Experimental Strategies for Better Diffraction Data

Michael Wedel and Horst Borrmann

Introduction

Precise determination of crystal structures and, in particular, experimental charge densities with x-ray diffraction methods requires data sets of extremely good quality, up to very high resolution. Instrumentation has clearly improved in course of the last decades, introducing high-precision area detectors, more powerful x-ray sources and optics, just to mention a few assets. Likewise, the availability of more powerful computers at much lower prices has lead to a large variety of data reduction software which greatly reduces the time used for these tasks. All of these developments have contributed to high resolution diffraction experiments being carried out in a matter of days instead of weeks or even months, without the need to travel to a synchrotron facility.

In order to be able to evaluate if advanced instrumentation indeed provides for vastly improved data quality, two main questions need to be tackled: "What characterizes an excellent data set?" and "How can such a data set be collected?"

Overview

Generally the goal is to provide a quantitative measure of data quality not based on any comparison with structural models. Most methods build on statistical analysis of symmetry-equivalent reflections finally reflected in a quality indicator called R_{merge} or R_{int} [1–3].

$$R_{\mathrm{merge}} = \frac{\sum_{hkl}\sum_{i}|I_i(hkl) - \overline{I(hkl)}|}{\sum_{hkl}\sum_{i} I_i(hkl)}$$

As a main drawback R_{merge} heavily depends on redundancy in the data set [1]. Even worse, higher redundancy results in higher values of R_{merge} and thus somehow contradicts "scientific intuition". Of course, this behavior is highly undesirable and modified indicators were proposed, e.g., an R factor which is independent of redundancy, called $R_{\mathrm{r.i.m.}}$ [3] or R_{meas} [1],

$$R_{\mathrm{r.i.m.}} = \frac{\sum_{hkl}\sqrt{\frac{N}{N-1}}\sum_{i}|I_i(hkl) - \overline{I(hkl)}|}{\sum_{hkl}\sum_{i} I_i(hkl)}$$

where N is the redundancy, which in this case means the number of times a reflection *hkl* has been measured. Another R factor that needs to be mentioned is $R_{\mathrm{p.i.m.}}$ [3], as it incorporates redundancy in a way more appropriate, since the value gets *lower*, as redundancy gets *higher*:

$$R_{\mathrm{p.i.m.}} = \frac{\sum_{hkl}\sqrt{\frac{1}{N-1}}\sum_{i}|I_i(hkl) - \overline{I(hkl)}|}{\sum_{hkl}\sum_{i} I_i(hkl)}$$

Clearly *redundancy* has to be considered most important. Commonly, this defines how often a certain reflection appears in a data set including symmetry equivalents and multiple measurements of the same reflection. Remeasuring a particular Bragg reflection in the very same position gives a good indication about reproducibility and stability of the experiment. However, recollecting individual reflections at different positions (orientations) gets additional information included, very similar to the situation with reflections equivalent by symmetry. Accordingly multiple measurements are widely used to account for systematic "errors" like absorption or anisotropic extinction. Along these arguments the terms "real redundancy", or "multiplicity of observation" (MoO [4]) have been defined to emphasize these particular conditions in a diffraction experiment.

Generally, the best quality indicator for a single measurement is its signal-to-noise ratio $I/\sigma(I)$, which reflects how well the signal is detectable above background. Any indicator based on $I/\sigma(I)$ should directly reflect data quality without any additional contributions from symmetry, redundancy and alike. A proper value is defined as R_σ (for example in SHELX [5]), however, it is not always used in this strict way:

$$R_\sigma = \frac{\sum_{i}\sigma(I_i)}{\sum_{i} I_i}$$

Table 1: The data of a cubic calcium hexaboride crystal collected with Mo Kα radiation were merged (with SORTAV [6]) in Laue groups $m\bar{3}m$ and $\bar{1}$ for comparison. Q closely resembles $I/\sigma(I)$ ($Q = \bar{I}/\sigma(I) \cdot \sqrt{N}$). Obviously very significant improvement for high redundancy.

