicates that while outliers inflate the r.m.s. distance value itself, they introduce negligible distortions in the solution of the superposition problem.

2.4. Isomorphous crystal structures

When a protein crystallizes in multiple crystal forms, the structures are expected to be largely isomorphous, with some variation driven by crystal contacts. For example, when tetragonal and triclinic forms of hen egg-white lysozyme [PDB entries 3a8z (Takafumi *et al.*, 2010) and 3lzt (Walsh *et al.*, 1998)] are compared, the structures are essentially identical but produce a relatively large overall r.m.s. distance of 1.34 Å. The distribution of interatomic distances (Fig. 4) clearly shows that the majority of atoms shift by less than 0.8 Å. Fitting the distribution to (1) produces a much smaller estimate of the underlying σ_r of ∼0.39 Å. Comparison of the lysozyme structure in the tetragonal lattice to that in monoclinic (PDB entry 1hf4; Vaney *et al.*, 2001) and orthorhombic (PDB entry 2aki; Artymiuk *et al.*, 1982) space groups produces similar results: the overall r.m.s. distances (0.94/1.10 Å) are much larger than the estimates based on the fit to (1) (0.28/0.26 Å).

The overall r.m.s. distance is reduced when structures refined for the same crystal form are compared. For example, Vaney *et al.* (1996) compared regular tetragonal lysozyme crystals with those grown under microgravity conditions. The overall r.m.s. distance is reduced to 0.15 Å, which is still significantly higher than the σ_r of ∼0.05 Å determined from fitting to (1). Similar results are observed for orthorhombic lysozyme crystals compared with isomorphous crystals grown in a strong magnetic field (Saijo *et al.*, 2005; overall r.m.s.d. of

$\sim$0.75 Å and σ_r of $\sim$0.11 Å). Interestingly, σ_r is close to the Cruickshank DPI of these structures ($\sim$0.07 Å), indicating that the variations in atomic positions are mostly a consequence of the limited precision of the refined models.

To account for the possibility that the crystal may contain subgroups of atoms that exhibit different degrees of positional variation, backbone and side-chain atoms were compared separately for the lysozyme structures described in Vaney *et al.* (1996). Fig. 5 compares the r.m.s. distance for various atom groups with the corresponding σ_r obtained by fitting the distance distribution to (1). For all the protein atoms the r.m.s. distance inflates the estimate of the underlying variation by threefold. The two measures are much closer when only the backbone atoms are compared. However, this does not mean that the side-chain atoms have a much higher positional variability, but rather that the majority of outliers are side-chain atoms.

It is expected that atoms with high B factors will be more likely to result in outliers. For instance, some such atoms are not well defined in the corresponding electron density and may be placed in rather different positions upon independent refinement. Removing 10% of atoms with the highest B factors from the comparison has an effect similar to that of removing side chains. In fact, for backbone atoms with low B factors the r.m.s.d. matches σ_r within less than 20% (0.055/0.047 Å). Thus, for perfectly isomorphous structural models most of the inflation of the r.m.s. difference arises from side chains with relatively high B factors. However, this does not extend to the structures discussed below, which contain actual differences that cannot be accounted for by model-building uncertainties.

2.5. Conformational changes

The percentile-based spread is independent of the amplitude of the atomic shifts caused by conformational changes (assuming that such movements are local and only cover a fraction of the protein molecule). Therefore, it allows the identification of conformational changes by comparing them with the baseline variation in the rest of the structure. Two examples of such analysis are discussed below.

2.5.1. Anti-cocaine antibody M82G2. Structural comparison of the variable domains of this antibody in apo and liganded forms (Pozharski *et al.*, 2005) shows rearrangement of the CDR loops upon ligand binding. The overall r.m.s.d. between the variable domains is 0.69 Å, which is significantly higher than the percentile-based spread (0.20 Å). Multivariate analysis shows the presence of a significant subpopulation ($\sim$20%) of atoms with a much higher underlying variation of $\sim$0.47 Å. Fig. 6(*a*) shows that these larger shifts are primarily associated with the CDR loops that interact with the ligand.

Remarkably, restricting the structural comparison to backbone atoms with B factors of less than 45 Å^2 (in both structures $\langle B\rangle \simeq 30$ Å^2) does reduce the r.m.s. difference but not to the extent seen above with isomorphous structures. Moreover, even in the limited structural comparison multivariate analysis indicates the presence of two major populations of atoms. Approximately an 85% majority of low-B-factor backbone atoms are governed by an underlying variation of 0.15 Å, as expected from the Cruickshank DPI of the two structural models. The remaining 15%, which are mostly located in the CDR loops, show positional shifts of $\sim$0.48 Å corresponding to the induced fit upon ligand binding.

