Crystal structure: How to use data quality Indicators to select a high resolution cutoff

M.AIT MOUHA*1; D.TLAMSAMANI1; K.YAMNI1
1 EMaMePS, Department of Chemistry, Faculty of Sciences, Moulay Ismail University, Meknes ,Morocco.

* Corresponding author: Tel: +212674031273, Email:aitmouha.m@gmail.com

Abstract: To get a good quality crystal structure, we should first do the initial analysis of the diffraction data. There are many global indicators of the quality of the diffraction data. Accentuated in this paper are the d_{min}, d_{eff}, C, R_{merge} and R_{r.i.m}. Based on the results obtained, this paper will overview in our understanding of how to use these indicators to select a high resolution cutoff that leads to the best crystal structure, using practical examples.

1. Notation

- $d_{\text{min}}$: the nominal resolution
- $d_{\text{eff}}$: the effective resolution
- $R_{\text{merge}}$: the merging factor
- $R_{\text{r.i.m}}$: the redundancy independent merging factor
- $\frac{I}{\sigma(I)}$: Signal-to-noise ratio
- $C$: the completeness
- $N_{\text{refl}}$: the number of reflections
- $N$: Redundancy.
- $CC^*$: the correlation coefficients
- $\text{Obs/param}$: N of reflections used/N of Isq parameters
- $wR$: The weighted R-factor based on $F^2$
- $R$: The unweighted residual factor based on $F$
- $S$: is the goodness of fit

2. Introduction

A series of operations are carried out on the data, which is normally referred to as ‘data reduction’ following integration of the spots on a set of X-ray diffraction images to produce a list of reflection intensities. Many papers (e.g. [1] [2]) have well documented algorithms and methods for data reduction. The findings will not be reiterated in this paper. We can analyze the quality of diffraction data when we have a set of scale factors to put all the intensities on a common scale and improved estimates of the error on the intensity.

The importance of confirming that the highest quality data set of a given crystal is obtained cannot be overemphasized, particularly in light of the fact that the diffraction experiment is the final real experimental step in the course of a structure determination. Many papers and comparative studies on the evaluation of the quality of the shaping diffraction data have been published [3] [4] [5]. They comparatively examine the evaluation of the quality of crystal structure.

This paper will give and describe an outline of the global indicators of the quality of diffraction data that have been described to date will be given. By using practical examples, this paper will also study their strengths as well as their usefulness to select a high-resolution cutoff.
3. Indicators of data quality

Taking for granted that the data reduction and the data collection strategy were carried out in a significant and practical way.

Several parameters, such as \( d_{\text{min}} \), \( C \), \( R_{\text{merge}} \), \( R_{\text{i.m}} \) and \( I/\sigma(I) \) can be used to probe the internal data quality.

3-1 The resolution of the data

The first essential indicator is the nominal resolution \( d_{\text{min}} \) of the data. This latter is defined as ‘the minimum distance at which two features in the corresponding electron density map can be resolved’ [5]. It is also defined as \( d_{\text{min}} \) as the resolution of the X-ray diffraction pattern where, following the Bragg equation: \( d_{\text{min}} = \frac{\lambda \sin \theta_{\text{max}}}{2} \), \( \theta_{\text{max}} \) being the maximum value of \( \theta \) for reflections contained within the measured data set. Atomic resolution parallels with a \( d_{\text{min}} \) of about 0.8 Å, for copper radiation.

Generally, a smaller \( d_{\text{min}} \) shows that smaller structural details are noticeable in the maps. For many crystallographic procedures, such as direct phasing, the resolution of the data set may be a restrictive parameter; Giacovazzo indicates that \( d_{\text{min}} \approx 1.4 \) Å is the resolution threshold below which a small molecule structure cannot be deciphered ab initio[6]; consequently, its precise measurement is important.

In some cases, an overestimation of the nominal resolution can also be identified based on the appearance of the Wilson [7] plot of the data set. In particular, a clear deviation of the plot from linearity at the nominal resolution side is often a good indication.

3-2 The completeness of data

An additional factor, the second one, is the completeness of the data set. In reality, the established measure \( d_{\text{min}} \) may change after the addition or removal of a few reflections. This measure does not reveal the incompleteness of the data set.

Weiss [5] proposed portraying the ‘effective resolution’ based on the nominal resolution \( d_{\text{min}} \) and the cube root of the completeness of the data set \( C \): \( d_{\text{eff}} = d_{\text{min}} C^{-1/3} \), with \( C = (N_{\text{ind}} / N_{\text{refl}}) \). This relation is attributed to the fact that missing reflections result in a deterioration of the final model parameters much in the same way as reduced resolution does [8].

It is possible to guess the number of reciprocal lattice points within the range \( \theta \) to \( \theta_{\text{max}} \) for a given crystal, the number of reflections \( N_{\text{refl}} \) is given by, \( N_{\text{refl}} = 33,510 V c \sin^{3} \theta_{\text{max}} / \lambda^{3} \), the number of reflections present \( N_{\text{ind}} \) may not be congruent with the expected number \( N_{\text{refl}} \) from the nominal value of \( d_{\text{min}} \).