d [Å]	N	N_{ind}		N/N_{ind}		R_{merge}		Q		$R_{\mathrm{p.i.m.}}$	
		$m\bar{3}m$	$\bar{1}$	$m\bar{3}m$	$\bar{1}$	$m\bar{3}m$	$\bar{1}$	$m\bar{3}m$	$\bar{1}$	$m\bar{3}m$	$\bar{1}$
d > 1.090	398	14	125	28.4	3.2	0.0154	0.0147	230.72	81.95	0.0031	0.0078
1.090 > d > 0.865	320	9	105	35.6	3.0	0.0132	0.0116	276.80	77.87	0.0023	0.0067
0.865 > d > 0.756	359	10	133	35.9	2.7	0.0190	0.0185	168.13	44.46	0.0029	0.0108
0.756 > d > 0.687	205	8	83	25.6	2.5	0.0196	0.0202	179.81	56.51	0.0040	0.0117
0.687 > d > 0.637	175	8	98	21.9	1.8	0.0261	0.0253	107.58	31.30	0.0056	0.0164
0.637 > d > 0.600	91	5	54	18.2	1.7	0.0218	0.0225	87.18	29.63	0.0048	0.0152
0.600 > d > 0.570	134	11	91	12.2	1.5	0.0231	0.0202	76.48	29.93	0.0064	0.0143
0.570 > d > 0.545	98	7	65	14.0	1.5	0.0229	0.0221	84.07	28.88	0.0063	0.0153
0.545 > d > 0.524	101	6	65	16.8	1.6	0.0384	0.0368	49.85	14.82	0.0095	0.0260
0.524 > d > 0.506	96	8	66	12.0	1.5	0.0352	0.0325	41.71	16.87	0.0095	0.0228
0.506 > d > 0.490	75	5	52	15.0	1.4	0.0321	0.0324	60.94	20.75	0.0079	0.0229
0.490 > d > 0.476	115	10	75	11.5	1.5	0.0375	0.0375	48.22	18.60	0.0111	0.0262
0.476 > d > 0.464	63	4	44	15.8	1.4	0.0387	0.0372	53.40	16.37	0.0098	0.0263
0.464 > d > 0.452	96	9	70	10.7	1.4	0.0525	0.0508	28.62	11.55	0.0157	0.0359
0.452 > d > 0.442	62	6	47	10.3	1.3	0.0623	0.0357	25.62	16.31	0.0187	0.0252
0.442 > d > 0.433	110	8	81	13.8	1.4	0.0652	0.0702	37.43	11.74	0.0175	0.0495
0.433 > d > 0.424	59	4	41	14.8	1.4	0.0438	0.0314	53.39	19.74	0.0109	0.0222
0.424 > d > 0.416	68	8	52	8.5	1.3	0.1047	0.0861	16.33	6.76	0.0359	0.0609
0.416 > d > 0.408	80	8	62	10.0	1.3	0.0871	0.0823	23.67	11.14	0.0286	0.0582
0.408 > d > 0.402	105	8	81	13.1	1.3	0.1055	0.1000	24.67	9.42	0.0287	0.0707

In contrast to the revised R factors as cited above, this value is derived by many data processing programs, although proper determination of $\sigma(I)$ is a general problem with area detectors since there is no counting statistics accessible. Nevertheless, despite a certain systematic deviation, R_σ is still useful in comparing the quality of data sets. Finally a data set with high $I/\sigma(I)$ but respectively low R_σ along with low $R_{\mathrm{r.i.m.}}$ and $R_{\mathrm{p.i.m.}}$ values we consider to be of truly high quality (Table 1).

Data Collection Strategies

Control software of practically any modern diffractometer is bundled with a routine to determine a certain sequence in data collection. Such programs are typically called strategy while representing a typical case of the well-known *travelling salesman problem*. Based on instrument parameters and predetermined crystal parameters an optimized sequence for data collection is determined. Due to the complexity of the problem, typically desired completeness and/or redundancy may usually be defined only rather crudely. Even with powerful modern computers the procedure quickly becomes quite time-consuming. Of course, cheap computer power has allowed for more advanced and new approaches which often tackle specific requirements of particular experiments. E.g. BEST [7, 8] takes estimations on radiation damage into account, or STRATEGY [9] suggests the shortest possible scan to reach a complete data set on a diffractometer with only one rotation axis. Other examples are RSPACE [10], LATTICEPATCH [11] or STRAT [12]. Development on this topic is still going on as indicated by most recent applications [13].

Own Developments

The aforementioned arguments provided the basis to develop a dedicated strategy software tool. Our main goal is to obtain data sets of utmost quality for charge density analysis, mainly on intermetallic compounds.

