# Advances in Accuracy and Automation of Data Collection and Processing

Wladek Minor
Department of Molecular Physiology and Biological Physics, University of Virginia,
Charlottesville, VA 22908
*wladek@iwonka.med.virginia.edu*

Zbyszek Otwinowski
Department of Biochemistry, Southwestern Medical Center, University of Texas,
5323 Harry Hines Blvd., Dallas, TX 55235
*zbyszek@mix.swmed.edu*

## Abstract

*Macromolecular Crystallography is expanding rapidly, with respect to the number of molecules studied and their significance to biology and medicine. This expansion is enabled by faster acquisition of high quality diffraction data, the development of efficient X-ray detectors and the expanded availability of high brilliance X-ray beam lines at synchrotron radiation facilities.*

## 1 Introduction

Macromolecular crystallographers apply the technique of X-ray crystallography to the study of the structure and function of proteins and viruses, as well as other biological macromolecules and assemblies. Among other achievements, these studies often elucidate molecular origins of disease and provide a molecular basis for the design of therapeutic agents.

At present, this field is expanding rapidly, with respect to the number of molecules studied and their significance to biology and medicine. This expansion is enabled by faster acquisition of high quality diffraction data, the development of efficient X-ray detectors and the expanded availability of high brilliance X-ray beam lines at synchrotron radiation facilities. Despite the great progress in protein crystallography, the enormous success of genetics during the last decade has overwhelmed biological scientists trained in macromolecular crystallography. Knowledge of the gene product (protein, RNA) 3-d structure is central to understanding of chemical interactions responsible for biological functions of the gene. Genes and their products are identified at much higher rate than new 3-d structures are solved. Rapid progress in genetics has produced an enormous backlog of structures that would be interesting to determine.

The difficulty originates in the low diffracting power of macromolecular crystals. The ability to collect and process data from low quality, disordered, and weakly diffracted crystals may provide the largest speed-up factor. The current construction of a new generation of synchrotrons such as the APS and ALS will provide strong X-ray sources but in the same time will increase the need to speed-up not only the data collection but the whole process which leads to the structure solution.

Synchrotron radiation allows the measurement of small, but important, diffraction amplitudes. Most often, the value of interest is a small difference between large observed X-ray diffraction intensities. The main need for precise measurements comes from phasing of macromolecular crystals, where only difference measurements are important. A case of enormous scientific significance is the measurement of diffraction phase by the Multiwavelength Anomalous Dispersion (MAD) method [1-6]. To make MAD and related methods applicable to the majority of protein crystals it is necessary to **dramatically increase the precision of diffraction experiment.** Recent advances in the reduction or elimination of radiation sensitivity of biological samples by the rapid freezing method have made it possible to expose crystals at the synchrotron to X-ray doses up to 1000 times larger than it is practical with laboratory sources. Such a large increase in beam dose should, by the laws of statistics, make it possible to measure a signal up to 30 times smaller. Practice shows large improvement in data quality while conducting experiments at the synchrotron, but it is still far from what is theoretically possible**.** The precision of today's experiments is limited by inadequate detectors, inadequate experimental procedures and inadequate processing techniques.

## 2 Cryo-Crystallography

Proteins, peptides, nucleic acids and virus crystals can be flash cooled to cryogenic temperatures [7-9]. The principal advantage of this treatment is in the virtual elimination of radiation damage. Frequently an improvement in microscopic crystal order is observed resulting in stronger high resolution diffraction. Crystals can also be frozen at the optimal time during their growth, which is especially significant if one has to wait for synchrotron time.

Flash freezing, in most cases, increases the mosaicity, the macroscopic crystal disorder. Perhaps

surprisingly, a moderate increase in mosaicity makes precise MAD experiments much easier. The reflection intensity is proportional to the average beam intensity over the range of angles for which the reflection is in the Bragg condition. The larger the reflection width, the more fluctuations of beam intensity are averaged out. The fact that precision of X-ray diffraction improves by increasing mosaicity (to about 0.5°) and decreasing extinction is well known in small molecule crystallography. The high precision experiments can only be done in practice with flash frozen crystals. It seems that different cryo-protectants (which prevent formation of ice crystals) can be introduced into most macromolecular crystals. In leading laboratories many crystals are successfully frozen at the first attempt. The technique is becoming increasingly popular and some laboratories are employing it in the majority of their work.

