Improved *R*-factors for diffraction data analysis in macromolecular crystallography

Kay Diederichs[1] and P. Andrew Karplus[2]

The quantity R_{sym} (also called R_{merge}) is almost universally used for describing X-ray diffraction data quality. Here, we prove that R_{sym} is seriously flawed, because it has an implicit dependence on the redundancy of the data. A corrected *R*-factor, R_{meas}, is introduced as the equivalent robust indicator of data consistency. In addition, we introduce R_{mrgd}, an *R*-factor that reflects the gain in accuracy upon averaging of equivalent reflections, as a useful indicator of the quality of reduced data. These new data quality indicators better reveal the benefits of highly redundant data and should stimulate improvements in data quality through increased merging of data from multiple crystals.

R_{sym} (sometimes called R_{merge}), is the most widespread statistic used to indicate data quality for macromolecular crystallography[1,2], and with the advent of area detectors for small molecule crystallography it is a standard quality indicator for that field as well[3]. It is defined as:

$$R_{sym} = \frac{\sum\limits_{h} | \hat{I}_h - I_{h,i} |}{\sum\limits_{h} \sum\limits_{i}^{n_h} I_{h,i}} \qquad \text{with } \hat{I}_h = \frac{1}{n_h} \sum_{i}^{n_h} I_{h,i}$$

Arndt[4] introduced R_{sym} as a reliability indicator for data collected by precession photography, where R_{sym} was specifically summed over symmetry-related intensities on the same film, and R_{sca}, calculated in an analogous fashion, reported the agreement of identical reflections measured on different films. As oscillation photography was introduced, so that symmetry related reflections were not commonly on the same film, it appears that the original R_{sym} and R_{sca} were combined into the present day R_{sym} which is summed over all observed equivalent reflections.

R_{sym} is commonly used to guide decisions during data reduction, such as determining to what resolution data are reliable, and whether two crystals are isomorphous, so that their data should be merged together. A single R_{sym} value is generally reported in publications to summarize the data quality. Overall R_{sym} values of <5%, 5–10%, 10–20% and >20% are taken to indicate good, usable, marginal, and questionable quality data respectively[2]. Here, we present empirical and mathematical analyses proving that R_{sym} is an inherently unreliable indicator of data quality. We also present alternate indicators that provide more robust measures of the quality of the individual measurements as well as of the final reduced data set. We expect that the application of the ideas described here will result in improved primary data quality, and ultimately in more accurate macromolecular structures.

Experimental data

The analyses presented here hold true for diffraction data measured with various detector/software combinations, but for simplicity, we present analyses based on three sets of data collected from crystals of the enzyme urease with Cys 319 from the α-chain mutated to Ala[5]. These crystals are isomorphous with wild-type urease and grow in space group $I2_13$ with a=170.8 Å[6,7]. Independent 2 Å resolution data sets were collected from each of three crystals which had been soaked at pH values of 6.5, 7.5 and 8.5; these are designated Ure_1, Ure_2, and Ure_3 respectively. Difference Fourier maps showed no apparent structural changes between the data sets, so for the purposes of this study, we are treating them as equivalent. The three crystals had approximate volumes of 0.036, 0.027 and 0.036 mm^3, and the data sets were successively collected using a Rigaku RU-200 rotating anode (Cu-Kα, 50kV, 150 mA) and a pair of SDMS MARKII multiwire detectors[8] placed at 2θ-values of 14° and 34° and at distances of 719 and 780 mm respectively. Each data set consisted of three 50° ω-sweeps, using either 0.1° or 0.08° steps and an ω-scan rate of 10 min degree^{-1}. The data from all sweeps were reduced and merged together using SCALEPACK[9]. Conventional statistics are reported in Table 1 for six different data reductions based on the data from these three crystals: data reductions for each of the sweeps of the first crystal (Ure_1A, Ure_1B, Ure_1C), the three complete data sets (Ure_1, Ure_2 and Ure_3), and a data set obtained by merging all three crystals together (Ure_123). As expected due to the larger sizes of the crystals, the Ure_1 and Ure_3 data sets yield slightly better statistics than the Ure_2 data set.

R_{sym} inherently depends on multiplicity

When data from multiple crystals are merged, the R_{sym} of the combined data set is commonly higher than those of the individual data sets. Similarly, R_{sym} usually rises as a function of frame number during a single measurement. Whereas this prop-

[1] Universität Konstanz, Fakultät für Biologie, Postfach 5560 (M656), D-78434 Konstanz, Germany [2] Section of Biochemistry, Molecular and Cell Biology, Cornell University, Ithaca, NY 14853, U.S.A.

Correspondence should be addressed to P.A. K. *pak4@cornell.edu* or K.D. *kay.diederichs@uni.konstanz.de*