2.5.2. Anti-morphine antibody 9B1. In this case, no dramatic structural rearrangements of the CDR loops were observed upon ligand binding (Pozharski *et al.*, 2004). For the constant domain overall, the r.m.s.d. is approximately twice the value of the p.b.s. (0.35/0.17 Å), which can be attributed to disordered side chains since the two measures converge when only the backbone atoms are considered (0.18/0.15 Å). Interestingly, a similar convergence is observed for the backbone atoms of the variable domain, but both measures show a

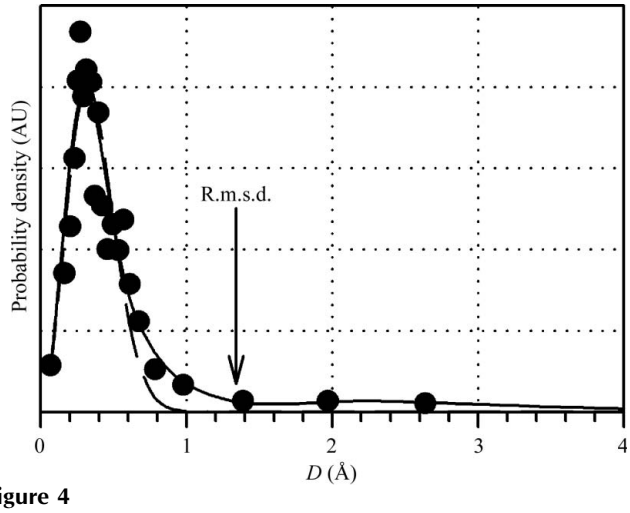


Figure 4
Distribution of atomic shifts for the alignment of tetragonal and triclinic forms of lysozyme. The arrow shows the r.m.s.d. The dashed line corresponds to the least-squares fit to (1).

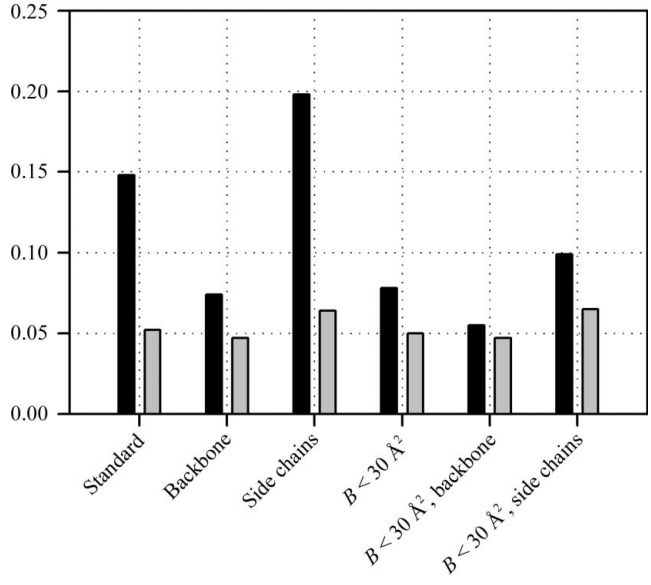


Figure 5
Variations in atomic positions for different groups of atoms comparing regular lysozyme crystals with those grown in microgravity. Black bars, r.m.s.d.; gray bars, σ_r as determined by fitting to (1).

relatively high variation in atomic positions (0.35/0.33 Å). This is significantly higher than what is expected based on the Cruickshank DPI and cannot be explained by a higher degree of disorder since the atoms in both domains have an identical average B factor ($\langle B \rangle \simeq 30$ Å^2).