In highly symmetric space groups multiplicity of general reflections is vastly larger than for zonal or even serial reflections. Accordingly, special emphasis is put on high redundancy of such reflections. At the same time high data quality should be maintained up to very high resolution, e.g. $\sin\theta/\lambda \leq 1.75\,\text{Å}^{-1}$. Clearly, careful measurement of individual reflections along with high redundancy up to highest resolution using an instrument in perfect shape is expected to provide excellent data. Of course, data collection should run in an efficient way and one may have to take care about special

conditions due to crystal shape, sample environment and alike. In total an entire set of reflections based on the respective orientation matrix is generated, these reflections are arranged in a set of lists before these lists are sorted to derive a sequence of scans which fulfill all requirements [14].

A large variety of diffraction experiments and associated instrumentation calls for extensibility and flexibility as a major part in the concept. Basically the program comprises a set of layers with all actual calculations and algorithms handled in the logic layer. The user interacts with the interface layer, which is linked to the logic layer through a management layer in between. The graphical interface offers precise control over the granularity of all calculation via construction of complex graphs.

Since the program uses runtime loadable plug-ins that do not depend on each other, it is very easy to add new features. This is especially useful to get own algorithms included, e.g. to select a strategy or to adapt to completely new requirements.

Independence from specific experimental setups is achieved by providing a way to enter characteristics of the diffractometer, such as goniometer axes, detector, wavelength and other parameters into the program. For internal calculations all involved matrices and directions are converted according to a common specification. Crystal data may be imported via different file formats, too.

Currently the program allows for basic strategy searches using the simulated annealing algorithm [15]. It randomly generates sets of scan parameters, performs a complete simulation of the corresponding diffraction experiment and calculates a score for the resulting data set. Via iteration, the score is minimized. Since the origin of the score is not relevant to the algorithm, the user may define it according to own needs. Speed and memory consumption of the algorithm are limited mainly by the number of simulated reflections and the number of required scans.

Outlook

In the future different methods of calculating scores for the existing algorithm will be tried out, as well as additional searching algorithms. The results will be compared to those from programs supplied by various manufacturers of instruments. Ultimate goal is always to derive reliable criteria for the collection of high quality data sets, which allow for more detailed structure analysis and even deeper insights.

References

[1] *K. Diederichs and P. A. Karplus*, Nat. Struct. Mol. Biol. **4** (1997) 269.

[2] *M. S. Weiss and R. Hilgenfeld*, J. Appl. Crystallogr. **30** (1997) 203.

[3] *M. S. Weiss*, J. Appl. Crystallogr. **34** (2001) 130.

[4] *P. Müller, R. Herbst-Irmer, A. L. Spek et al.*, Crystal Structure Refinement: A Crystallographer's Guide to SHELXL, vol. 8 of *International Union of Crystallography Monographs on Crystallography*, Oxford University Press, Oxford (2006).

[5] *G. M. Sheldrick*, Acta Crystallogr. **A64** (2008) 112.

[6] *R. H. Blessing*, Crystallogr. Rev. **1** (1987) 3.

[7] *A. N. Popov and G. P. Bourenkov*, Acta Crystallogr. **D59** (2003) 1145.

[8] *G. P. Bourenkov and A. N. Popov*, Acta Crystallogr. **D62** (2006) 58.

[9] *R. B. G. Ravelli, R. M. Sweet, J. M. Skinner, A. J. M. Duisenberg, and J. Kroon*, J. Appl. Crystallogr. **30** (1997) 551.

[10] *M. R. Harris, M. Fitzgibbon, and F. Hage*, J. Appl. Crystallogr. **22** (1989) 624.

[11] *A. L. Klinger and R. H. Kretsinger*, J. Appl. Crystallogr. **22** (1989) 287.

[12] *X.-J. Zhang and B. W. Matthews*, J. Appl. Crystallogr. **26** (1993) 457.

[13] *J. Zikovsky, P. F. Peterson, X. P. Wang, M. Frost, and C. Hoffmann*, J. Appl. Crystallogr. **44** (2011) 418.

[14] *Nokia*, Qt 4, http://qt.nokia.com/, We decided to write the program in C++, using the Qt library as the basis for the GUI, but also for other modules. With the QtConcurrent module it provides convenient support to parallelize many operations commonly used when working with lists.

[15] *S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi*, Science **220** (1983) 671.