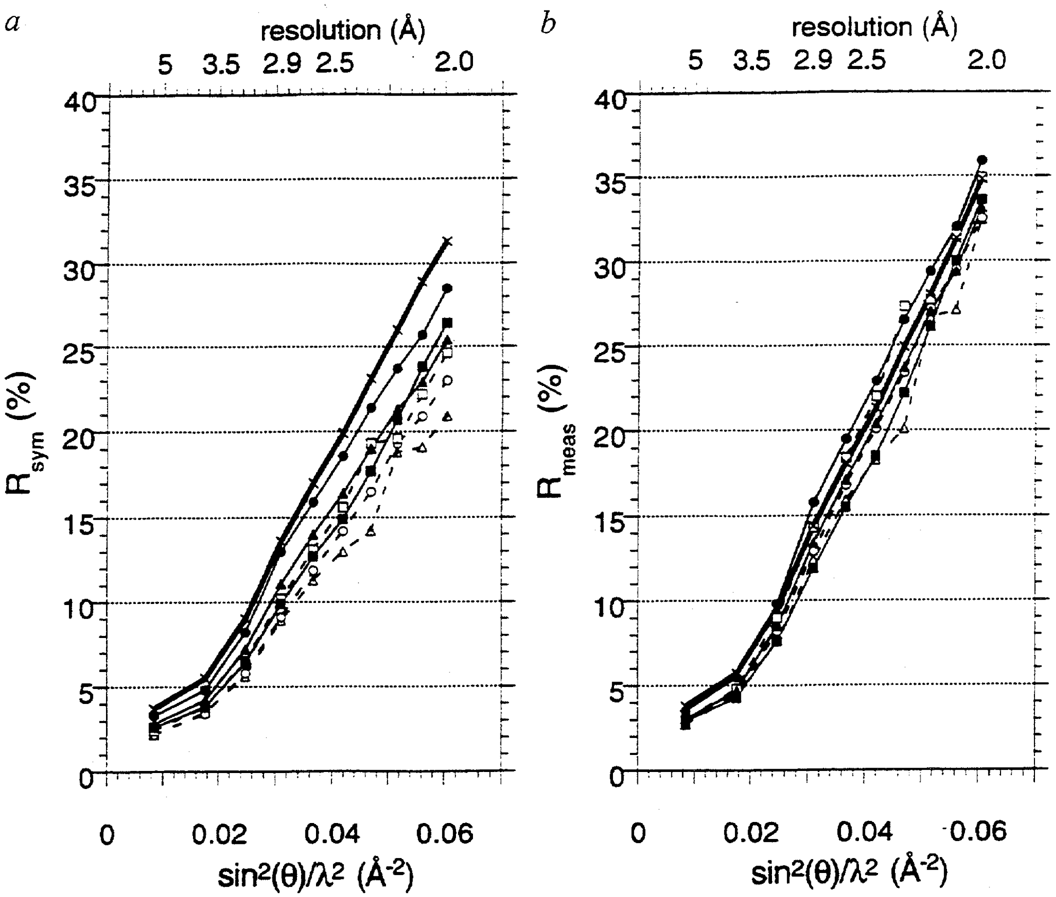

Fig. 1 R_{sym} and R_{meas} as a function of resolution for various reduced urease data sets. *a*, R_{sym} is shown for data sets Ure_1A (open triangle), Ure_1B (open circle), Ure_1C (open square) Ure_1 (filled triangle), Ure_2 (filled circle), Ure_3 (filled square), and URE_123 (X). Within the two groups of three data sets with similar redundancy, the R_{sym} values are comparable and reflect that Ure_2> Ure_1, Ure_3, consistent with the crystal sizes. However, the artifact produced by R_{sym} is that the partial data sets appear much better, and the data set merging all three crystals gives the highest R_{sym}. *b*, R_{meas} shown for the same six data sets (same symbols). Here , the Ure_2 data set is clearly seen to have the lowest quality as is appropriate due to the small crystal size, and the partial data sets from crystal 1 are seen to be of similar quality as the full dataset Ure_1. The close overlap of the curves provides evidence that there is little systematic difference between the individual data sets.

erty of R_{sym} has been noted[10,11], it has not been comprehensively documented nor formally deduced. Such increases in R_{sym} are generally attributed to slight non-isomorphism or systematic errors (mainly absorption, radiation decay) between the data segments, and large increases are taken as grounds for not merging data sets, for discarding the final frames of a data collection run, and/or for lowering the threshold for outlier rejection. This behaviour of R_{sym} has contributed to the common practice which favours using complete data sets from single crystals whenever possible.

As is illustrated in Table 1 and Fig. 1*a*, our urease test data show this common behaviour. The R_{sym} values are lowest for the smallest segments of data (the individual sweeps Ure_1A, Ure_1B, and Ure_1C), they are intermediate for the single data sets (Ure_1, Ure_2, and Ure_3) and they are highest for the merged data set (Ure_123). The higher R_{sym} of Ure_123 compared to Ure_1A, Ure_1B, and Ure_1C (~1.7-fold increase overall and ~1.5-fold increase at high resolution) would normally be interpreted to indicate systematic differences between the data sets being merged. However, an analysis of R_{sym} as a function of multiplicity suggests that the cause of the increased R_{sym} values is not systematic differences between portions of the data, but rather an undesirable dependence of R_{sym} on multiplicity. Using a single set of reflections, so that the only variable is multiplicity, the value of R_{sym} increases smoothly in an asymptotic manner from 13.4% to near 19% as the multiplicity increases from two to twelve (Fig. 2). Control

Table 1 Data collection statistics summary[1]

Data set	Unique reflections	Completeness (%)	Multiplicity	R_{sym}	R_{meas}	$<I/\sigma>^2$	R_{mrgd-F}
Ure_1A	37776	68(40)	1.6(1.2)	4.1(22.9)	5.5(32.4)	8.6(2.6)	8.0(28.6)
Ure_1B	37049	67(36)	1.6(1.3)	4.6(23.0)	6.2(32.5)	8.6(2.6)	9.2(28.4)
Ure_1C	38502	69(38)	1.5(1.2)	4.5(24.6)	6.0(34.9)	8.0(2.4)	8.7(30.2)
Ure_1	53925	97(70)	3.3(2.0)	6.0(25.4)	7.1(33.1)	11.6(3.2)	10.3(27.4)
Ure_2	52787	95(64)	3.3(2.0)	6.8(28.5)	7.9(35.9)	10.4(2.9)	10.9(27.3)
Ure_3	53814	96(68)	3.3(2.0)	5.5(26.4)	6.5(33.6)	12.7(3.4)	9.5(27.1)
Ure_123	54941	98(85)	9.6(4.7)	7.7(31.3)	8.1(34.8)	20.1(4.9)	6.9(20.2)

[1] Overall values of crystallographic indicators are given for all measured data to 2.0 Å resolution. Numbers in brackets indicate the values in the highest resolution range (2.07–2.0Å).
[2] $\langle I/\sigma \rangle$ of the merged intensities.

