Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Quality of protein crystal structures

Eric N. Brown and S. Ramaswamy*

University of Iowa, Department of Biochemistry, Iowa City, IA, USA

Correspondence e-mail: s-ramaswamy@uiowa.edu

Received 11 June 2007
Accepted 10 July 2007

The genomics era has seen the propagation of numerous databases containing easily accessible data that are routinely used by investigators to interpret results and generate new ideas. Most investigators consider data extracted from scientific databases to be error-free. However, data generated by all experimental techniques contain errors and some, including the coordinates in the Protein Data Bank (PDB), also integrate the subjective interpretations of experimentalists. This paper explores the determinants of protein structure quality metrics used routinely by protein crystallographers. These metrics are available for most structures in the database, including the *R* factor, $R_{\rm free}$, real-space correlation coefficient, Ramachandran violations *etc.* All structures in the PDB were analyzed for their overall quality based on nine different quality metrics. Multivariate statistical analysis revealed that while technological improvements have increased the number of structures determined, the overall quality of structures has remained constant. The quality of structures deposited by structural genomics initiatives are generally better than the quality of structures from individual investigator laboratories. The most striking result is the association between structure quality and the journal in which the structure was first published. The worst offenders are the apparently high-impact general science journals. The rush to publish high-impact work in the competitive atmosphere may have led to the proliferation of poor-quality structures.

1. Introduction

In 1990, Carl Brändén and Alwyn Jones published an article on subjectivity in the crystallographic models deposited in the Protein Data Bank (Brändén & Jones, 1990). 17 y and thousands of structures later, we analyze here the quality of structures deposited in the PDB. Much has changed in the way protein crystallography is practiced. Routine production, purification and crystallization of proteins has led to an explosion of protein structures available in the Protein Data Bank. These have found many uses in structure-based drug design, molecular modeling and general biochemistry.

Since its inception in 1971 at Brookhaven National Laboratory, the database has progressively grown from seven structures to over 40 000 in 2007 (Bernstein *et al.*, 1977; Berman *et al.*, 2000, 2003). The vast majority of the data collected for structure determination were obtained using synchrotron radiation (Jiang & Sweet, 2004).

Through the use of high-throughput cloning, expression and purification methods, more and more proteins are amenable to

doi:10.1107/S0907444907033847

structure determination *via* crystallography (Abola *et al.*, 2000). At the extreme of the high-throughput spectrum is the automation available at some pharmaceutical companies. Robots handle cloning, expression, purification, crystallization and data collection for dozens to hundreds of protein targets, all with minimal user intervention. These data can then be processed in a semi-automated fashion using popular integration packages such as *d*TREK* from Rigaku, *HKL*-2000 from HKL Research (Otwinowski & Minor, 1997) or even the *MOSFLM*/*SCALA* packages from CCP4 (Collaborative Computational Project, Number 4, 1994; Leslie, 1992). Finally, automated software can perform model building and structure refinement. A bare minimum of human intervention during determination is the selection of targets, the preparation of initial DNA and looping of crystals produced. All other steps can be automated.

In addition to proprietary structural studies performed by pharmaceutical companies, various structural genomics projects have been initiated (Peat *et al.*, 2002; Geerlof *et al.*, 2006; Rupp *et al.*, 2002; Liu *et al.*, 2005). 18 structural genomics centers currently exist and have deposited over 2000 structures in the Protein Data Bank. The most productive of these, the Midwest Center for Structural Genomics (MCSG), the Joint Center for Structural Genomics (JCSG), the RIKEN Structural Genomics/Proteomics Initiative (RSGI), the Structural Genomics Consortium (SGC) and the New York Structural Genomics Research Consortium (NYSGRC), have each solved over 200 structures.

An immediate downside to increased automation is inferior quality structures being deposited, distributed and used by other disciplines, since human intuition and reasoning are taken out of the process (Brändén & Jones, 1990). It is becoming increasingly easy to incorrectly use a protein structure. Simulations ranging from homology modeling to ligand and protein docking to kinetics simulators often take the structural model as 'gospel', ignoring any interpretation that went into refining the structure. Atomic displacement parameters are most likely ignored. Large-order displacements such as TLS or multiple conformations may be overlooked. Finally, disordered termini and loops can be forgotten.

Another unfortunate side-effect of the proliferation of crystallographically determined structures is the increasingly limited peer review that they elicit. New structures are becoming a minor character in the discourse of a paper and thus of the few reviewers recruited to read the paper, fewer may be qualified or be able to evaluate the quality of the structure. A further hindrance is that the structure factors and coordinates are often not part of the reviewing process, making critical review difficult, if not impossible. Although validation is becoming increasingly important during structure deposition in the PDB, policy dictates that the PDB cannot refuse a structure upon the author's insistence.

This paper explores the determinants of protein structure quality metrics. Using a combination of common and easy-to-compute descriptors of the protein being solved, an *a priori* estimate of various structure-quality metrics can be made. Any significant deviation of the observed metrics from these expected values is thus an additional sign that careful evaluation of the structure is necessary prior to its use.

2. Protein structure quality metrics

A multitude of qualitative and quantitative metrics have been devised to evaluate crystallographic models during or following refinement (Brändén & Jones, 1990). The metrics used in this study include the *R* factor, $R_{\rm free}$, real-space *R* factor, real-space electron-density correlation coefficient, average occupancy-weighted *B* value and number of Ramachandran violations. Some of these values can be pulled from the header information found in the PDB entry. Others are available *via* computational servers on the Internet.

The most common quality metric is the *R* factor. Computed as the relative deviation of calculated structure factors from those observed, its value is tied to the quality of not only the model but also the data. It is commonly understood that an acceptable *R* factor depends on the completeness of the model and the resolution limits of the data: a more complete model or one with higher resolution data should have a lower *R* factor.

Statistics provides a few methods for preventing the overinterpretation of the data by the model (overfitting). By randomly partitioning the data into two sets, a working and a testing data set, refinement changes that decrease the working error but increase the testing error can point to overfitting. The $R_{\rm free}$ measure thus reports the deviations of the calculated model as it applies to this smaller testing data set (Brünger, 1993). Unfortunately, the $R_{\rm free}$ reflections are often used in map construction during model building, thereby decreasing the effectiveness of $R_{\rm free}$ as a measure of true unbiased structural quality. Additionally, noncrystallographic symmetry increases the correlation between reflections in the working and testing sets (Fabiola *et al.*, 2006).