## 3 New high intensity X-ray sources

Synchrotron sources have evolved towards dedicated (low emittance) high energy storage rings with an insertion device producing X-rays. The current wiggler beamline at NSLS is about 1000 times stronger than the strongest laboratory sources. APS and ALS beamlines are expected to be another factor of 10 - 100 stronger. For the purpose of the proposed experiments the stable flux through 0.3 mm pinhole is the most important beam parameter.

Radiation damage to the frozen crystal limits the total exposure to about $5*10^{16}$ photons / $mm^2$ [43, 44]. This dose corresponds to one day of wiggler time at NSLS or about an hour at ALS and APS. Although very intense beam potentially allows for fast data collection (for example in seconds), acquiring highly precise data will require large increase in X-ray dose and the corresponding increase in data collection time. For this reason very precise data collection should be done only one order of magnitude faster than it is done in laboratory. Data collection at such rate can be accomplished by evolutionary changes in today's methods of data acquisition and processing. Wiggler and undulator beamlines, particularly at APS and ALS, will be capable of doing several experiments in a day. Such a rate requires efficient data analysis methods in order for the work to proceed smoothly.

## 4 Detectors

Area detectors have been used from the very beginning of X-ray diffraction studies in the year 1912. However, the detector technology has evolved since that time and now includes, apart from x-ray film, electronic and IP (phospholuminescent, best known by trade name Image Plate) area-sensitive detectors. Crystallographic detectors measure X-ray flux simultaneously at a large number (millions) of pixels. The major requirements from the detector for very precise data collection are:

- high quantum efficiency,
- high saturation,
- stability.

Charge Coupled Device (CCD) and Image Plate (IP) detectors satisfy these requirements but require corrections to be applied to each pixel to achieve high accuracy. The main corrections are for geometric distortion, background (dark current) and sensitivity. Coupling of distortion and sensitivity corrections limits the precision of current calibration methods of CCD's to about 2-3%. Calibration of IP scanners does not have problem with this coupling, however, some IP detectors (e.g. MAR, R-AXIS-II) vibrate during readout making independent pixel-by-pixel sensitivity correction impossible. The new calibration techniques methodology has to be based on a vibration free detector with a very large dynamic range (high saturation point), but not necessarily a very fast detector, as accumulation of a high number of scattered X-ray photons will take minutes rather than seconds of exposure

The 2D detectors and related software are now used predominantly to measure and integrate diffraction from single crystals of biological macromolecules. However, their usefulness in small-molecule, high-resolution crystallography is growing rapidly allowing fast solution of small molecule structures. For instance, newly developed Nonius detector has an potential to solve a 'typical' small molecule structure in 30 minutes including data collection, processing and actual structure solution.

## 5 Data Reduction

Several computer programs were developed to analyze single-crystal diffraction data. The analysis and reduction of a single crystal diffraction data consists of seven major steps. These are:

1) Visualization and preliminary analysis of the original, unprocessed, detector data.
2) Indexing of the diffraction pattern
3) Refinement of the crystal and detector parameters,
4) Integration of the diffraction maxima,
5) Finding the relative scale factors between measurements,
6) Precise refinement of crystal parameters using whole data set.
7) Merging and statistical analysis of the measurements related by space group symmetry.

Among the computer programs that were used widely for data reduction are *MOSFLM* and related programs [10-14], *XDS* [15-17], *OSC* [18-20], *DENZO* [21], *MADNES* [22,23], the San Diego programs [24], *XENGEN* and *X-GEN* [25] and others. For full list of

programs look into authors paper in newest edition of Methods in Enzymology. The theory behind the data reduction methods is complex enough that a series of European Economic Community workshops were dedicated to this task only [26,27]. The proceedings from these workshops contain a fairly complete presentation of the theory.

The authors of this review developed three programs: *DENZO* and *SCALEPACK* to integrate and scale the data and *XDISPLAYF* to analyze the process visually. Together, these programs form the *HKL* or the *MAC-DENZO* software suite. The programs can estimate Bragg intensities from single-crystal diffraction data that are recorded on two-dimensional, position-sensitive x-ray (also potentially neutron-diffraction or electron-diffraction) detector, for example film, IP scanners, or charge-coupled device (CCD) area detectors. The programs allow for data collection by oscillation, Weissenberg and precession methods. The detector can be either flat or cylindrical. The detector readout can be either a rectilinear or spiral, although spiral coordinates must be converted to rectilinear before processing. The programs allow for random changes in the position and the sensitivity of the detector between consecutive exposures. The programs *DENZO*, *XDISPLAYF* and *SCALEPACK* implement most of the ideas discussed at the EEC Cooperative Programming Workshop on Position Sensitive Detector Software [26,27]. In particular, the programs feature profile fitting, weighted refinement, eigenvalue filtering and universal definition of detector geometry.