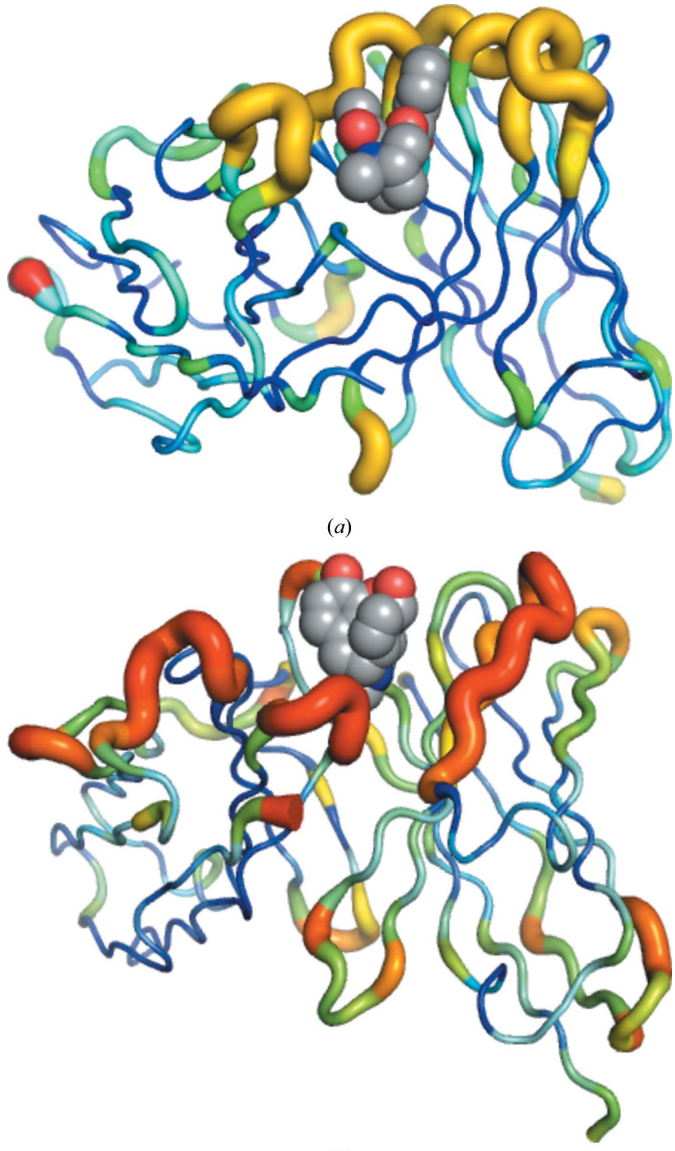
Further analysis reveals that the r.m.s.d. and p.b.s. are dramatically reduced when single immunoglobulin-fold domains, V_L and V_H, are considered. The corresponding values of the r.m.s.d. and p.b.s. are 0.25/0.21 Å and 0.20/0.16 Å, respectively. This supports the conclusion reached in Pozharski *et al.* (2004) regarding the domain-closure motion in 9B1 upon ligand binding. Unlike M82G2, the observed structural variations are not localized in the binding site but rather are spread throughout the structure, as shown in Fig. 6(*b*).

2.6. Model ensembles generated by automated rebuilding

The refinement of macromolecular crystal structures is known to suffer from significant systematic errors that prevent the corresponding measure of refinement quality, the R factor, from reaching the theoretical minimum (Vitkup *et al.*, 2002). It is widely believed that the two major problems are inadequate modeling of bulk solvent and heterogeneity of the protein structure owing to the anharmonic nature of atomic motions that cannot be captured by existing models. Indeed, the structures of small molecules which do not contain solvent and can be adequately described by harmonic dynamics are routinely refined to much lower R factors than their macromolecular counterparts.

Attempts have been made to introduce model ensembles to capture the heterogeneity of protein structures (DePristo *et al.*, 2004; Furnham *et al.*, 2006; Gill *et al.*, 2002; Levin *et al.*, 2007; Pellegrini *et al.*, 1997; Terwilliger *et al.*, 2007; van den Bedem *et al.*, 2009). It has been suggested that iterative rebuilding produces ensembles of models that are sensitive to protein dynamics (DePristo *et al.*, 2004); this claim has subsequently been disputed (Terwilliger *et al.*, 2007). The automated model rebuilding produces model ensembles that are characterized by r.m.s. variations in atomic positions that far exceed the expectations based on measures of overall model precision such as the Cruickshank DPI (but more closely match the maximum-likelihood estimate of coordinate error according to the report; such estimates are expected to be higher and are somewhat meaningless for an unfinished structural model; Murshudov & Dodson, 1997). Based on this observation, it has been widely assumed that the systematic errors in macromolecular models exceed the statistical errors owing to variation in the underlying data. Analysis of the distribution of variations in atomic positions in model ensembles generated in Terwilliger *et al.* (2007) indicates that outliers contribute disproportionately to the r.m.s. measure. Results for structures at different resolutions are shown in Table 1. It is noteworthy that the core variation is lower than the Cruickshank DPI. This suggests that properly rebuilt fragments of the model ensemble vary according to the actual instability of the refinement protocol, which is much less than the DPI measure which reflects the statistical error of the diffraction experiment. The only exception to this is the lowest resolution structure in the set, 1c1z (Schwarzenbacher *et al.*, 1999). The ensembles generated for this low-resolution structural model showed severe geometry problems (for instance, ∼14% of residues are found in disallowed regions of the Ramachandran map).