If the number of reflection essentially measured is considerably lower than the expected value, it can be anticipated that the resolution of the analysis will be correspondingly lower although it is not easy to estimate by how much. For example, a nominal \( d_{\text{min}} = 1.4 \) Å data set which is only 70% complete will effectually be a 1.6 Å data set.

To collect additional complete diffraction data sets, a positive side effect of a requirement for authors to report the efficient resolution of their data set instead of the nominal resolution might likewise be an increased motivation. Indicators of accumulated statistical properties are needed since the number of reflections in a crystallographic experiment is high. For a long period of time, the ‘merging’ R value has been used to measure the consistent regularity of measurements.

3-3 The internal consistency

The merging statistics offer a valuable source of information about the quality of a data set prior to merging equivalent reflections. The most regularly reported descriptor of data quality is the usual
merging R factor \( R_{\text{merge}} = \sum_h \sum_l |I_{hl} - \langle I_h \rangle|/\sum_h \sum_l \langle I_h \rangle \) which measures the spread of \( n \) independent measurements of the intensity of a reflection, \( I_{(hkl)} \), around their average, \( \bar{I}_{(hkl)} \). These R factors cannot be paralleled with the R factors in refinement (\( R_{\text{work}} \) or \( R_{\text{free}} \)), the \( R_{\text{merge}} \) is fundamentally dependent on the redundancy of the data [5]. Low redundancy will always generate a lower \( R_{\text{merge}} \), but at the same time lead to less exact data. Diederichs & Karplus [3], show that to give a result that is independent of the average multiplicity \( N \), each term of the numerator has to be adjusted by a factor of \( \sqrt{N/(N-1)} \). The resulting quantity is called \( R_{\text{meas}} \) (or \( R_{\text{r.i.m}} \); [5]). These merging R factor (\( R_{\text{merge}} \) or better \( R_{\text{r.i.m}} \)) are not suitable measures for setting a resolution cutoff [4]. \( R_{\text{r.i.m}} = \sum_h \sum_l |I_{hl} - \langle I_h \rangle| \sqrt{\left( \frac{N}{N-1} \right)}/\sum_h \sum_l \langle I_h \rangle \) Subsequently, it was indicated that by introducing an additional factor of \( (1/N)^{1/2} \) into each term of the \( R_{\text{meas}} \) numerator, the quality of the merged data could also be expected [5], as this explains the expected increase inaccuracy related with averaging N measurements. The resulting quantity is called \( R_{\text{p.i.m}} \). (The precision-indicating merging R; [5]) which defines the precision of the averaged measurement.

\[ R_{\text{p.i.m}} = \sum_h \sum_l |I_{hl} - \langle I_h \rangle| \sqrt{\left( \frac{1}{N-1} \right)}/\sum_h \sum_l \langle I_h \rangle \] .The R factor should deliver the most information when it comes to guessing the performance of a given data set in structure determination, given the fact that the course of structure determination and refinement, averaged intensities or amplitudes are typically used.

The conventional criteria for examining the data resolution limit are the \( R_{\text{merge}} \) and \( R_{\text{r.i.m}} \) and \( R_{\text{p.i.m}} \). Yet, they are not adopted universally and are not required by journals that publish X-ray structures. Lately, (Karplus & Diederichs, [9]) we have advocated that the Pearson correlation coefficient of two ‘half’ data sets \( CC_{1/2} \) might be better appropriate than merging R factors for assessing and judging data quality.

Some of the advantages of a correlation constant are that it has a clear and accurate range: +1.0 for a good correlation and 0 for no correlation. \( CC_{1/2} \) is commonly close to 1 at low resolution and falls sharply to near zero at higher resolution as the intensities become weaker [10]

4. Analysis of data quality

The results for the practical examples tested are summarized in table 1

| Crystal data | Chemical formula | \( C_2 H_2 N_2 O \) |
|--------------|------------------|------------------|
| Pressure (K) | \( T \) | 347 \pm 2 |
| Space group | Crystal system | Monoclinic, \( P2_1/n \) |
| Temperature (K) | \( T \) | 293 |
| a, b, c (\text{Å}) | \( \beta \) | 12.2410 (4), 5.8386 (2), 25.2112 (9) |
| \( V \) (\text{Å}^3) | \( V \) | 1832.69 (11) |
| Z | 4 |
| Radiation type | Cu Kα |

| Crystal data | Chemical formula | \( C_2 H_2 N_2 O \) |
|--------------|------------------|------------------|
| Pressure (K) | \( T \) | 347 \pm 2 |
| Space group | Crystal system | Orthorhombic, \( P2_l 2_l 2_l \) |
| Temperature (K) | \( T \) | 293 |
| a, b, c (\text{Å}) | \( \beta \) | 9.7570 (3), 32.0691 (13), 5.8643 (2) |
| \( V \) (\text{Å}^3) | \( V \) | 1834.93 (11) |
| Z | 4 |
| Radiation type | Mo Kα |

Data 1: ot-Cu:Slassi et al.IUCrData (2019)  Data 2: ot-Mo:Slassi et al.IUCrData (2017)
An interesting feature was observed by comparing the two data sets. According to a conventional Signal-to-noise ratio, $I/\sigma(I)>2$ criterion:

- The nominal resolution $d_{\text{min}}$ of the tow data sets would be 0.84 and 0.7 Å respectively.