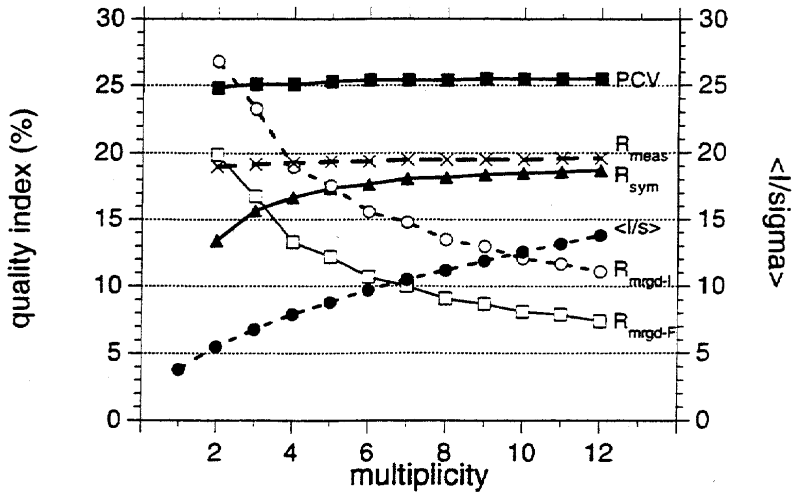

Fig. 2 The behaviour of data quality indicators as a function of multiplicity. Solid lines are used for the three indicators of the quality of individual measurements: R_{sym} (open square); R_{mease} (closed squares); PCV (closed triangles). Dashed lines are used for the indicators of reduced data quality: $R_{mrgd\text{-}I}$ (closed circles); $R_{mrgd\text{-}F}$ (open circles); $<I/\sigma_I>$ (X). Note that the value of 3.8 observed for the $<I/\sigma_I>$ of the individual measurements (multiplicity of 1) and the PCV of 25.5% are consistent with the relation $PCV=100*<I/\sigma_I>^{-1}$ (see box). This analysis was carried out using all 3137 reflections which were between 3 and 2 Å resolution and were observed ≥12 times in the Ure_123 dataset. For these high-multiplicity reflections, the observations in excess of the 12th were discarded. Then, for each multiplicity value n given on the abscissa, n out of the 12 observations of each reflection were selected randomly and used for calculating the statistical indicators. This approach allowed each of the 12 observations to contribute equally for all multiplicity values.

calculations using fictitious data with perfectly Gaussian error prove that this ~1.4-fold increase reflects an inherent property of R_{sym} (Fig. 3). Since the ability to accurately estimate data quality must improve with increased number of data, the R_{sym} value truly reflecting data accuracy is the asymptotic value obtained at high redundancy.

Two robust alternate indicators: R_{meas} and PCV

Mathematical analysis (see Box) makes explicit how the contributions of reflections to R_{sym} depend on their multiplicity, and leads us to propose two alternate well-behaved measures of data quality. The first is an adjusted R_{sym} which we have dubbed R_{meas} because it should accurately reflect the reliability of individual measurements, independent of multiplicity. The mathematical analysis (see Box) shows that a robust variant of R_{sym} can be obtained by adjusting each reflection's contribution by a factor of $\sqrt{n_h/(n_h-1)}$, where n_h is the multiplicity:

(2)

$$R_{meas} = \frac{\sum_h \sqrt{\frac{n_h}{n_h-1}} \sum_i^{n_h} |\hat{I}_h - I_{h,i}|}{\sum_h \sum_i^{n_h} I_{h,i}} \quad \text{with } \hat{I}_h = \frac{1}{n_h}\sum_i^{n_h} I_{h,i}$$

For data sets with fixed redundancy *n*, this is equivalent to multiplying R_{sym} by the factor $\sqrt{n/(n-1)}$, but for typical real data sets, it is important that the factor be placed inside the sum so that the contributions from the individual reflections are appropriately weighted according to their multiplicity. The magnitude of the scaled difference terms $\sqrt{n_h/(n_h-1)}\ |\hat{I}_h - I_{h,i}|$ is not correlated with the multiplicity n_h of a reflection, and R_{meas} values for low redundancy data sets are as high as those of high redundancy data sets (Fig. 3). This also means that the merging R_{meas} of two data sets should be close to the average of their individual, internal R_{meas} values, if no systematic differences (for example, anisomorphism) exist between the data sets. This is in sharp contrast with the behaviour of R_{sym}.

A second robust indicator of data quality, which is commonly used in statistics, is the pooled coefficient of variation (PCV), in which the pooled standard deviation (the statistically valid measure of the noise level) is divided by the sum of the intensities (the signal level):

(3)

$$PCV = \frac{\sum_h \sqrt{\frac{1}{n_h-1}\sum_i^{n_h} (I_{h,i} - \hat{I}_h)^2}}{\sum_h \hat{I}_h}$$

A mathematical analysis shows that for Gaussian distributed error, PCV should be exactly a factor of $\sqrt{\Pi/2}$ (~ 1.25-fold) larger than R_{meas} (see Box and Fig. 3). We suspect that despite the greater statistical information content of the PCV, crystallographers will prefer to use R_{meas} because it gives values which can be compared with the past literature. PCV is related to other 'quadratic' *R*-factors used in popular data reduction software[9,11,12], but it includes the important factor $1/(n_h-1)$ that makes it robust with respect to multiplicity.

For the urease data, examining the behaviours of both R_{meas} and PCV as a function of multiplicity (Fig. 2) it can be seen that both are relatively constant with respect to redundancy, and the PCV is, indeed, about 1.25-fold higher than R_{meas}. Using R_{meas} to assess the six merged urease data sets shows that the misleading behaviour seen for R_{sym} is abolished (compare Fig. 1*b versus* Fig. 1*a*). Independent of how much data is merged together, the R_{meas} values match closely the R_{sym} values seen for the high multiplicity Ure_123 data set and indicate that there are no large systematic differences between the data sets.