### 5.1 Visualization of the diffraction space

A diffraction data set forms an image of three-dimensional reciprocal space. This image is formed by a series of two-dimensional diffraction images, each of them representing a different, curved, slice of reciprocal space. In order to accurately integrate the diffraction maxima, they must appear as separated (non-overlapping) spots in the individual 2-d images. Unless the data are collected by the precession method, the diffracted image contains a distorted view of reciprocal space. This distortion of the image is a function of the data collection method, the diffraction geometry, and the characteristics of the detector. For the data reduction to be successful, the distortion of reciprocal space as viewed by the detector has to be correctly accounted for by the program. The distortion of the image of reciprocal space can vary even between images collected on the same detector. This is because the position of the detector, the X-ray wavelength, the oscillation range, pixel size, scanner gain, and the exposure level all affect the detector representation of diffraction space.

If problems exist with the detector or other components of the data-collection system, the display option[28] helps to discover these before all the data are recorded. The examination of the image may reveal if there are extraneous sources of x-ray background. There are other statistics that can be provided instantly by *XDISPLAYF* which may indicate for example A/D converter malfunction.

### 5.2 Indexing

There is enormous literature regarding indexing of 2-D images[29-31]. *HKL* (*MAC-DENZO*) package offers two indexing methods: automatic and interactive. The automatic method, applicable in most cases, is fast and simple. The first step in the automatic method is the peak search, which chooses the spots to be used by the autoindexing subroutine. Ideally, the peaks should come from a diffraction by a single crystal. The *DENZO* program accepts peaks for autoindexing only from a single oscillation image. It is important that the oscillation range be small enough (it can even be zero, i.e. a still) so that the lunes (spots on one reciprocal plane, roughly perpendicular to the beam direction) formed by diffraction peaks are resolved. Otherwise reflections can have more than one index consistent with a particular position on the detector. On other hand, oscillation range should be large enough to have sufficient number of spots, for the program to be able to establish periodicity of the diffraction pattern. This may require at least 0.5 degree oscillation for a small unit cell protein crystal and 2-3 degree oscillation in the case of small molecule crystals.

The second step in the autoindexing is the mapping of the found diffraction maxima onto reciprocal space. Because the precise angles at which reflections diffract are a priori unknown for oscillation data, the center of the oscillation range is used as the best estimate of the angle at which the diffraction occurs.

The autoindexing in *DENZO* is based on a novel algorithm: a complete search of all possible indices of all reflections, found by peak search or manually selected. When the program finds values (integer numbers) of one index (for example, h) for all reflections, this is equivalent to finding one real space direction of the crystal axis (in this case, a). For this reason such indexing is called “real space indexing.” Finding one real space vector is logically equivalent to finding periodicity of reciprocal lattice in the direction of this vector. Search for real space vectors is performed by a Fast Fourier Transform (FFT) and takes advantage of the fact that finding all values of one index (e.g. h) for all reflections is independent of finding all values of another index (e.g. k). The *DENZO* implementation of this method is not dependent on prior knowledge of the crystal unit cell; however, for efficiency reasons, the search is restricted to reasonable range (obtained by default from requirement of spot separation) of unit cell lengths.

After the search for real space vectors is completed, the program finds the three linearly independent vectors, with minimal determinant (unit cell volume), that would index all (or more precisely almost all) of the observed peaks. These three vectors are unlikely to form a standard basis for a description of the unit cell. The process of finding a standard basis is called "cell reduction." The program finds the best cells for all of the fourteen Bravais lattices. The transformation of the primitive cell to a higher symmetry cell may require some distortion of the best triclinic lattice that fits the peak search list. Due to experimental errors the fit is never perfect for the correct crystal lattice. Sometimes the observed reflections can fit a higher symmetry lattice than one defined by space group symmetry. Such condition is called lattice (or metric tensor) pseudo-symmetry. If this happens the lattice determination and assignment of lattice symmetry may get complicated. The procedure in such case is to index the data in lowest symmetry lattice that does not introduce wrong lattice symmetry (triclinic lattice is always a safe choice) and look for symmetry of intensity pattern during scaling of symmetry related reflections. The program *DENZO* calculates the distortion index for all fourteen of the Bravais lattices. It is up to the user to define the lattice and space group symmetry, as the program at this stage of the calculation cannot distinguish lattice symmetry from pseudosymmetry.