- The number of measured, independent and observed $|I|/2\sigma(I)$ reflections would be

| Data    | $N_{\text{meas}}$ | $N_{\text{ind}}$ | $N_{\text{obs}}$ |
|---------|-------------------|------------------|------------------|
| Data 1  | 5548              | 3487             | 3037             |
| Data 2  | 25352             | 3590             | 2978             |

- The completeness increased with higher resolution Figure 1. It has long been realized that any missing reflection, be it one that has not been measured leads to a deterioration of the structure parameters

![Figure 1](image1.png)

**Figure 1:** Plots from Excel of completeness $N_{\text{obs}}(hkl)$ and $N_{\text{theo}}(hkl)$ against resolution in Data1

- The Signal-to-noise ratio $I/\sigma(I)$ increased with lower resolution (Figure 2 and Figure 3), $I/\sigma(I)$ is a good criterion for resolution cutoff.

![Figure 2](image2.png) ![Figure 3](image3.png)

**Figure 2:** plot from Excel of $I/\sigma(I)$ against $d_{\text{min}}$ in Data1

**Figure 3:** plot from Excel of $I/\sigma(I)$ against $d_{\text{min}}$ in Data2

- The $R_{\text{merge}}$ or $R_{\text{sym}}$ value is generally increased with higher resolution and completeness, overall $R_{\text{sym}}$ values of <5% (Figure 4 for $d_{\text{min}}>0.9$), 5-10% (Figure 5 for $d_{\text{min}}>0.9$) are taken to indicate good, usable quality data [3]. $R_{\text{sym}}$ is commonly used to guide decisions during data reduction, such as determining to what resolution data are reliable.
5. Application of resolution cutoff

How can we decide where to apply a resolution cutoff? To examine these questions, a number of tests were carried out using two data. By comparing refinement against measured data with refinement against simulated data, we can judge the resolution point at which the measured data become no better than simulated data and compare this with the data-processing statistics.

The crystal structure of Data2 was then refined to 0.7 Å resolution against the observed data, R values (and wR) for the experimental data increase with increasing resolution using all data;

| N_{ind} | N_{obs} | R    | wR | S   | obs/para |
|---------|---------|------|----|-----|----------|
| Data2   | 3590    | 2978 | 0.037 | 0.10 | 1.04 | 12.5 |
| Cutoff  | 1895    | 1632 | 0.035 | 0.08 | 1.04 | 7 |

A better structure has a lower R values and better data allow obtaining a better structure as shown in Table2. Similar results were obtained with example Data1, as shown in Table3;

| N_{ind} | N_{obs} | R    | wR | S   | obs/para |
|---------|---------|------|----|-----|----------|
| Data1   | 3487    | 3037 | 0.084 | 0.245 | 1.08 | 12.8 |
| Cutoff  | 500     | 454  | 0.05  | 0.15  | 1.08 | 2 |

Choice of high-resolution cutoff is better because it gives lower R-values and better structure.

Figure4: plot from Excel of R_{merge} against d_{min} in Data2

Figure5: plot from Excel of R_{merge} against d_{min} in Data1

Figure6: The structure of crystal Data1

Figure7: The structure of crystal Data2
6. Conclusions

The results can be summarized in the following statements:

- Merging many data sets usually improves the completeness and signal-to-noise ratio in the data; \( I/\sigma(I) \) is a good criterion for resolution cutoff.
- Choice of high-resolution cutoff is better because it gives lower R-values.
- Structure quality depends on data quality.

7. Keywords

Diffraction data, crystal structure, resolution, global indicators, R values, cutoff.

8. References

[1] Fox, G. C. & Holmes, K. C. 1966. *Acta Cryst.* **20**, 886–891.
[2] Kabsch, W. 2010. *Acta Cryst.* D **66**, 133–144.
[3] Diederichs, K. & Karplus, P. A. 1997. *Nature Struct. Biol.* **4**, 269–275.
[4] Diederichs, K. & Karplus, P. A. 2015. *Current Opinion in Structural Biology*. **34**, 60–68.
[5] Weiss, M. S. 2001. *Appl. Cryst.* **34**, 130–135.
[6] Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Siliqi, D. 2005a. *Acta Cryst.* D **61**, 1080–87.
[7] Wilson, A. J. C. 1949. *Acta Cryst.* **2**, 318–321.
[8] Evans 2011. *Acta Cryst.* D **67**, 282–292.
[9] Karplus, P. A. & Diederichs, K. 2012. *Science*. **336**, 1030–33.
[10] Evans & Murshudov 2013. *Acta Cryst.* D **69**, 1204–14.