Furthermore, the majority of the outliers correspond to areas of the model with either poor geometry or poor fit to the electron density (or both). Some examples are shown in Fig. 7 and the quality of the individual ensemble models deteriorates when the resolution is lower. It appears that the variation in the model ensembles reflects the quality of the automated rebuilding, not the precision of the crystal structure. Introduced systematic errors can potentially reduce the quality of the rest of the model, in addition to the detrimental effect of



(*a*)

(*b*)

Figure 6
Spatial location of atoms that belong to the 'larger shift' population (corresponding to thicker tubes) in the variable domain of antibodies M82G2 (*a*) and 9B1 (*b*).

Table 1
Model ensembles obtained by automated rebuilding in Terwilliger *et al.* (2007).

R.m.s.d. and σ_r are shown for all atoms, for backbone (bb) and for side chains (sc). DPI was calculated using the statistics reported in the PDB and the following expression: DPI = $(N_{\rm atoms}/N_{\rm reflections})^{1/2}[(d_{\rm min}R_{\rm free})/({\rm completeness})^{1/3}]$.

			Ensemble r.m.s.d. (Å)			σ_r (Å)			
PDB code	$d_{\rm min}$ (Å)	DPI (Å)	All	bb	sc	All	bb	sc	Reference
1a0j	1.70	0.131	0.357	0.054	0.560	0.014	0.013	0.018	Schrøder *et al.* (1998)
1a3n	1.80	0.128	0.319	0.025	0.476	0.017	0.015	0.021	Tame & Vallone (2000)
1bmb	1.80	0.116	0.330	0.052	0.487	0.012	0.010	0.014	Ettmayer *et al.* (1999)
1aof	2.00	0.138	0.311	0.053	0.472	0.018	0.016	0.022	Williams *et al.* (1997)
1c2t	2.10	0.261	0.462	0.135	0.654	0.045	0.041	0.054	Greasley *et al.* (1999)
1uyi	2.20	0.155	0.369	0.080	0.540	0.023	0.021	0.029	Wright *et al.* (2004)
1rg5	2.50	0.140	0.366	0.150	0.549	0.027	0.024	0.031	Roszak *et al.* (2004)
1p4t	2.55	0.230	0.545	0.286	0.724	0.048	0.043	0.060	Vandeputte-Rutten *et al.* (2003)
1cqp	2.60	0.295	0.449	0.125	0.650	0.042	0.037	0.049	Kallen *et al.* (1999)
1c1z	2.88	0.207	1.203	0.634	1.622	0.209	0.196	0.240	Schwarzenbacher *et al.* (1999)

removing water molecules and ligands [some of the structural models discussed in Terwilliger *et al.* (2007) had large ligands such as heme which were removed prior to analysis].

The model ensembles were inspected to correct various errors introduced by automated rebuilding. To assist the model-ensemble repair, a cluster analysis-based algorithm was used. In this approach, groups of atoms (residues, backbone and side-chain atoms) are treated as points in $3N$-dimensional space. Instances from ensemble models are clustered using k-means clustering. The main (largest) cluster is identified and the distances between clusters are normalized by the r.m.s. of their radii. If this normalized distance between clusters is below a predefined cutoff value (*i.e.* if there is a substantial gap between clusters), the models that belong to the minor cluster are shifted to a random position within the main cluster. The r.m.s. diameter of the main cluster must be below a certain cutoff level (0.5 Å in this work) to assure that disordered residues are ignored. This algorithm was very efficient in correcting certain types of errors in model ensembles, such as symmetrical side-chain flips and lone incorrect models. The resulting models were subjected to refinement in *REFMAC* (Murshudov *et al.*, 1997) and visually inspected in the context of the corresponding electron density. Several common types of errors were identified and are listed below.

(i) Singular incorrect models. This is the most obvious type of error and its contribution to the overall r.m.s.d. reflects the inaccuracy of automated rebuilding. Some of these errors may be the consequence of an incorrect choice of initial rotamer, while others may be influenced by elements that are excluded from the structure, such as cofactors, water molecules and unconventional amino acids.