The Uppsala University Electron Density Server provides the rest of the metrics used in this study (Kleywegt *et al.*, 2004). The first of these is the real-space *R* value (Brändén & Jones, 1990). The real-space *R* value is the percentage deviation of the calculated $\sigma_{\rm A}$-weighted $2F_{\rm obs} - F_{\rm calc}$ map from the $F_{\rm calc}$ map in the vicinity of nonwater residues (Brändén & Jones, 1990; Srinivasan, 1966). These residue *R* values are then averaged to give a whole-structure real-space *R* value. Using a similar computation, the real-space correlation coefficient can be calculated for each nonwater residue and an average for the whole structure. Unlike the traditional *R* factors, these real-space versions are somewhat resistant to overfitting noise in the electron-density map.

Each individual structure will have a distribution of residue-based real-space *R* values and correlation coefficients. The average is used as a whole-structure value. The spread of the distribution can be used to identify potentially incorrect residues. If an individual residue's *R* value (or correlation coefficient) is more than 3σ from the mean, it is marked as an outlier. The percent of outliers of *R* values and correlation coefficients are a further measure of the reasonableness of a

Table 1
Coefficients for fitted quality metric models for data mined from PDB headers (equations 1 and 2).

Metrics shown are those determined to be significant during model construction.

Quality metric	Variable	Value	Standard error	p value
R factor	Intercept	1.44×10^{-1}	2.06×10^{-3}	<0.001
	Resolution	2.60×10^{-2}	9.55×10^{-4}	<0.001
	Source	-2.95×10^{-2}	2.22×10^{-3}	<0.001
	Heavy atoms	-1.10×10^{-6}	1.35×10^{-7}	<0.001
	Similarity index	-6.56×10^{-5}	6.74×10^{-6}	<0.001
	Resolution × source	1.74×10^{-2}	1.04×10^{-3}	<0.001
	Heavy atoms × similarity index	9.01×10^{-9}	1.69×10^{-9}	<0.001
	Resolution × heavy atoms	2.34×10^{-7}	5.16×10^{-8}	<0.001
$R_{\rm free}$	Intercept	2.19×10^{-1}	1.36×10^{-2}	<0.001
	Resolution	2.55×10^{-2}	6.30×10^{-3}	<0.001
	Heavy atoms	-6.79×10^{-7}	3.12×10^{-8}	<0.001
	Date	-6.11×10^{-6}	1.10×10^{-6}	<0.001
	Source	-2.09×10^{-2}	2.29×10^{-3}	<0.001
	Resolution × source	1.18×10^{-2}	1.07×10^{-3}	<0.001
	Resolution × date	1.73×10^{-6}	5.09×10^{-7}	<0.001

structure. The 3σ cutoff is mostly arbitrary and was chosen from the Uppsala Density Server's output.

The Uppsala server reports two additional quality metrics for a large number of proteins: the number of Ramachandran plot violations and the average occupancy-weighted B value. The first of these is computed using the *MOLEMAN*2 program (Kleywegt & Jones, 1996). The acceptable areas of the Ramachandran plot were calculated as the area containing 98% of a large collection of φ and ψ angles. Since this quality indicator is commonly used as a check during refinement, its usefulness may be diminished; however, structures with many Ramachandran violations are clearly suspect. The final quality metric is the occupancy-weighted B value. Thus, the importance of residues with incomplete occupancy or multiple conformations will be correctly diminished.

A number of additional quality metrics have been devised to evaluate protein structures. Some of these aim to generalize the R factor to more than a scalar quantity (Parkin, 2000) or are similar to Uppsala's real-space R factor (van der Akker & Hol, 1999). Finally, some are specific checks which identify known deficiencies or difficulties in protein structures (Vaguine *et al.*, 1999). The above nine quality metrics were chosen owing to their wide application and their scalar nature. Since all macromolecular crystallographic refinement programs include stereochemical restraints as part of their optimization routines, we have not included any additional quality metrics along this line. For a good discussion of these types of restraints, see the recent paper by Jaskolski *et al.* (2007).

3. Methods

To determine the expected quality metrics for a given structure, the statistical package R was used to evaluate linear models of the quality metrics as a function of general information about the structure (R Development Core Team, 2006). The response value of each tested statistical model was one of the quality metrics. These are the R factor, $R_{\rm free}$, real-space R value, number of 3σ outliers from the real-space R value, real-space correlation coefficient, number of 3σ outliers from the real-space correlation coefficient, occupancy-weighted B value, number of 3σ outliers from the occupancy-weighted B value and percentage of Ramachandran violations identified. The distribution of correlation coefficients are far from Gaussian, so correlation coefficients were first transformed as $R_{\rm CC} = \frac{1}{2}[\ln(2 - {\rm CC}) - \ln({\rm CC})]$. This transformation had the additional benefit of changing the ordering of values so that smaller values indicate better quality (higher correlation coefficient).

The linear predictor of each quality metric was a function of metadata mined from the PDB and Uppsala websites on 30 March 2007. All of these metadata values are known or can be estimated prior to structure refinement. The explanatory variables were the date of deposition (days from 1 January 1970), the highest resolution of refinement, the X-ray source used for data collection (binary synchrotron *versus* rotating-anode value), the number of non-H atoms and an index of the novelty of the structure. The first three variables were extracted from PDB headers. The number of non-H atoms was in the Uppsala EDS summary. Finally, to differentiate structures which are newly described *versus* structures which had previously been deposited, the automated clustering performed by the PDB was used. The list of 11 373 clusters with 50% sequence identity was used to assign each of 28 321 X-ray crystallographic structures to a cluster. The structures belonging to a given cluster were then sorted by date of deposition and the position of a structure in this list was saved as an integer similarity index. Therefore, a structure whose sequence was novel when it was deposited would be assigned a similarity index of 1, while the tenth structure of a given protein would be assigned an index of at least 10.

Over 16 000 protein structures had data available for all quality metrics and the five explanatory variables. Using these data, a linear model was first generated explaining each quality metric in terms of the resolution alone: Q = intercept + β_1 × resolution. The effect of adding or removing a single criterion (resolution, date *etc.*) on the likelihood of the current model is used to iteratively add or remove terms from the model (Venables & Ripley, 2002). Using the Bayesian Information Criterion (BIC) to add and remove terms attempts to maximize the fit while minimizing the chance of producing an overly complex model. The generated model maximizes the likelihood of the fit while penalizing overly complex models. Following the identification of significant explanatory variables, the model for each quality metric was refitted with a robust least-squares method (Huber, 1981) utilizing an M-estimator and Huber's φ function (Huber, 1981) to minimize the effect of outliers. The resultant linear model estimates the *a priori* quality metrics as a function of relevant

Table 2
Coefficients for fitted quality metric models for metrics derived from the Uppsala Electron Density Server (equations 3–9).