## 5.3 Refinement of the crystal and detector parameters

The integration of reflections requires knowledge of their index and position. The weak reflections can only be found by prediction based on the information obtained from strong reflections. The autoindexing step provides only approximate orientation of the crystal and the result may be imprecise if the initial values of the detector parameters are poorly known. The least squares refinement process is used to improve the prediction .

The parameters describing measurement process have to be either known "a priori" or estimated, by manual or automatic refinement procedure, from diffraction data. Depending on the particulars of the experiment, the same parameters (e.g. crystal to detector distance) may be more precisely known "a priori" or are better estimated from the data. *DENZO* allows for the choice of the method for each of the parameters separately. This flexibility is handy under special circumstances; however, using it well requires considerable knowledge of diffraction experiments. Fortunately, the "fit all" option and detector specific default values seem to be reliable under most conditions.

The initial crystal and detector orientation parameters require refinement for each processed image. The refinement can be simple, for a series of images collected with an on-line detector, or more complexed, if the detector orientation is only crudely known and varies from image to image, as it is in the case of off-line scanners. The refinement is controlled by the user and can consist of several steps. Both detector and crystal parameters can be fitted simultaneously by the fast-converging least squares method. The refinement is done separately for each image to allow for the processing of data even when the crystal (or the detector) slips considerably during data collection. Occasionally the refinement can be unstable due to a high correlation among some parameters. High correlation makes possible for errors in one parameter to partially compensate errors in other parameters. If the compensation is 100%, the parameter would be undefined, but the error compensation by other parameter would make the predicted pattern correct. In such cases eigenvalue filtering (the same method as Singular Value Decomposition, described in Numerical Recipes[32]) is employed to remove the most correlated components from the refinement and make it numerically stable. Eigenvalue filtering works reliably when starting parameters are close to correct values but may fail to correct large errors in the input parameters, if the correlation is close to, but not exactly 100%. Once the whole data set is integrated, the global refinement (sometimes called postrefinement) [33,34] can refine crystal parameters (unit cell and orientation) more precisely and without correlation with detector parameters. Unit cell used in publications should come from global refinement (in *SCALEPACK*) and not from *DENZO* refinement.

A correct understanding of the detector geometry is essential to accurate positional refinement. Unfortunately, most detectors deviate from perfect flat or cylindrical geometry. These deviations are detector specific. The primary sources of error include misalignment of the detector position sensors (IP scanners), non-planarity of the film or IP during exposure or scanning, inaccuracy of the wire placement and distortions of the position readout in multi-wire proportional counters (MWPC), optical distortion (which can also be due to a magnetic field acting upon the image intensifier) in the TV or CCD based detectors. If the detector distortion can be parametrized, then these parameters should be added to the refinement.

With film and IP's handled manually in cassettes, the technique still used at many synchrotrons, the biggest problem lies in keeping the detector flat during exposure and subsequent scanning. In the manual systems, it is much harder to model the possible departures from ideal flat or cylindrical geometry, and *DENZO*, like most programs, makes limited attempts to correct for such distortions. Non-ideal film or IP geometry is one of the main factors behind the variable quality of data collected with the manual systems.

## 5.4 Integration of the Diffraction Maxima: Profile Fitting

The accurate prediction of spot positions is necessary to achieve a precise integration of Bragg peaks. The most important reason for accurate prediction of the spot positions arises from the application of profile fitting[35,36]. Profile fitting is a two step process. First, the profile is predicted based on the profiles of the other reflections within a chosen radius. The predicted profile in *DENZO* is an average of profiles shifted by the predicted separation between the spots, so that they are put on top of each other. If the predicted positions are in error, then the average profile will be broadened and/or displaced from the actual profile of the reflection. In the second step the information from the actual and the predicted profile is combined by the following process:

The observed profile $M_i$ is a sum of the Bragg peak and background. The estimate of $M_i$ - $P_i$ - is expressed by the formula

$$P_i = B_i + Const. * p_i \qquad (1)$$

where $B_i$ is the predicted value of the background and $p_i$ is predicted profile. Profile fitting minimizes the function:

$$\Sigma (M_i - P_i)^2/V_i \qquad (2)$$

where $V_i$ is variance ($\sigma^2$) of $M_i$. $V_i$ is a function of the expected signal in a pixel, which in the case of a counting detector is $P_i$. The index *i* represents all pixels in a two-dimensional profile; however, the same formulation of profile fitting applies to one and three-dimensional profiles. The predicted profile can be arbitrarily normalized; the most natural definition is that the sum of $p_i$'s is equal to one. Such a choice makes the constant in expression (1) the fitted intensity I, i.e. I = *Const.*

The profile fitting increases the accuracy (decreases the statistical error) of the measurement, but it may introduce an error due to lack of precision of the predicted profiles. *DENZO* applies the averaging of profiles in detector coordinates and, unlike other programs that use profile fitting method, averages profiles separately for each spot. The prediction of profile shape is never exact, due to errors in the positional refinement, due to averaging of different shapes, due to truncation of pixel shifts or interpolation, etc.

To calculate the diffraction intensity the background under the Bragg peak has to be estimated and then subtracted from the reflection profile. The standard method used to estimate the background value is to calculate an average detector signal in the neighborhood of a specific reflection. In *DENZO* it is assumed that the background is a linear function of the detector coordinates. Robust statistics (as discussed in the Numerical Recipes[32]) is applied to remove the contribution of pixels that deviate more than 3 sigma from the best fit to the background function. If too many background pixels are flagged as outliers from background function the whole reflection is removed from the integration. *DENZO* ignores pixels in three other cases: when they have been flagged as no measurement by an auxiliary program, when they have special value or when they are in the spot area (based on the predicted, rather than the measured position) of an adjacent reflection.

A correction for the non-linear response function of the detector to the photon flux is applied internally in *DENZO* so that it can read the original data without the need for any transformations, with the exception of the data from spiral scanners. Pixel values can represent two special cases: no measurement or detector overload. Overloaded pixels are assumed to be close to the center of gravity of the diffraction spots and as such they are used in determining the spot centroids. Pixels that are either overloaded or have no measurement are ignored in calculating the spot intensity by the profile fitting method, but the existence of such pixels in the spot area is flagged by a negative sign applied to the sigma estimate. Profile-fitted intensities seem to be reliable independent of the existence of such pixels in the spot area.

## 5.5 Scaling and Merging

The scaling and merging of different data sets as well as global refinement of crystal parameters (postrefinement) in *HKL* program suite is performed by program *SCALEPACK*. The scaling algorithm is one described by Fox and Holmes[37]. *SCALEPACK* differs in the definition of the estimated error of measurement. In *SCALEPACK*, unlike in other procedures, the estimated error is enlarged by a fraction of expected, rather than observed, intensity. The *SCALEPACK* method reduces the bias existing in other programs towards reflections with integrated intensity below the average. This program calculates single, isotropic scale and B factors for each of the batches of processed data.

*SCALEPACK* program is based on Bayesian reasoning process behind the error estimation. The essence of Bayesian reasoning in *SCALEPACK* is that you bring $\chi^2$ (or technically speaking, the goodness-of-fit, which is related to the total $\chi^2$ by a constant) close to 1.0 by manipulating the parameters of the error model. $R_{merge}$, on the other hand, is an *unweighted* statistic which is independent of the error model. It is sensitive to both intentional and unintentional manipulation of the data used to calculate it, and may not correlate with the quality of the data. An example of this is seen when collecting more and more data from the same crystal. As the redundancy goes up, the final averaged data quality definitely improves, yet the $R_{merge}$ also goes up. As a result, $R_{merge}$ is only really useful when comparing data which has been accumulated and treated the same.