(ii) Flipping of quasi-symmetrical residues. Asp, Glu, Asn, Gln and His can be placed into the same fragment of electron density in two ways. Sometimes the correct orientation may be deduced from B factors and/or hydrogen-bonding patterns, while for Asp and Glu the two orientations are equivalent and simply refer to the naming of the oxygen atoms. Alternating the orientations of side chains in the model ensemble artificially inflates the r.m.s.d. while not providing an actual alternative model.

(iii) Alternate conformers. For residues that in fact adopt alternate conformations, the model ensemble obtained by automated rebuilding will include both rotamers. While this is an important feature of a crystal structure, its contribution to the inflation of the r.m.s.d. is misleading. The correct way to describe the situation is to introduce alternate conformations; indeed, the atomic positions are known with much better accuracy than the distances between alternate conformers. In our analysis, a single conformer was introduced whenever possible. This left the model incomplete, but allowed us to dissect the contribution of this type of error to r.m.s.d. inflation.

(iv) Disordered residues. When no electron density is found to support a particular conformation, automated rebuilding produces a diverse set of rotamers. These may inflate the r.m.s.d. by as much as an order of magnitude (see below), resulting in a misleading estimate of the uncertainty in the coordinates of individual models. As described below, disordered atoms were removed at the last step of the model-ensemble analysis.

The results of the repair of model ensembles from Terwilliger *et al.* (2007) are shown in Table 2. After correcting the modelling errors described above, except for the disordered residues, an average of a 3.6-fold reduction in the r.m.s.d. was observed. 1bmb (Ettmayer *et al.*, 1999) exhibited the smallest reduction (∼1.5-fold). This is the structure of a small protein refined against high-quality data, resulting in very few misplaced residues. In contrast, the r.m.s.d. for the 1c1z model ensemble was reduced almost eightfold. This is the lowest resolution (2.87 Å) structure included in the analysis and it contained multiple regions with poorly defined density where automated rebuilding produced models with impossible geometry.

Removing disordered atoms from the models resulted in a dramatic decrease of the r.m.s.d. (average of ∼ninefold). Interestingly, the most significant decrease (∼30-fold) was seen for 1bmb, while the decrease for 1c1z was relatively modest (∼fourfold). Some correlation is observed between the fraction of the model that is disordered and the reduction in r.m.s.d. upon the removal of disordered residues, but the latter is also influenced by various random factors such as the

degree of diversity in the rotamers produced by automated rebuilding.

Interestingly, the p.b.s. also decreased upon model repair/correction. The effect was smaller compared with that exerted on the r.m.s.d. (∼eightfold *versus* ∼24-fold average decrease). Nevertheless, it indicates that model errors affect the rest of the structure, resulting in a global decline in model quality.

The average final r.m.s.d. of the model ensembles is ∼0.02 Å, which is significantly less than the initial variation of ∼0.5 Å. The gap between the r.m.s.d. and the p.b.s. is reduced when model errors are corrected (the average final p.b.s. for all the models analyzed here was ∼0.01 Å). It is important to emphasize that these values include all atoms in the model. This suggests that the instability of positional refinement contributes significantly less (∼0.01 Å) to the precision of crystallographic models and the inflation of variation in model ensembles produced by automated rebuilding arises from the introduction of systematic errors. It is likely that statistical errors (*e.g.* the uncertainty in measured intensities) constitute the major component of the uncertainty in crystal structure models (the average DPI of the analyzed models is ∼0.2 Å).