Metrics shown are those determined to be significant during model construction.

Quality metric	Variable	Value	Standard error	*p* value
Real-space *R* value	Intercept	1.62×10^{-1}	1.99×10^{-2}	<0.001
	Resolution	-4.04×10^{-2}	9.37×10^{-3}	<0.001
	Source	-1.44×10^{-1}	2.27×10^{-2}	<0.001
	Date	-1.36×10^{-5}	1.64×10^{-6}	<0.001
	Similarity index	-3.93×10^{-4}	6.13×10^{-5}	<0.001
	Heavy atoms	-2.42×10^{-6}	1.41×10^{-7}	<0.001
	Resolution × source	8.73×10^{-2}	1.08×10^{-2}	<0.001
	Resolution × date	8.66×10^{-6}	7.71×10^{-7}	<0.001
	Date × similarity index	3.18×10^{-8}	4.43×10^{-9}	<0.001
	Resolution × heavy atoms	8.36×10^{-7}	5.67×10^{-8}	<0.001
	Source × date	1.15×10^{-5}	1.85×10^{-6}	<0.001
	Resolution × similarity index	-4.90×10^{-5}	1.11×10^{-5}	<0.001
	Similarity index × heavy atoms	3.30×10^{-8}	7.61×10^{-9}	<0.001
	Resolution × source × date	-6.58×10^{-6}	8.83×10^{-7}	<0.001
	Resolution × similarity index × heavy atoms	-9.35×10^{-9}	2.66×10^{-9}	<0.001
Transformed real-space CC	Intercept	3.75×10^{-2}	3.72×10^{-3}	<0.001
	Resolution	3.85×10^{-3}	7.17×10^{-4}	<0.001
	Source	4.26×10^{-3}	4.40×10^{-3}	0.33
	Heavy atoms	-8.26×10^{-7}	1.10×10^{-7}	<0.001
	Similarity index	-4.73×10^{-4}	5.73×10^{-5}	<0.001
	Date	1.68×10^{-6}	2.67×10^{-7}	<0.001
	Resolution × source	8.88×10^{-3}	7.65×10^{-4}	<0.001
	Resolution × heavy atoms	2.26×10^{-7}	4.22×10^{-8}	<0.001
	Similarity index × date	2.75×10^{-8}	4.38×10^{-9}	<0.001
	Resolution × similarity index	4.43×10^{-5}	8.20×10^{-6}	<0.001
	Source × date	-1.49×10^{-6}	3.25×10^{-7}	<0.001
Real-space *R*-value outliers	Intercept	2.61	0.02	<0.001
	Resolution	−0.30	0.01	<0.001
Real-space CC outliers	Intercept	2.32	2.26×10^{-2}	<0.001
	Resolution	-1.02×10^{-1}	9.92×10^{-3}	<0.001
	Heavy atoms	-1.54×10^{-5}	2.22×10^{-6}	<0.001
	Similarity index	-6.06×10^{-4}	1.07×10^{-4}	<0.001
	Resolution × heavy atoms	4.94×10^{-6}	8.01×10^{-7}	<0.001
Occupancy-weighted *B* value	Intercept	36.2	4.32	<0.001
	Resolution	−10.2	2.17	<0.001
	Source	−10.8	8.46×10^{-1}	<0.001
	Date	-3.36×10^{-3}	3.55×10^{-4}	<0.001
	Heavy atoms	-6.27×10^{-4}	6.63×10^{-5}	<0.001
	Similarity index	6.81×10^{-2}	1.09×10^{-2}	<0.001
	Resolution × source	7.37	4.25×10^{-1}	<0.001
	Resolution × date	2.14×10^{-3}	1.79×10^{-4}	<0.001
	Resolution × heavy atoms	2.00×10^{-4}	2.75×10^{-5}	<0.001
	Heavy atoms × similarity index	-8.55×10^{-6}	3.75×10^{-6}	0.023
	Resolution × similarity index	-4.62×10^{-2}	5.84×10^{-3}	<0.001
	Resolution × heavy atoms × similarity index	6.00×10^{-6}	1.51×10^{-6}	<0.001
B-value outliers	Intercept	2.75	2.86×10^{-2}	<0.001
	Resolution	-6.74×10^{-1}	1.25×10^{-2}	<0.001
	Source	-1.63×10^{-1}	1.43×10^{-2}	<0.001
	Heavy atoms	8.39×10^{-6}	9.23×10^{-7}	<0.001
Ramachandran outliers	Intercept	-6.41×10^{-1}	2.44×10^{-1}	0.009
	Resolution	2.34	4.01×10^{-2}	<0.001
	Source	3.24×10^{-1}	2.55×10^{-2}	<0.001
	Date	-1.78×10^{-4}	1.87×10^{-5}	<0.001
	Heavy atoms	-2.65×10^{-4}	3.82×10^{-5}	<0.001
	Resolution × heavy atoms	7.66×10^{-5}	5.34×10^{-6}	<0.001
	Date × heavy atoms	7.50×10^{-9}	2.90×10^{-9}	0.01

structure information that can be determined prior to starting refinement.

The expected quality metrics generated using the above models were computed for all structures in the PDB that had the requisite explanatory variables available. The difference of the true metric of the structure from the expected metric was calculated and converted into a PDB-wide Z value, which was normally distributed, centered at 0 and with a standard deviation of 1. Thus, the quality of any and all structures can be compared regardless of whether they were collected at the same resolution, X-ray source *etc.* A positive Z value indicates a quality metric that is worse than expected.

A principal component analysis was used to combine the nine Z values into a single scalar quality value. A three-dimensional subspace of the nine-dimensional Z-score metric was identified that described 50% of the variation in the metrics. Each coordinate in this subspace was renormalized. Finally, the Euclidean distance in this three-dimensional space from the origin was assigned as the final scalar quality metric. This method is robust with respect to correlations between the initial quality metrics.

To explore the variations in quality amongst the structures in the Protein Data Bank, the scalar quality metric was compared between different subsets of the entire PDB using the appropriate statistical test (using Holm corrections for multiple comparisons). The *Zelig R* package with robust standard errors implemented with the sandwich package was used extensively (Imai *et al.*, 2006; Zeileis, 2004).