The discrepancies between R_{sym} and R_{meas} are largest for data with low multiplicity and can be as large as a factor of $\sqrt{2}$. Although modern (area detector) data sets often have high redundancy so that the problems with R_{sym} are lessened, not all do. A recent 1.95 Å resolution haemoglobin structure was based on data with a completeness of 76% (38% in the highest resolution bin) and a multiplicity of 1.7 (1.1 in the highest bin)[13]. This is quite similar to the completeness and multiplicity of the Ure_1A, B, or C data sets reported here, and indicates that the reported overall R_{sym} of 4.1 strongly overestimates the data quality. Another case is the structure of the C-reactive protein, in which data were carefully selected from 33 crystals to yield a data set with 74% completeness and 1.5-fold multiplicity with a high R_{sym} of 25.5%[14]. Given the low multiplicity, the true data quality (as measured by R_{meas}) would be significantly worse. Finally, we note that even for those data sets with high multiplicity, it is common that the

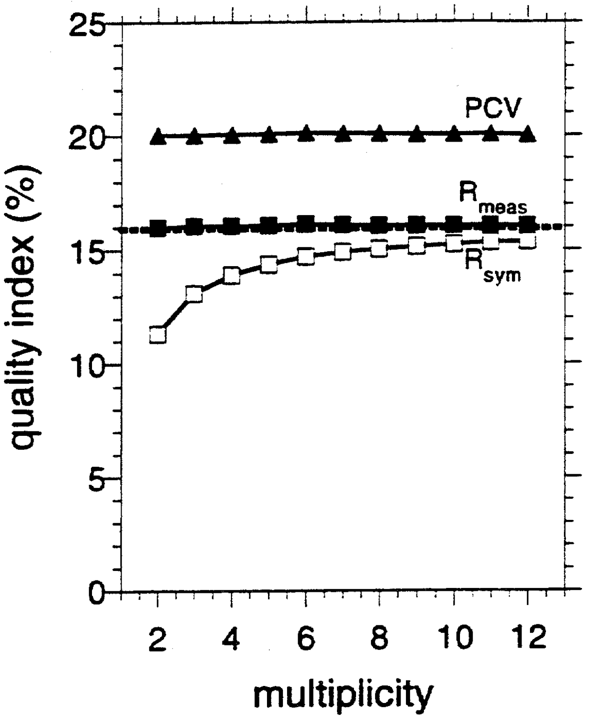

Fig. 3 The figure is equivalent to Fig. 2 but reports statistics based on a computer generated set of 10,000 reflections which have intensities that are normally distributed with a $\bar{I}_{true}$= 50, σ = 10. Plotted are R_{sym} (open squares), R_{rmeas} (filled squares), and PCV (triangles). The dashed line marks the value for R_{sym} obtained for infinite multiplicity (calculated by substituting $\bar{I}_{true}$ for $<I>$ in equation 1).

highest resolution bin has the lowest multiplicity. Thus, exactly where one should be most concerned about data quality, the inherent properties of R_{sym} make it least reliable.

For crystals of unknown space group, usually the symmetry of intensities in a given Bravais lattice is found by comparing R_{sym} values after reducing a set of frames in alternative space groups (for example, $P4_{(0,1,2,3)}$ *versus* $P4_{(0,1,2,3)}2_{(0,1)}2_{(0,1)}$). In this scenario, R_{sym} will always favour the space group(s) with lower symmetry, as the average multiplicity of reflections will be less than in the high symmetry space group(s). Use of R_{meas}, on the other hand, would give unbiased indications toward the highest symmetry compatible with the diffraction pattern, and would therefore help to avoid space group assignment errors[15].

An additional problem with 'overall' reliability factors

The use of R_{meas} (or PCV) instead of R_{sym} removes misleading impressions of data quality that make less redundant data appear better. However, it does not fix an additional problem inherent in overall R_{meas} values. This additional problem occurs because reflections contribute to the overall R_{meas} in proportion to their multiplicity, and multiplicity may be distributed differently in various data sets. In the urease data sets, this effect can be seen in the variation of the overall R_{meas} values in Table 1 from 5.5 to 8.1, despite the fact that the R_{meas} values as a function of resolution vary much less (Fig. 1*b*).

Within a narrow resolution range, R_{meas} has meaning, because multiplicity is fairly constant and (more importantly) should not be correlated with a reflection's reliability. However, when data across wide ranges of resolution are combined, the overall R_{meas} depends heavily on how the multiplicity is distributed *versus* resolution (Fig. 4). Although the Ure_1C and Ure_123 data sets have nearly identical R_{meas} in all resolution ranges (Fig. 1*b*), the Ure_1C data set has a lower overall R_{meas} (6.0 % *versus* 8.1 %) because it has higher relative multiplicity at low resolution. In extreme cases, the overall index could vary between values approximating the R_{meas} of the low resolution data and values approximating the R_{meas} of the high resolution data.

For 'typical' data sets, the low resolution (more accurate) data have higher multiplicity and dominate to yield a low overall R_{meas}. However, many data sets are not typical. For instance merging a lower resolution data set with a complete high resolution data set (as in the report of a camel antibody structure[16]) would tend to skew the multiplicity toward lower resolution reflections, whereas the merging of a data set consisting of exclusively higher resolution reflections (such as may be obtained by swinging out a detector) would tend to skew the multiplicity toward higher resolution[17]. In many published reports, the importance of such effects are hard to assess. For instance, the distribution of multiplicity is not described in the C-reactive protein study cited above[14], but if the multiplicity involves largely the higher resolution reflections, the overall R_{sym} of 25.5% would not be as bad as it appears.

Indicators of reduced data quality

The above discussion shows that R_{meas}, as opposed to R_{sym}, provides a robust measure of the consistency of individual measurements. While that is important, it is also desirable to have measures which estimate the reliability of the *reduced* data. $<I/\sigma_I>$ is such an indicator , but no generally accepted *R*-factors for this purpose exist. The reduced data will, in general, be more accurate than individual measurements, because the averaging of multiple observations leads to increased accuracy (a theoretical factor of $\sqrt{n_h}$ for reflections with multiplicity n_h) that is not reflected in R_{meas} (or the other measures)[18]. For a number of years, one of us (P.A.K.) has used a statistic called R_{int} which does reflect much of the accuracy gained through high redundancy, because it is calculated from data that have been partially merged[19,20]. R_{int} is simply the *R*-factor between the amplitudes of Friedel pairs,

$$(4) \qquad R_{int} = \frac{\sum |F_{h+} - F_{h-}|}{0.5 * \sum F_{h+} + F_{h-}}$$

and can be calculated from the output of a special data reduction run that does not merge the Friedel pairs. R_{int} underestimates the final data quality because of an additional improvement (up to a factor of $\sqrt{2}$) to come from the merging of F^+ and F^-.