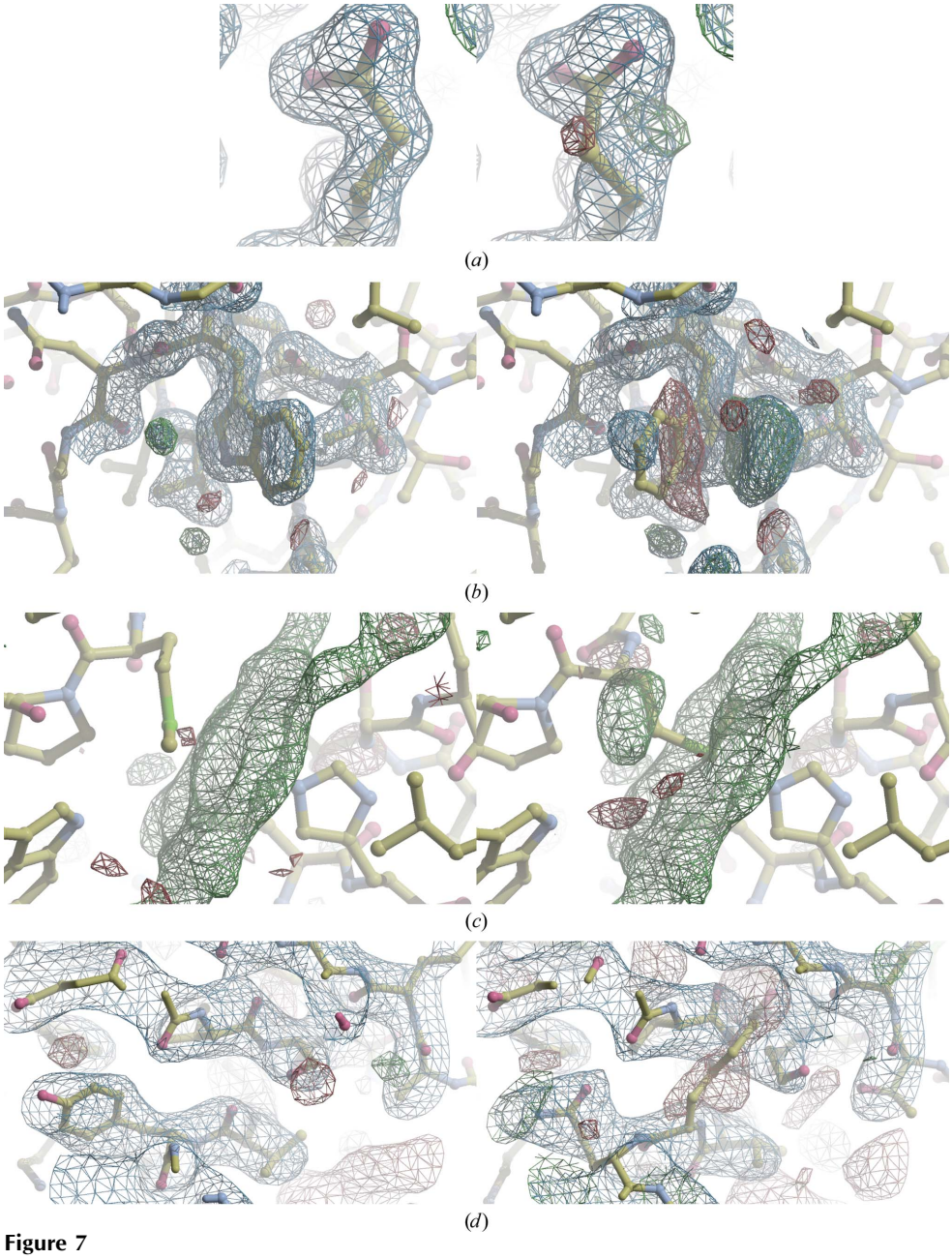


Figure 7
Examples of systematic model errors introduced by automated rebuilding. Each panel shows the correctly placed model on the left compared with incorrect placement on the right. Electron density was calculated using *PHENIX* (Adams *et al.*, 2010) and rendered with *Coot* (Emsley *et al.*, 2010). (*a*) Glu77 from PDB entry 1a0j (incorrect side-chain rotamer). (*b*) Trp522 from PDB entry 1aof (incorrect side-chain rotamer which may be a consequence of the nearby water molecule). (*c*) Met106 from PDB entry 1aof (side chain misplaced to account for the electron density of a missing heme ligand; only the difference density map is shown for clarity). (*d*) Tyr256 from PDB entry 1c1z (unacceptable geometry at low resolution).

2.7. Two sources of sliding variance: model precision and actual changes

The above considerations are independent of the actual source of the deviation of the distance distribution from the theoretical prediction given by (1). Our major concern so far was the correct estimation of the variation corresponding to the main (principal) component. Several possible sources of the outliers are discussed below.

Naturally, actual structural differences that exceed the precision of the structural model will produce deviations from the theoretical prediction. To this end, outlier analysis may help to identify conformational changes in protein structures. However, this should be performed with caution since different parts of the model may be determined with varying precision. Such variations throughout the model are likely to correlate with the atomic displacement parameters (Cruickshank, 1999). In other words, some atoms may show higher positional variation simply because they are not as well defined by electron density. This factor on its own may produce multivariate distance distributions.

Another possible source of outliers are modeling errors. An example of this discussed above is

Table 2
Results of error correction in model ensembles obtained by automated rebuilding in Terwilliger *et al.* (2007).

The total number of residues in the initial model (N) and the total number of residues removed in the last step of model correction (ΔN) are shown. (In a single instance, an iron was added to the model for 1rg5 to prevent the distortion of metal-coordinating residues.) The all-atom r.m.s.d. and p.b.s. are shown at three stages of model correction: prior to correction (initial), after all errors were corrected but prior to removal of disordered residues (repaired) and after disordered residues were removed and models subjected to one last cycle of refinement (final).