4. Results

4.1. Quality models

The initial model used to estimate each quality metric (R factors, Ramachandran violations *etc.*) was fitted using only the resolution of the structure. Explanatory variables (date, number of non-H atoms *etc.*) were then added in an iterative fashion to minimize the Bayesian Information Criterion (BIC; Lindsey & Jones, 1998). The final models, omitting multiplicative coefficients for the linear equations, predicting the quality metrics are given below. In these equations R is an R-factor-like value ($R_{\rm CC}$ is a correlation-coefficient transformed into an R-factor-like value), B is the occupancy-weighted B value, O is the percentage of outliers, C is a constant intercept, $r_{\rm high}$ is the maximum resolution, S is 1 if the data were acquired at a synchrotron and 0 otherwise, N is

Table 3
Results of principal component analysis showing the relative variance described by each component and the quality metrics composing each component.

	Principal components								
	1	2	3	4	5	6	7	8	9
Variance explained (%)	21.6	13.8	12.4	11.2	10.7	9.7	9.3	5.6	5.5
R factor	0.474	−0.034	−0.274	0.172	−0.047	−0.185	0.270	0.286	0.691
$R_{\rm free}$	0.444	−0.077	−0.280	0.266	−0.084	−0.253	0.322	−0.489	−0.483
Real-space *R* value	0.443	0.134	0.043	−0.318	0.242	0.249	−0.432	−0.546	0.278
Outliers	−0.004	0.509	−0.006	−0.055	−0.023	0.627	0.585	−0.038	−0.008
Real-space CC	0.478	0.108	−0.196	−0.038	0.078	0.243	−0.322	0.596	−0.442
Outliers	−0.010	0.713	−0.022	−0.105	−0.517	−0.406	−0.216	−0.014	0.006
Occupancy-weighted *B* value	0.285	0.015	0.598	−0.475	0.216	−0.358	0.350	0.153	−0.120
Outliers	−0.180	0.429	−0.177	0.250	0.780	−0.284	−0.015	0.021	−0.013
Ramachandran violations	0.197	0.109	0.645	0.703	−0.053	0.114	−0.145	−0.010	0.039

the number of non-H atoms, I is the similarity index and D is the date (days since 1 January 1970). The coefficients for the fitted equations can be found in Tables 1 and 2.

$$\begin{aligned}
R\ \text{factor} &= C + r_{\rm high} + S + N + I \\
&\quad + r_{\rm high} \times (S + N) + N \times I, && (1)\\
R_{\rm free} &= C + r_{\rm high} + N + D + S + r_{\rm high} \times (S + D), && (2)\\
R_{\text{real-space}} &= C + r_{\rm high} + S + D + I + N \\
&\quad + r_{\rm high} \times (S + D + I + N) \\
&\quad + D \times (I + S) + I \times N \\
&\quad + r_{\rm high} \times (S \times D + I \times N), && (3)\\
O_{\text{real-space}} &= C + r_{\rm high}, && (4)\\
R_{\rm CC} &= C + r_{\rm high} + S + D + I + N \\
&\quad + r_{\rm high} \times (S + I + N) + D \times (I + S), && (5)\\
O_{\rm CC} &= C + r_{\rm high} + I + N + r_{\rm high} \times N, && (6)\\
B &= C + r_{\rm high} + S + D + I + N \\
&\quad + r_{\rm high} \times (S + D + I + N) + N \times I \\
&\quad + r_{\rm high} \times N \times I, && (7)\\
O_{B\text{-value}} &= C + r_{\rm high} + S + N, && (8)\\
O_{\rm Ramachandran} &= C + r_{\rm high} + S + D + N \\
&\quad + r_{\rm high} \times N + D \times N. && (9)
\end{aligned}$$

4.2. Final quality metric

The nine quality Z scores assigned to each crystallographic structure depend on how the observed quality metrics deviate from their expected values (determined by the models outlined above and normalized to give a standard deviation of 1). Since this nine-dimensional space is difficult to comprehend and some quality metrics evaluate similar features, the quality information was projected onto a three-dimensional space using a principal component analysis. The first three principal components accounted for 22, 14 and 12% of the variability of the Z scores, respectively (Table 3). The first of these components was primarily a combination of the R factor, $R_{\rm free}$, real-space R value and real-space correlation coefficient, *i.e.* global quality features. The second principal component was a combination of the three outlier metrics, *i.e.* local quality features. The third principal component was a combination of the occupancy-weighted B value and Ramachandran violations metrics.

Since these three principal components of the nine-dimensional quality Z values can be seen as arising from different features in the structure and their relative importance cannot be determined, they were equally weighted in determining the overall scalar quality metric. To ease understanding the final scalar quality metric, it was normalized. Therefore, a structure that has a quality metric of 0.0 is average (*i.e.* the quality of its refined structure is what would be expected for a structure of its resolution and complexity). A structure with a quality metric of 1.0 has worse refined quality metrics than 66% of the structures in the PDB. Conversely, a structure with a quality metric of less than zero has been determined better than expected compared with the average structure in the PDB.

4.3. Quality over time

As Tables 1 and 2 show, the date of deposition does not have an appreciable effect on the R factor or other quality metrics. According to the quality models, if an average 2.0 Å structure were deposited in 2007 instead of 1997, the $R_{\rm free}$ would be expected to decrease by 0.2%, the real-space correlation coefficient would slightly decrease to 93.2%, the occupancy-weighted B value would increase by 3.4 Å^2, the number of Ramachandran violators would decrease by 0.5% and all other metrics would stay the same.

4.4. Effect of synchrotron data acquisition

The synchrotron has had a perplexing effect on structure quality, with the change in quality being dependent upon structure resolution (Fig. 1). For an average 2.0 Å structure the expected R factor and $R_{\rm free}$ and two real-space metrics are higher when a synchrotron is used for data collection. The only metrics that are expected to be lower (better) for the synchrotron data are the occupancy-weighted B value (decreased by 4 Å^2) and the percentage of Ramachandran outliers (decreased by 0.3%).

On the other hand, at higher resolutions, data acquired at a synchrotron is expected to be significantly better than data acquired using a rotating anode. For example, at 1.0 Å structures acquired at a synchrotron had expected R factors and $R_{\rm free}$ values that were 1% lower. The occupancy-weighted B value is lower by 3.4 ± 0.5 Å^2 for synchrotron data at this resolution and there were fewer outliers. The real-space R

Table 4
Quality of X-ray crystallographic structures determined by all structural genomics efforts represented in the PDB.