Here, we generalize the concept of R_{int} by randomly assigning the n_h observations of a unique reflection *h* (originating from subsets of a single data collection run, or from two or more data collection runs with possibly different crystals) to two disjoint sets P,Q with $n_{h,P} = \mathrm{int}(n_h/2)$ and $n_{h,Q} = n_h - n_{h,P}$ members, and average observations in these sets separately (for simplicity, equations (5) and (8) use unweighted averages, but in practice weighting with the experimental σ_I values would be most appropriate). We thus calculate

$$(5) \qquad R_{mrgd\text{-}I} = \frac{\sum |I_{h,P} - I_{hQ}|}{0.5 * \sum I_{h,P} + I_{h,Q}} \text{ with } I_{h,P} = \frac{1}{n_{h,P}} \sum_{i \in P}^{n_{h,P}} I_{h,i}$$

$$\text{and } I_{h,Q} = \frac{1}{n_{h,Q}} \sum_{i \in Q}^{n_{h,Q}} I_{h,i}$$

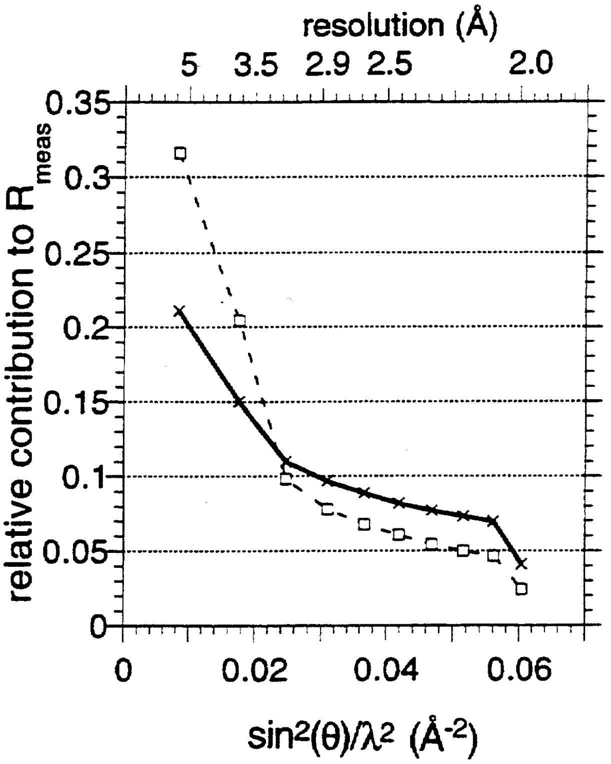

Fig. 4 Variability of the contribution to R_{meas} as a function of resolution. The distribution is shown for the Ure_1C (open square) and the Ure_123 (x) data sets. The total number of reflections contributing to the R_{meas} sums for the two data sets are 36569 and 525934 reflections, respectively. The fractional contribution of each resolution bin was calculated by dividing the number of contributing observations in that bin by the total number of contributing observations for the whole data set. Although Ure_1C and Ure_123 have nearly identical values in each resolution bin (see fig. 1*b*), the overall R_{meas} of Ure_123 is ~30% larger than that of Ure_1C (8.1% *versus* 6.0%) due to the different distribution of multiplicity. It should be noted that the relative contribution to R_{meas} is not identical to multiplicity because when multiplicity is 1.0, *no* reflections contribute to R_{meas}. In addition to the multiplicity related bias, it should be noted that all 'overall' *R*-factors have an intrinsic bias in which the largest reflections have the largest influence. Overall *R*-factors calculated on ***F*** are less influenced by this bias because structure factors have a much smaller range than intensities.

$I_{h,P}$ and $I_{h,Q}$ represent the partially averaged, and therefore improved, estimates of the true intensity. The term $R_{mrgd\text{-}I}$ is chosen to indicate that this *R*-factor reflects the quality of the merged intensities. As for R_{int}, the final estimates of the $\bar{I}_h$ may be up to a factor of $\sqrt{2}$ more accurate than $R_{mrgd\text{-}I}$ indicates. Fig. 2 includes the properties of $R_{mrgd\text{-}I}$ as a function of multiplicity. At a multiplicity of 4, $R_{mrgd\text{-}I}$ equals R_{meas}, and the drop seen for $R_{mrgd\text{-}I}$ with increasing multiplicity closely matches the expected factor of $\sqrt{(n/2)}$.

Most crystallographic calculations, such as phasing and structure refinement, are carried out using structure factor amplitudes rather than intensities. For this reason, it is also relevant to know the reliability of the structure factors on a scale useful for comparisons with the *R*-factors used for assessing the level of signal in heavy atom derivatives (R_{iso}), and those used to judge the progress of model refinement (R_{cryst}, R_{free}). We suggest that an R_{mrgd} calculated using structure factors rather than intensity data, $R_{mrgd\text{-}F}$, is an appropriate measure, because it is exactly analogous to the *R*-factors R_{cryst} (or R_{free}) and R_{iso} between two data sets A and B:

(6)
$$R_{cryst} = \frac{\sum_{h} | F_{h,\,calc} - F_{h,\,obs} |}{\sum_{h} F_{h,\,obs}}$$

(7)
$$R_{iso} = \frac{\sum | F_{h,A} - F_{h,B} |}{0.5 * \sum F_{h,A} + F_{h,B}}$$