			Ensemble r.m.s.d. (Å)			p.b.s. (Å)		
PDB code	N	ΔN	Initial	Repaired	Final	Initial	Repaired	Final
1a0j	892	32	0.357	0.135	0.009	0.021	0.007	0.003
1a3n	572	12	0.319	0.152	0.018	0.022	0.013	0.007
1bmb	106	6	0.330	0.221	0.008	0.017	0.009	0.002
1aof	1074	79	0.311	0.147	0.022	0.025	0.013	0.007
1c2t	418	4	0.462	0.093	0.044	0.066	0.023	0.020
1uyi	208	3	0.369	0.140	0.017	0.032	0.008	0.004
1rg5	824	15	0.366	0.135	0.034	0.038	0.011	0.006
1p4t	155	5	0.545	0.101	0.017	0.083	0.011	0.005
1cqp	364	4	0.449	0.100	0.024	0.061	0.017	0.010
1c1z	326	17	1.203	0.157	0.041	0.406	0.042	0.023

iterative automated model rebuilding, which tends to introduce model errors. These significantly inflate the r.m.s. variation and, if not numerous, can be easily identified and corrected.

3. Conclusions

How precise are macromolecular crystal structures? Several measures have been devised over the years to determine the overall uncertainty of model coordinates, with the Cruickshank DPI and its maximum-likelihood-based variant being the currently accepted measures of overall model precision. It is also recognized that when structures that are expected to be similar or even identical are compared, the r.m.s. distance between models often exceeds what it should be assuming that statistical error is the main contribution to the structural uncertainty. This fact is routinely ignored or is explained by crystal-to-crystal variations. More recently, it has been proposed that systematic errors in crystal structures far exceed the statistical errors arising from variation in the underlying diffraction data. The instability of the refinement process leading to model degeneracy and the heterogeneity of possible models owing to anharmonic motion have been cited as possible sources of such systematic error.

Here, it is proposed that the problem lies in the r.m.s. distance measure itself. The Cruickshank DPI and other measures of coordinate uncertainty refer to the positional error of an 'average atom' or, more specifically, a hypothetical atom with a B factor matching the average value in the structure. However, the r.m.s. distance will only match such an error in the absence of outliers, when all the atoms obey the same underlying distribution. It is well understood that this is not the case in protein crystal structures since the positions of atoms found in the areas of weaker density are less well defined. Outliers will also be present when actual conformational changes do take place.

Two alternative approaches to the evaluation of the structural differences are proposed. Firstly, the percentile-based spread (p.b.s.) is introduced and is shown to be less sensitive to the presence of outliers than the traditional r.m.s. difference. In the situation where a small percentage of interatomic distances are characterized by much higher variation, the p.b.s. is particularly accurate. If two or more groups of atoms are present with positional variations that are comparable, the inflation of this measure is proportional to the fraction of corresponding atom groups in the population.

More detailed analysis is possible when interatomic distances show the presence of several groups with different underlying variations. Direct nonlinear regression allows the determination of the corresponding variances and the relative number of atoms in each group. It appears that at least in some cases these subgroups can be interpreted as more *versus* less flexible atomic groups (*e.g.* backbone *versus* side chains, surface *versus* protein core *etc.*). Because of the significant overlap between these groups, direct assignment of individual atoms is not always trivial.

The main conclusion proposed here is that using the r.m.s. distance to characterize differences between crystal structures will produce misleading results in most cases and that the percentile-based spread should be used instead.

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.
Artymiuk, P. J., Blake, C. C. F., Rice, D. W. & Wilson, K. S. (1982). *Acta Cryst.* B**38**, 778–783.
Bedem, H. van den, Dhanik, A., Latombe, J.-C. & Deacon, A. M. (2009). *Acta Cryst.* D**65**, 1107–1117.
Blow, D. M. (2002). *Acta Cryst.* D**58**, 792–797.
Chambers, J. L. & Stroud, R. M. (1979). *Acta Cryst.* B**35**, 1861–1874.
Cruickshank, D. W. J. (1999). *Acta Cryst.* D**55**, 583–601.
DePristo, M. A., de Bakker, P. I. W. & Blundell, T. L. (2004). *Structure*, **12**, 831–838.
Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.
Ettmayer, P., France, D., Gounarides, J., Jarosinski, M., Martin, M.-S., Rondeau, J.-M., Sabio, M., Topiol, S., Weidmann, B., Zurini, M. & Bair, K. W. (1999). *J. Med. Chem.* **42**, 971–980.
Furnham, N., Blundell, T. L., DePristo, M. A. & Terwilliger, T. C. (2006). *Nature Struct. Mol. Biol.* **13**, 184–185.