Genomics group	Structures†	Quality‡	p value
JCSG	331 (355)	−0.30 ± 0.06	<0.001
CESG	59 (59)	−0.3 ± 0.1	0.019
SGC	281 (283)	−0.27 ± 0.06	<0.001
NYSGXRC	79 (79)	−0.2 ± 0.1	0.058
SGPP	35 (36)	−0.2 ± 0.2	0.153
MCSG	568 (568)	−0.14 ± 0.05	0.011
RSGI	303 (303)	−0.12 ± 0.06	0.051
NYSGRC	224 (224)	−0.08 ± 0.07	0.310
NESG	191 (195)	−0.05 ± 0.09	0.537
SECSG	79 (79)	0.0 ± 0.1	0.592
TBSGC	132 (137)	0.0 ± 0.1	0.601
BSGI	21 (21)	0.0 ± 0.2	0.765
SPINE	13 (13)	0.0 ± 0.3	0.779
BSGC	71 (71)	0.3 ± 0.1	0.020

† Number of structures from the structural genomics group used in fitting the model and the total number of structures from this group (in parentheses). ‡ The fitted overall quality (and standard error) of the structures submitted by the structural genomics group: negative values indicate genomics efforts that have produced structures better than the PDB as a whole, while positive values indicate genomics efforts where the average structures are worse than the PDB as a whole.

value, however, is still marginally worse for data collected at a synchrotron (an absolute increase of 1.0 ± 0.2%) and the percentage of Ramachandran violations also is expected to be larger by 0.33 ± 0.03%.

Table 5
Relationship between the quality metric and the journal of primary citation.

Journal	Structures†	Quality‡	p value
Eur. J. Biochem.	117 (159)	−0.5 ± 0.1	<0.001
Protein Eng.	51 (96)	−0.3 ± 0.2	0.103
Biochemistry (US)	2468 (3346)	−0.25 ± 0.05	<0.001
Chem. Biol.	134 (154)	−0.19 ± 0.09	0.046
Proteins	398 (541)	−0.18 ± 0.07	0.011
J. Mol. Biol.	2956 (3855)	−0.12 ± 0.05	0.018
Acta Cryst. D	974 (1074)	−0.12 ± 0.06	0.036
Protein Sci.	600 (771)	−0.10 ± 0.07	0.174
Bioorg. Med. Chem. Lett.	184 (195)	−0.10 ± 0.09	0.270
J. Struct. Biol.	56 (83)	−0.1 ± 0.2	0.475
Biophys. J.	60 (71)	−0.1 ± 0.2	0.511
J. Biol. Inorg. Chem.	68 (81)	−0.1 ± 0.1	0.528
Biochem. J.	67 (67)	−0.1 ± 0.1	0.635
J. Biol. Chem.	2475 (2849)	−0.09 ± 0.05	0.082
J. Am. Chem. Soc.	271 (324)	−0.06 ± 0.08	0.444
Structure	1197 (1412)	−0.05 ± 0.06	0.400
FEBS Lett.	137 (173)	0.0 ± 0.1	0.667
J. Bacteriol.	110 (111)	0.0 ± 0.1	0.749
Bioorg. Med. Chem.	52 (53)	0.0 ± 0.2	0.959
J. Med. Chem.	417 (450)	0.04 ± 0.08	0.607
Nature Struct. Biol.	637 (768)	0.08 ± 0.08	0.323
Proc. Natl Acad. Sci. USA	1117 (1324)	0.10 ± 0.06	0.091
J. Virol.	80 (86)	0.1 ± 0.2	0.388
Biochem. Biophys. Res. Commun.	100 (103)	0.2 ± 0.1	0.112
EMBO J.	768 (828)	0.24 ± 0.07	<0.001
Nucleic Acids Res.	166 (199)	0.3 ± 0.2	0.069
Nature	611 (807)	0.35 ± 0.08	<0.001
Mol. Cell	415 (422)	0.36 ± 0.08	<0.001
Science	381 (571)	0.4 ± 0.1	<0.001
Cell	436 (488)	0.5 ± 0.1	<0.001

† Number of structures from the journal used in fitting the model and the total number of structures from this journal (in parentheses). ‡ The fitted overall quality (and standard error) of the structures first published in the journal: negative values indicate journals with structures that are better than the PDB as a whole, while positive values indicate structures worse than the PDB as a whole.

4.5. Structural genomics structure quality

The efforts of structural genomics groups have led to thousands of crystal structures being solved and deposited in the PDB. The scalar quality metric was used to compare structures deposited by structural genomics efforts with the average structure in the PDB. The mean quality of structures from structural genomics groups is better than the structures not from genomics groups (−0.136 *versus* +0.018, $p < 10^{-11}$). Looking at the 14 largest structural genomics efforts, all of which had deposited more than 50 structures at the time of this study, pairwise t-tests showed that CESG, JCSG, SGC, MCSG and BSGC were significantly different from nonstructural genomics structures ($p < 0.05$; Table 4 and Fig. 2).

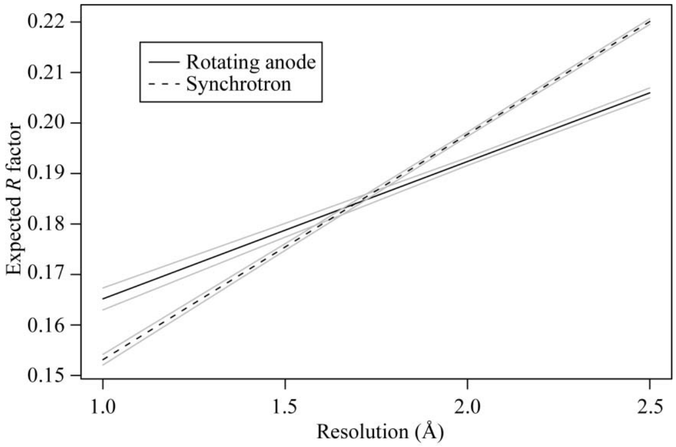

Figure 1
Change in expected R factor for an average structure determined using a rotating-anode source and a synchrotron X-ray source at various resolutions. Shown are the average expected R factor and 95% confidence intervals based on the model fitted using 16 609 structures in the PDB.

4.6. Journal structure quality

Many scientific journals now publish crystal structures. Each structure that contained a primary citation in the header information was assigned to that journal. Then, to evaluate the quality of structures by an individual journal, all structures initially published in that journal were examined. A one-way ANOVA (analysis of variance) shows that all journals do not publish structures of the same quality ($p < 0.001$, *i.e.* at least one journal publishes better or worse structures).