To overcome the problem of negative intensities for which the square root is not defined, we suggest the use of pseudo-amplitudes

$$A_I = \begin{cases} \sqrt{I} \text{ if } I \geq 0 \\ -\sqrt{I} \text{ if } I < 0 \end{cases}$$

solely for use in the $R_{mrgd\text{-}F}$ equations[21]. Pseudoamplitudes are not physically meaningful, but they have the desirable property that even negative reflections contribute to the overall $R_{mrgd\text{-}F}$ in a sensible way. We therefore define

(8)
$$R_{mrgd\text{-}F} = \frac{\sum | A_{I_{h,P}} - A_{I_{h,Q}} |}{0.5 * \sum A_{I_{h,P}} + A_{I_{h,Q}}} \text{ with } I_{h,P} \text{ and } I_{h,Q} \text{ as above}$$

as a measure for the quality of the reduced amplitudes. Once again, $R_{mrgd\text{-}F}$ does not reflect the accuracy gain (up to a factor of $\sqrt{2}$) due to merging the P and Q subsets, and thus may somewhat underestimate the data quality. Being conservative and reporting $R_{mrgd\text{-}F}$ as the data quality is somewhat like the convention of reporting d_{min} as the resolution limit, even though it is known that for perfect data the resolution is technically 0.92* d_{min}[22].

$R_{mrgd\text{-}F}$ is a useful quantity as it can be compared with the many *R*-factors calculated during the course of a crystallographic structure solution. Obviously, the quality of the reduced data limits the accuracy of the final model[23,24]. In the past, R_{sym} has been used for comparison with R_{cryst} and R^{iso}, but this is not appropriate. At high resolution, R_{sym} is often seen to be much higher than the final R_{cryst}[25] or R_{free}[26]; R_{sym} can be larger than the isomophous change in a useful heavy atom derivative[27,28]; and in a MAD phasing analysis R_{sym} can be much higher than the level of anomalous signal[18]. In practice, $R_{mrgd\text{-}F}$ should provide an approximate lower limit on R_{free}, the cross-validated R_{cryst}, which is most useful for evaluating refinement progress and model accuracy. Such a measure allows one to assess when phasing accuracy or refinement progress is truly being limited by the data quality.

As noted above, the values of R_{mrgd} are only reliable in the absence of large systematic differences between subsets of the data being merged. With R_{meas} now available as a robust indicator of data quality, significant differences between two data sets will be flagged by merging statistics showing an increase in R_{meas} compared to the individual subsets. An additional test for systematic errors is to compare the R_{mrgd} value for the combined data (with random assignments to the disjoint sets P and Q), with an R_{iso} calculation between the separately reduced subsets of data (equivalent to an R_{mrgd} calculation in which the disjoint sets correspond to the data from two different subsets of data A, B). Systematic differences between the datasets A,B would (statistically) cause R_{iso} to be higher than R_{mrgd} (provided $n_{h,A}$,$n_{h,B}$ >1 for the common *h*). In cases for which the difference is small, merging of datasets is justified and R_{mrgd} can be considered a good estimate of the reliability of the reduced data. We expect that practical experience with these indicators will be required to decide how this information is best applied (that is, how much systematic error is too much).

A recommendation

Based on these observations, we suggest that crystallographic data reduction and scaling programs should report R_{meas}, PCV and $R_{\mathrm{mrgd\text{-}F}}$: R_{meas} will enable crystallographers to better assess the internal consistency of their measured data sets as well as the agreement between different data sets; PCV (or R_{meas} if errors are normally distributed) can be used to evaluate whether the reported $<I/\sigma>$ values match the true scatter in the measurements (see Box and Fig. 2); and $R_{mrgd\text{-}F}$ is the indicator of the quality of the final merged structure factor amplitudes, and makes direct comparisons with R_{cryst} and R_{iso} possible. $R_{\mathrm{mrgd\text{-}F}}$, as opposed to R_{meas}, should be considered as the most important indicator of final data quality in crystallographic publications, since it is the reduced data set which is used to determine structures. Finally, rather than reporting a single overall value for data quality, it is important to provide information about data quality as a function of resolution, at a minimum including separate statistics for the highest resolution data.

In addition to allowing a more accurate assessment of data quality, based on our experience, the use of R_{meas} and $R_{\mathrm{mrgd\text{-}F}}$ will reveal that systematic errors between data sets from different crystals are often rather small compared to the random errors, especially in the higher resolution range. This revelation should stimulate a shift in data collection strategies, so that the current bias toward using single crystals for complete data sets whenever possible will shift to favour multiple crystal data sets which have increased multiplicity and hence more accurate reduced structure factors. Such a change in strategy has great potential to improve the quality and completeness of the high resolution data and serve to enhance the accuracy of details seen in macromolecular structures.

Acknowledgements

We thank Matthew Pearson for preparing the urease data sets for analysis, Hugh Gauch for suggesting the use of the PCV, and Alexander Wlodawer for comments on the manuscript.

Received 25 October 1996; Accepted 31 January 1997

Note: A program that reads a SCALEPACK[9] *or a XDS*[29] *output file and calculates all R-factors mentioned in this paper can be requested by e-mail from one of the authors (kay.diederichs@uni-konstanz.de).*