Gill, H. S., Pfluegl, G. M. U. & Eisenberg, D. (2002). *Biochemistry*, **41**, 9863–9872.
Greasley, S. E., Yamashita, M. M., Cai, H., Benkovic, S. J., Boger, D. L. & Wilson, I. A. (1999). *Biochemistry*, **38**, 16783–16793.
Kabsch, W. (1976). *Acta Cryst.* A**32**, 922–923.
Kallen, J., Welzenbach, K., Ramage, P., Geyl, D., Kriwacki, R., Legge, G., Cottens, S., Weitz-Schmidt, G. & Hommel, U. (1999). *J. Mol. Biol.* **292**, 1–9.
Levin, E. J., Kondrashov, D. A., Wesenberg, G. E. & Phillips, G. N. (2007). *Structure*, **15**, 1040–1052.
Murshudov, G. N. & Dodson, E. J. (1997). *CCP4 Newsl.* **33**, 31–39.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Pellegrini, M., Grønbech-Jensen, N., Kelly, J. A., Pfluegl, G. M. U. & Yeates, T. O. (1997). *Proteins*, **29**, 426–432.
Pozharski, E., Moulin, A., Hewagama, A., Shanafelt, A. B., Petsko, G. A. & Ringe, D. (2005). *J. Mol. Biol.* **349**, 570–582.
Pozharski, E., Wilson, M. A., Hewagama, A., Shanafelt, A. B., Petsko, G. & Ringe, D. (2004). *J. Mol. Biol.* **337**, 691–697.
Rashin, A. A., Rashin, A. H. L. & Jernigan, R. L. (2009). *Acta Cryst.* D**65**, 1140–1161.
Roszak, A. W., McKendrick, K., Gardiner, A. T., Mitchell, I. A., Isaacs, N. W., Cogdell, R. J., Hashimoto, H. & Frank, H. A. (2004). *Structure*, **12**, 765–773.
Saijo, S., Yamada, Y., Sato, T., Tanaka, N., Matsui, T., Sazaki, G., Nakajima, K. & Matsuura, Y. (2005). *Acta Cryst.* D**61**, 207–217.
Schrøder, H. K., Willassen, N. P. & Smalås, A. O. (1998). *Acta Cryst.* D**54**, 780–798.
Schwarzenbacher, R., Zeth, K., Diederichs, K., Gries, A., Kostner, G. M., Laggner, P. & Prassl, R. (1999). *EMBO J.* **18**, 6228–6239.
Takafumi, U., Satoshi, A., Tomomi, K., Takahiro, O., Tatsuo, H. & Yoshihito, W. (2010). *Chem. Eur. J.* **16**, 2730–2740.
Tame, J. R. H. & Vallone, B. (2000). *Acta Cryst.* D**56**, 805–811.
Terwilliger, T. C., Grosse-Kunstleve, R. W., Afonine, P. V., Adams, P. D., Moriarty, N. W., Zwart, P., Read, R. J., Turk, D. & Hung, L.-W. (2007). *Acta Cryst.* D**63**, 597–610.
Vandeputte-Rutten, L., Bos, M. P., Tommassen, J. & Gros, P. (2003). *J. Biol. Chem.* **278**, 24825–24830.
Vaney, M. C., Broutin, I., Retailleau, P., Douangamath, A., Lafont, S., Hamiaux, C., Prangé, T., Ducruix, A. & Riès-Kautt, M. (2001). *Acta Cryst.* D**57**, 929–940.
Vaney, M. C., Maignan, S., Riès-Kautt, M. & Ducruix, A. (1996). *Acta Cryst.* D**52**, 505–517.
Vitkup, D., Ringe, D., Karplus, M. & Petsko, G. A. (2002). *Proteins*, **46**, 345–354.
Walsh, M. A., Schneider, T. R., Sieker, L. C., Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1998). *Acta Cryst.* D**54**, 522–546.
Williams, P. A., Fülöp, V., Garman, E. F., Saunders, N. F. W., Ferguson, S. J. & Hajdu, J. (1997). *Nature (London)*, **389**, 406–412.
Wright, L. *et al.* (2004). *Chem. Biol.* **11**, 775–785.