A linear model with the quality metric as the response variable and the journal as the single explanatory variable shows the relationship between each journal and structural quality (Fig. 3). Here the null hypothesis in the significance test is that the quality is average (the coefficient is zero). Protein structures first published in the *European Journal of Biochemistry*, *Biochemistry* (*US*), *Chemistry and Biology*, *Proteins*, *Journal of Molecular Biology* and *Acta Crystallographica Section D* were all significantly better than the average protein found in the PDB ($p < 0.05$). However,

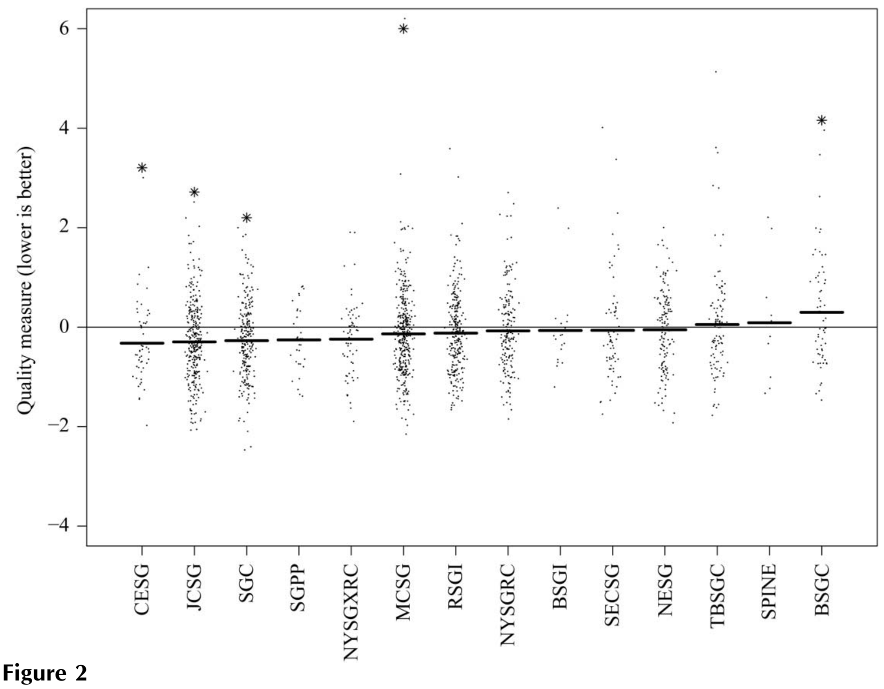

Figure 2
Scatter plot showing the range of quality for the structural genomics groups represented in the PDB. Each dot represents a structure. The thin line along zero is the mean for the entire PDB. The thick line for each genomics endeavor indicates that project's mean structural quality. Stars indicate structural genomics groups with average qualities significantly different from zero ($p < 0.05$). Note: the star is positioned adjacent to the dot representing the poorest quality structure determined by the structural genomics group.

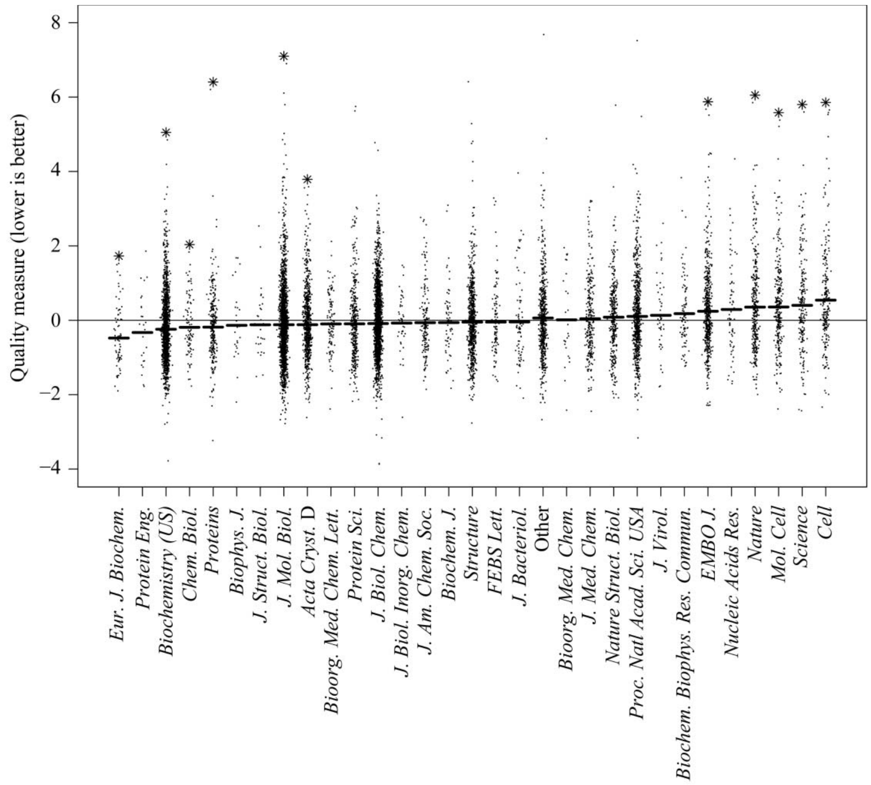

Figure 3
Scatter plot showing the range of quality for the 30 journals with the most primary citations in the PDB. Each dot represents a structure. The thin line along zero is the mean for the entire PDB. The thick line for each journal endeavor indicates that journal's mean structural quality. Stars indicate journals with average qualities significantly different from zero ($p < 0.05$). Note: the star is positioned adjacent to the dot representing the poorest quality structure published in the journal.

structures first published in *EMBO Journal*, *Nature*, *Molecular Cell*, *Science* and *Cell* were significantly worse than the average PDB structure ($p < 0.001$). It is possible to argue that the presence of a few structures of significantly poor/good quality will lead to a bias in these calculations. Another way of looking at the data would be to compare the percentage of structures that are better than the global average reported in every journal (Fig. 4). The y axis is shifted to the 50% position. If the quality of structures is random one would anticipate no large deviations from the line at 50%. There are a number of journals with large bars extending on both sides of the 50% line. While only about 35% of the structures published in *Cell* are above the global average, over 68% of the structures published in *Eur. J. Biochem.* are better than the global average. The significance level of these differences is the corresponding p value reported in Table 5.

5. Discussion

Clearly, a number of factors contribute to the quality of an X-ray crystallographic structure. In addition to the quality of the raw data, which was not a factor examined in this study, the resolution of the structure, the complexity of the structure and the proficiency of the crystallographer all have an impact on the product. Ultimately, structures are refined until the crystallographer is satisfied with the final model and the researchers are able to draw scientific conclusions from the structure.

During refinement, quantitative measures and visual inspection guide the process. The most common quantitative measure monitored during refinement is the R factor; however, $R_{\rm free}$ also is utilized by the crystallographer to prevent overfitting to the data, to determine the optimum balance between geometric restraints and the use of raw data, and to give an indication as to the true quality of the structural model. Because these values are minimized throughout structure refinement, they cannot be examined alone to analyze the quality of structures following refinement. Therefore, the real-space R value, real-space correlation coefficient, occupancy-weighted B value and Ramachandran violations were also examined in this study. Each of these metrics is easily and widely computed following structure deposition.