1. Blundell, T.L. & Johnson, L.N. *Protein crystallography*. p. 331 . Academic Press, New York (1976).
2. McRee, D.E. *Practical protein crystallography* . p. 101. Academic Press, San Diego (1993).
3. Ladd, M.F.C & Palmer, R.A. *Structure determination by X-ray crystallography*, Third ed. p. 509. Plenum Press, New York (1994).
4. Arndt, U.W., Crowther, R.A. & Mallet, J.F.W. A computer-linked cathode ray tube microdensitometer for X-ray crystallography. *J. Phys. E: Sci. Instr.* **1**, 510–516 (1968).
5. Martin, P.R. & Hausinger, R.P. Site-directed mutagenesis of the active site cysteine in *Klebsiella aerogenes* urease. *J. Biol. Chem.* **267**, 20024–20027 (1992).
6. Jabri, E., Lee, M.H., Hausinger, R.P. & Karplus, P.A. Crystallization of urease. *J. Mol. Biol.* **227**, 934–937 (1992).
7. Jabri, E., Carr, M.B., Hausinger, R.P. & Karplus, P.A. Crystal structure of urease. *Science* **268**, 998–1004 (1995).
8. Hamlin, R.C. Multiwire area X-ray diffractometers. *Meths Enzymol.* **114**, 416–451 (1985).
9. Otwinowski, Z. (1993) "Oscillation Data Reduction program", in *Proceedings of the CCP4 Study Weekend: Data Collection and Processing*. 29–30 January 1993, compiled by L. Sawyer, N. Isaacs and S. Bailey, SERC Daresbury Laboratory, England, pp.56–62.
10. M.S. Weiss (1996) personal communication.
11. D. Gewirth (1995) The HKL manual. A description of the programs Denzo, XDisplayF, Scalepack. Edition 4, Yale University New Haven Connecticut. p. 94.
12. Howard, A.J., Nielsen, C. & Xuong, N.H. Software for a diffractometer with multiwire area detector. *Meths. Enzymol.* **114**, 452–472 (1985).
13. Mylvaganam, S.E., Bonaventura, C., Bonaventura, J. & Getzoff, E.D. Structural basis for the Root effect in haemoglobin. *Nature Struct. Biol.* **3**, 275–283 (1996).
14. Shrive, A.K. *et al*. Three dimensional structure of human C-reactive protein. *Nature Struct. Biol.* **3**, 346–353 (1996).
15. Kleywegt, G.T., Hoier, H. & Jones, T.A. A revaluation of the crystal structure of chloromuconate cycloisomserase. *Acts Crystallogr.* D**52**, 858–863 (1996).
16. Desmyter, A. *et al*. Crystal structure of a camel single-domain VH antibody fragment in complex with lysozyme. *Nature Struct. Biol.* **3**, 803–811 (1996).
17. Jabri, E. & Karplus, P.A. Structures of the *Klebsiella aerogenes* urease apoenzyme and two active site mutants. *Biochemistry* **35**, 10616–10626 (1996).
18. Glover, I.D. *et al*. Structure determination of OppA at 2.3 Å resolution using multiple-wavelength anomalous dispersion methods. *Acta Crystallogr.* D**51**, 39–47 (1995).
19. Van Duyne, G.D., Standaert, R.F., Karplus, P.A., Schreiber, S.L. & Clardy, J.C. Atomic structures of the human immunophilin FKBP-12 complexes with FK-506 and rapamycin. *J. Mol. Biol.* **229**, (1993).
20. Bruns, C.M. & Karplus, P.A. Refined crystal structure of spinach ferredoxin reductase at 1.7 Å resolution: oxidized, reduced and 2-phospho-5'-AMP bound states. *J. Mol. Biol.* **247**, 125–145 (1995).
21. Gonschorek, W. Intensities, structure factors and their variances. *Acta Crystallogr.* A**41**, 189–195 (1985).
22. Stenkamp, R.E. & Jensen, L.H. Resolution revisited: limit of detail in electron density maps. *Acta Crystallogr.* A**40**, 251– 256 (1984).
23. Brünger, A.T. The free *R*-value: a more objective statistic for crystallography. *Meths Enzymol.* in the press **(AUTHOR: STATUS?)**.
24. Kleywegt, G.J. & Brünger, A.T. Checking your imagination: applications of the free *R* value. *Structure* **4**, 897–904 (1996).
25. Diederichs, K. & Schulz, G.E. The refined structure of the complex between adenylate kinase from beef heart mitochondrial matrix and its substrate AMP at 1.85 Å resolution. *J. Mol. Biol.* **217**, 541–549 (1991).
26. Boisvert, D.C., Wang, J., Otwinowski, Z., Horwich, A.L. & Sigler, P.B. The 2.4 Å crystal structure of the bacterial chaperonin GroEL complexed with ATPgS. *Nature Struct. Biol.* **3**, 170–177 (1996).
27. Canady, M.A., Larson, S.B., Day, J. & McPherson, A. Crystal structure of turnip yellow mosaic virus. *Nature Struct. Biol.* **3**, 771–781 (1996).
28. Garcia, K.C. *et al*. An αβ T cell receptor structure at 2.5 Å and its orientation in the TCR-MHC complex. *Science*. **274**, 209–219 (1996).
29. Kabsch, W. Evaluation of Single-Crystal X-ray Diffraction Data from a Position-Sensitive Detector. *J. Appl. Crystallogr.* **21**, 916–924 (1988).
30. Papoulis, A. Probabiltity, random variables and stochastic processes. International Student Edition, McGraw Hill, Kogakusha Ltd, Tokyo. p. 147 (1965)

Mathematical derivation of the dependance of R_{sym} on multiplicity

As the true intensity $\hat{I}_h$ known, but can be approximated by the average $\hat{I}_h$ of the $I_{h,i}$, the variance of the R_{sym} numerator is related to the multiplicity n_h of the contributing reflections *h*. This is shown as follows: reformulating $\Delta_{h,i} = I_{h,i} - \hat{I}_h$, the sum of whose absolute values constitute the numerator of the R_{sym} formula

$$\Delta_{h,i} = I_{h,i} - \frac{1}{n_h}\sum_{j}^{n_h} I_{h,j}$$

$$= I_{h,i} - \frac{1}{n_h}\left(I_{h,i} + \sum_{j,j\neq i}^{n_h} I_{h,j}\right)$$

we find

$$\Delta_{h,i} = \frac{n_h - 1}{n_h} I_{h,i} - \frac{1}{n_h}\sum_{j,j\neq i}^{n_h} I_{h,j}$$

As the $I_{h,i}$ and $I_{h,j}$ (jfii) are independent, the variance $s_{\Delta_h}^2$ of the $\Delta_{h,i}$ can be calculated as the sum of the variances of the two terms on the right-hand side of the last equation:

$$s_h^2 = \left(\frac{n_h - 1}{n_h}\right)^2 s_h^2 + (n_h - 1)\frac{1}{n_h^2} s_h^2$$

$$= \frac{n_h - 1}{n_h} s_h^2$$

where s_h^2, the sample variance of $I_{h,i}$, is defined as $s_h^2 = \frac{1}{n_h - 1}\sum_{j}^{n_h}(I_{h,j} - \hat{I}_h)^2$