Since all of these quality metrics are in part dependent upon the resolution and structure complexity, accounting for these factors prior to analysis is necessary for comparisons between structures. The quality metric models were devised to control these variables and to allow comparison between structures differing in resolution and complexity. Additionally, since each of these variables can be estimated prior to starting refinement, the expected quality metrics can be used to track refinement progress. However, this application could bias the individual quality metrics, making comparisons between structures difficult in the future.

By themselves, the quality metric models allow some inferences to be made regarding what factors affect structure quality. For example, (1) and Table 1 validate the common wisdom that with all other factors being the same, higher resolution structures have lower final R factors. Additional insight can be seen in trends over time and owing to synchrotron use.

It might be first assumed that over time the quality of crystal structures in the PDB has improved. However, there seems to be little dependence of the quality metrics on the date of deposition. Among the potential explanations for this lack of a trend is that as a consequence of the exponential increase in structure depositions there may not be enough older structures or there may be too much variability for a linear statistical model to accurately determine a trend over time. Alternatively, as improvements have been made to the science and art of crystallography, researchers may have used these advances to push the resolution and complexity of structures to an ever-increasing degree while maintaining the same quality of the final structure. Finally, since the process of refinement ends when the scientist is satisfied that the structure is able to explain a scientific question, progress in crystallography may be expected to lead to faster and more accessible methods of structure determination, not quality of research.

The effect of synchrotron use upon structure quality is also perplexing. For the average low- to moderate-resolution structure, the quality metrics are actually poorer if a synchrotron was used for data collection. However, as resolution increases, synchrotrons have a positive impact on structure quality. From a technological standpoint, the rapid data acquisition possible at a synchrotron may prompt researchers to stretch data to higher resolutions than would be attempted at a laboratory source (where the required time would be prohibitive.) Thus, a data set that would have been solved as a good 2.5 Å structure using a rotating-anode generator may be refined as an average 2.0 Å or poor 1.5 Å structure at a synchrotron. Additionally, crystals that diffract poorly or not at all at a home source can sometimes have data collected at a synchrotron, biasing the structures collected at a synchrotron to include small, difficult or poorly diffracting crystals.

It is reassuring to observe that the quality of most structures from structural genomics efforts are as good or better than the rest of the PDB. This trend may continue to improve as these groups solve additional structures and further refine quality-control procedures. The significantly better structural genomics consortia include some of the most productive genomics efforts. One potential strength of the JCSG's methods is their inclusion of an explicit and independent quality-control step in the structure pipeline. In addition to the standard validation steps preformed prior to structure deposition, JCSG also utilizes the programs *PROCHECK* (Laskowski *et al.*, 1993; Morris *et al.*, 1992), *SFCHECK* (Vaguine *et al.*, 1999), *WHAT_CHECK* (Hooft *et al.*, 1996), *ERRAT* (Colovos & Yeates, 1993), *DDQ* (van der Akker & Hol, 1999), *PROVE* (Pontius *et al.*, 1996) and *WASP* (Nayal & Di Cera, 1996). The output of these programs includes metrics similar to those used by this study (perhaps biasing their quality to be better than expected).

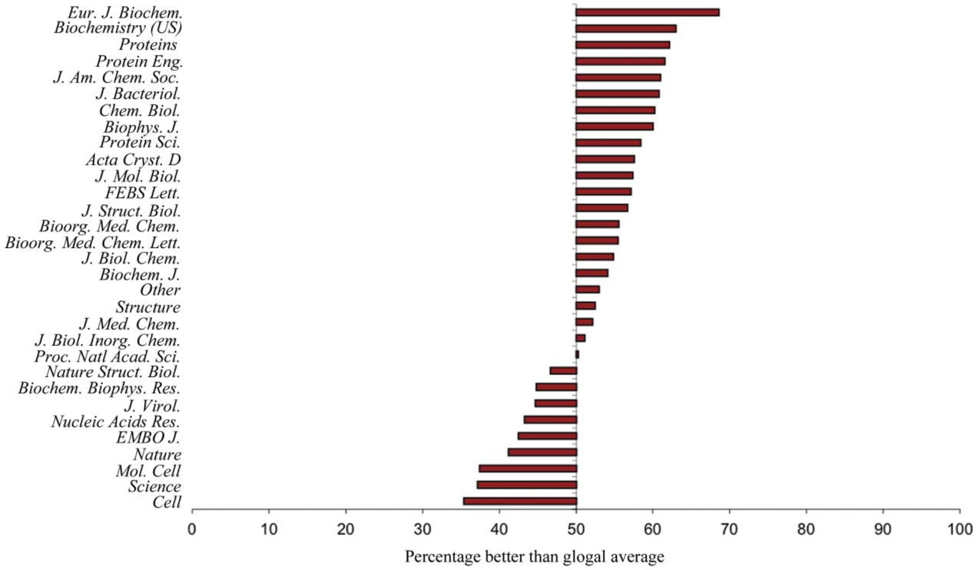

Figure 4
Plot showing the percentage of structures better than average for the different journals. Bars pointing to the right indicate journals with more than 50% structures better than the global average.

The final comparison of structures published by various journals is worrisome. There is a definite bias for the most prestigious general science journals to have published structures that are much worse than expected. This cannot be attributed to difficult structures, effects of resolution *etc.*, since these are accounted for in predicting the expected quality metrics. Furthermore, as illustrated by Table 5, there is not an obvious selection bias in the structures analyzed in this study for the various journals. The journals that were better than average were as well represented in creating the model as those that were worse than average. Combining all nine quality metrics clearly shows that structures published in *EMBO Journal*, *Nature*, *Molecular Cell*, *Science* and *Cell* are below average overall.

There are a few potential criticisms of our conclusions. For example, it may be argued that 'prestigious journals may publish difficult or new structures'. The careful creation of the quality metric models, however, attempts to systematically control for the size of the structure and the novelty of the structure in the analysis. For example, the expected *R* factor only slightly depends on whether the protein was solved for the first or fiftieth time. The results show that this statement is not sufficient to explain the differences between journals.

There are, in the opinion of the authors, four main reasons for the decreased overall quality of structures published in prestigious journals compared with the PDB as a whole. They relate to the complexity inherent in evaluating crystallographic results and the limited resources available to journals during review.

Firstly, in the premier general scientific journals research is published on a wide range of topics. For a manuscript to be accepted, it must be novel and important for science as a whole, not solely crystallography. Thus, papers published in these journals use crystallography as just one aspect of their research. While this reflects the widespread use of protein X-ray crystallography, less educated or experienced researchers may be performing the refinement. The structures may only be refined to the degree that the authors' point can be made, with less effort being made to produce a good structure overall. This is only exacerbated by the pressure to publish rapidly. The risk of being 'scooped' may be great enough to publish results that are good enough rather than good. This does not directly imply that the scientific conclusions published by these articles are unfounded, only that the structures have restricted utility to the scientific community owing to their decreased quality and increased prevalence of errors.

Secondly, also owing to the general nature of these journals, the reviewers (and editors) to whom the papers are sent most are most likely not seasoned structural biochemists. Thus, it is more difficult for the reviewers to insist upon robust structures. Checks that may be required for publication in crystallographic journals may not be performed. Instead, papers clearly lacking in crystallographic quality may be accepted for other reasons.

Thirdly, when manuscripts are reviewed by expert crystallographers, there will often be too little information upon which to effectively judge the quality of the crystal structure. Journals rarely require the submission of protein coordinates and structure factors along with the manuscript, although they may be requested by the reviewer. The more general and high-impact journals often have fewer pages in which the authors can describe their results, further exacerbating the difficulties of the reviewers. It is the opinion of the authors that access to both the atomic coordinates and structure factors are a requirement for the evaluation of a structure. Without both of these it is very difficult to determine whether a structure is accurate. The recent rash of common or disastrous errors that have slipped through the peer-review process, such as the use of the incorrect hand for the electron density, are a testament to this necessity (Chang & Roth, 2001; Pornillos *et al.*, 2005; Chang *et al.*, 2006; Matthews, 2007; Wang *et al.*, 2007).

The most reasonable method for improving the quality of published crystallography structures is to require that coordinates and structure factors be submitted along with the manuscript. Even with diligent authors and reviewers, quality will always vary and structures will exhibit a variety of imperfections. Therefore, any computational study must evaluate the quality of the structures used with respect to their specific methodology. For example, it is ridiculous to base precise conclusions upon a region of a homology model missing in the original template structure. Likewise, in evaluating the binding mode of ligands such as natural cofactors or synthetic drugs, the local electron density needs to be carefully examined. Often, identification and placement of water molecules, anions and cations are ambiguous. More importantly, ligands are often partially occupied. While this may be sufficient for interpretation in the context of the initial experiment, extrapolation of this structure for use in modeling may be ill-conceived.

We hope to develop a web service where users will be able to upload coordinates and structure factors to determine the quality of the model. The service will allow users to choose the group of structures with which they would like their model compared.

We thank Daniel Ferraro, Lokesh Gakhar, Shahram Khademi, Adrian Elcock and Bryce Plapp for their ideas and assistance. ENB is a University of Iowa MSTP trainee and would like to acknowledge financial support through a fellowship from the University of Iowa Center for Biocatalysis and Bioprocessing.

References

Abola, E., Kuhn, P., Earnest, T. & Stevens, R. C. (2000). *Nature Struct. Biol.* **7**, 973–977.

Akker, F. van den & Hol, W. G. J. (1999). *Acta Cryst.* D**55**, 206–218.

Berman, H. M., Henrick, K. & Nakamura, H. (2003). *Nature Struct. Biol.* **10**, 980.

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.

Brändén, C.-I. & Jones, T. A. (1990). *Nature (London)*, **343**, 687–689.

Brünger, A. T. (1993). *Acta Cryst.* D**49**, 24–36.

Chang, G. & Roth, C. (2001). *Science*, **293**, 1793–800.

Chang, G., Roth, C. B., Reyes, C. L., Pornillos, O., Chen, Y.-J. & Chen, A. (2006). *Science*, **314**, 1875.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Colovos, C. & Yeates, T. O. (1993). *Protein Sci.* **2**, 1511–1519.

Fabiola, F., Korostelev, A. & Chapman, M. S. (2006). *Acta Cryst.* D**62**, 227–238.

Geerlof, A. *et al.* (2006). *Acta Cryst.* D**62**, 1125–1136.

Hooft, R. W. W., Vriend, G., Sander, C. & Abola, E. E. (1996). *Nature (London)*, **381**, 272.

Huber, P. J. (1981). *Robust Statistics.* New York: Wiley.

Imai, K., King, G. & Lau, O. (2006). *Zelig: Everyone's Statistical Software.* http://gking.harvard.edu/zelig.

Jaskolski, M., Gilski, M., Dauter, Z. & Wlodawer, A. (2007). *Acta Cryst.* D**63**, 611–620.

Jiang, J. & Sweet, R. M. (2004). *J. Synchrotron Rad.* **11**, 319–327.

Kleywegt, G. J., Harris, M. R., Zou, J., Taylor, T. C., Wählby, A. & Jones, T. A. (2004). *Acta Cryst.* D**60**, 2240–2249.

Kleywegt, G. J. & Jones, T. A. (1996). *Structure*, **4**, 1395–1400.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.

Leslie, A. G. W. (1992). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **26**.

Lindsey, J. K. & Jones, B. (1998). *Stat. Med.* **17**, 59–68.

Liu, Z.-J. *et al.* (2005). *Acta Cryst.* D**61**, 679–684.

Matthews, B. (2007). *Protein Sci.* **16**, 1013–1016.

Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1992). *Proteins*, **12**, 345–364.

Nayal, M. & Di Cera, E. (1996). *J. Mol. Biol.* **256**, 228–234.

Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.

Parkin, S. (2000). *Acta Cryst.* A**56**, 157–162.

Peat, T., de La Fortelle, E., Culpepper, J. & Newman, J. (2002). *Acta Cryst.* D**58**, 1968–1970.

Pontius, J., Richelle, J. & Wodak, S. J. (1996). *J. Mol. Biol.* **264**, 121–136.

Pornillos, O., Chen, Y.-J., Chen, A. P. & Chang, G. (2005). *Science*, **310**, 1950–1953.

R Development Core Team (2006). *The R Project for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org.

Rupp, B., Segelke, B. W., Krupka, H. I., Lekin, T., Schäfer, J., Zemla, A., Toppani, D., Snell, G. & Earnest, T. (2002). *Acta Cryst.* D**58**, 1514–1518.

Srinivasan, R. (1966). *Acta Cryst.* **20**, 143–144.

Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *Acta Cryst.* D**55**, 191–205.

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S*, 4th ed. New York: Springer-Verlag.

Wang, J., Wlodawer, A. & Dauter, Z. (2007). *Acta Cryst.* D**63**, 751–758.

Zeileis, A. (2004). *J Stat. Softw.* **11**, 1–17.