However, the root mean square width of a Gaussian G(x) with zero mean is proportional[30] to its average width $<|x|>$

$$<x^2> = \frac{\pi}{2} <|x|>^2$$

Thus, as the $\Delta_{h,i}$ are assumed to be normally distributed around a mean of zero,

$$<|\Delta_h|> = \frac{s_{\Delta h}}{\sqrt{\pi/2}}$$

and therefore

(9)
$$<|\Delta_h|> = \frac{\sqrt{\frac{n_h - 1}{n_h}}}{\sqrt{\pi/2}} s_h$$

demonstrating that the contribution of each reflection *h* to the numerator of R_{sym} is proportional to a function of its multiplicity n_h. This is the reason why R_{sym} values for low average multiplicity are overly optimistic, and R_{sym} as a function of data collection progress (multiplicity) is bound to rise, even if only statistical errors are present.

Relation of R_{meas}, PCV and R_{mrgd-I} to the average I/s ratio of the measured and reduced data

Data reduction programs calculate the estimates of the variance $s_{h,i}^2$ for each $I_{h,i}$ from counting statistics and background level, and most report the <*I*/*s*> ratio as a function of resolution. From a statistical standpoint, the sample variance s_h^2 of the $I_{h,i}$ should be consistent with these $s_{h,i}^2$, if no systematic differences between observations exist[20]. In this case, it follows from (9) that both R_{meas} (Eq. 2) and PCV (Eq. 3) are related to the inverse of the average signal-to-noise ratio $<I/s_I>$ of the observations

$$R_{meas} = PCV/\sqrt{\pi/2} = \frac{1}{\sqrt{\pi/2}}\frac{<s>}{<I>} \cong \frac{1}{\sqrt{\pi/2}}\frac{1}{<I_{h,i}/\sigma_{I_{h,i}}>}$$

Likewise, R_{mrgd-I} is related to the average signal-to-noise ratio of the merged intensities $\hat{I}_h$,

$$R_{mrgd-I} \cong \frac{2}{\sqrt{\pi/2}}\frac{1}{<\hat{I}_h/\sigma_{\hat{I}_h}>} \quad \text{with} \quad \sigma_{\hat{I}_h} = \sqrt{\frac{1}{\sum_i \frac{1}{\sigma_{I_{h,i}}^2}}}$$

Compared to, $<\hat{I}_h/\sigma_{\hat{I}_h}>$, which rises according to a pure $\sqrt{n}$ law even if systematic errors are present, R_{mrgd} values after merging of non-isomorphous datasets will be worse than those after merging of datasets without such errors, because the averages of the separate subsets of data will be distinct.

erratum

Improved *R*-factors for diffraction data analysis in macromolecular crystallography

Kay Diedrichs and P. Andrew Karplus

Nature Structural Biology **4**, no. 4, 269–275 (1997).

The Rsym definition on p. 269 should read:

$$(1) \quad \mathrm{Rsym} = \frac{\sum_{h}\sum_{i}^{n_h} |\hat{I}h - Ih,i|}{\sum_{h}\sum_{i}^{n_h} Ih,i} \quad \text{with } \hat{I}_h = \frac{1}{n_h}\sum_{i}^{n_h} |\hat{I}h - Ih,i|$$

(thanks to Clemens Vonrhein for finding the missing summation sign)

The first sentences of the legend for Fig. 2 should read: "The behaviour of data quality indicators as a function of multiplicity. Solid lines are used for the following three indicators: R_{sym} (filled triangle); PCV (filled square); $R_{mrgd\text{-}F}$ (open square). Dashed lines are used for R_{meas} (X); $R_{mrgd\text{-}I}$ (open circle); $<1/\sigma_I>$ (closed circle)."

In Fig. 2, it should be $<1/\sigma_I>$ (*not* $<1/s>$).

On p. 272, there are only space groups of the form $P4_{(0,1,2,3)}2_{(0,1)}2$ (*not* $P4_{(0,1,2,3)}2_{(0,1)}2_{(0,1)}$).

On p. 275 (box), several typesetting errors make understanding difficult:

line 2: "true intensity $\tilde{I}_h$" (*note the tilde*). As the true intensity $\tilde{I}_h$ is not known, ... ← another error !

lines 9–12: "As the $I_{h,i}$ and $I_{h,j}$ ($j \neq i$) (*note: not jfii*) are independent, the variance $s_{\Delta_h}^2$ of the $\Delta_{h,i}$ can be calculated as the sum of the variances of the two terms on the right-hand side of the last equation:

$$s_{\Delta_h}^2 = \left(\frac{n_h - 1}{n_h}\right)^2 s_h^2 + (n_h - 1)\frac{1}{n_h^2} s_h^2 = \frac{n_h - 1}{n_h} s_h^2 \text{"} \quad (\textit{note: this defines } s_{\Delta_h}^2, \textit{ not } s_h^2)$$

The final paragraph on p. 275 should start as follows:

Relation of R_{meas}, PCV, and $R_{mrgd\text{-}I}$ to the average I/σ ratio of the measured and reduced data

Data reduction programs calculate the estimates of the variance of $\sigma_{h,i}^2$ for each $I_{h,i}$ from counting statistics and background level, and most report the $<I/\sigma>$ ratio as a funciton of resolution. From a statistical standpoint, the sample variance s_h^2 of the $I_{h,i}$ should be consistent with these $\sigma_{h,i}^2$, if no systematic differences between observations exist[20]. In this case, it follows from (9) that both Rmeas (eqn. 2) and PCV (eqn. 3) are related to the inverse of the average signal-to-noise ratio $<I/\sigma_I>$ of the observations. [Note the *s* versus σ confusion in the article as printed]

Update: Reference 10 is Weiss, M.S. & Hilgenfeld, R. *J. Appl. Cryst.* 30, 203–205 (1997).