## 5.6 Global Refinement (Postrefinement)

Due to correlation between crystal and detector parameters the values of unit cell parameters refined from a single image may be quite imprecise. This lack of precision is of little significance to the process of integration, as long as the predicted positions are on target. There is no contradiction here, because at some crystal/detector orientation the positions of reflections may only weakly depend on a value of a particular crystal parameter. At the end of data reduction process one would wish to get a precise unit cell values. This is done in a procedure referred to as a global refinement or postrefinement [15]. The implementation of this method in program *SCALEPACK* allows for separate refinement of the orientation of each image, but with the same unit cell value for the whole data set. In each batch of data (typically one image) a different unit cell parameter may be poorly determined; however, in a typical data set there are enough orientations to determine precisely all unit lengths and angles. The global refinement is also more precise than the processing of the single image in determination of the crystal mosaicity and orientation.

## 5.7 Experimental Feedback

The data collection is a highly interactive process. Immediate data processing can provide useful fast feedback during data collection. One has to decide on the position of the detector, the speed and the angular range of data collection. Most macromolecular crystallographic projects go through iterative stages of improving crystal and data collection strategy. Typically most of the data collection time and effort is spent before the optimal point is reached. Even then, if data collection is going well, there is a pressure from other users of the detector to use the expensive resource efficiently. The graphical interface allows visualization of the data instantly in their original form, and it can be used to view the progress of data reduction. Displaying raw data makes it possible to grasp the significance of complex patterns that would be hard to analyze numerically. This allows for a quick assement of problems in the collected data. Generally, problems may originate with the crystal, the detector, or the data reduction procedure.

Macromolecular crystals are often non-perfect. They undergo radiation damage, can be microscopically disordered or exhibit one of many kinds of macroscopic disorder - twinning, cracking, high mosaicity. The detector may be positioned too far or too close, or may be misaligned. The X-ray source may be non-uniform, incorrectly focused or non-monochromatic. Sometimes detectors fail, but still produce diffraction patterns. In order to fix the detector, its failures, which sometimes lower significantly the data quality, must be first recognized.

In the traditional approach, one collects data first and then begins analysis of the result. This strategy involves the risk that there may be a gross inefficiency in the setup of the experiment. For example: the data set may be incomplete, the reflections may overlap, the zones may overlap, a large percentage of the reflections may be overloaded etc. At that stage the only solution is to repeat the experiment, which may be difficult with unique crystals or with experiments which require synchrotron source.

Macromolecular crystallography has not developed benchmarks of acceptable performance. Sometimes lysozyme or a similar crystal is used to check the X-ray and detector system. The value of a test with such a good crystal depends on how it is analyzed. One should expect very high data quality from test crystals. An Anomalous difference fourier map should identify all sulfurs in lysozyme. All detector parameters should refine with a very small spread (tens of microns, hundreths of a degree) from one image to another. Such tests may require mounting the test crystal in such a way as to avoid slippage and minimize absorption. R-merge statistics in the range 2-3%, based on high redundancy (4 fold or higher) and high resolution (2Å or better) should be expected. Only very few (less than 0.1%) outliers should be found during merging. Worse results than above indicate a problem with the test crystal or with the experimental setup. Preferably test crystal should be kept at 100K to avoid radiation damage. Problems with the test crystal may mask detector problems. For instance, test crystal slippage makes it very difficult to notice spindle motor backlash or of the X-ray shutter malfunction.

Poor test results obviously point to experimental set-up problems. The most dangerous policy is to accept results from the test with significant number of reflections flagged as outliers, even if the R-merge statistics seem to be good. This is almost a sure sign of a serious problem and unless the problem is well understood, it may be a serious obstacle in structure solution. To understand the nature of outliers one should locate them in the detector space in order to recognize problems, like electronic failure producing single pixel spikes, a damaged detector surface or cosmic radiation. Frequently the way in which the detector test is performed may mask a significant problem. For example, if the test data are collected in a large oscillation angle mode, a shutter opening delay or spindle motor backlash may affect fewer partials than if the data are collected in a narrow frame mode. The test with tetragonal lysozyme gave acceptable results (R=4.3% for 1.7 Å data set). Subsequent data collection from Collicin E1 crystal (crystals were grown in Cynthia Stauffacher laboratory) did produce much worse statistics with 5% of observations rejected. The summary of reflections intensities and R-factors by shells are presented in the following table:

| *Resolution Shell* | | *Average Intensity* | *Average error* | *Average Stat.* | $\chi^2$ | $R_{linear}$ | $R_{Square}$ |
|---|---|---|---|---|---|---|---|
| 40.00 | 6.78 | 19929.8 | 430.7 | 199.1 | 3.721 | 0.051 | 0.065 |
| 6.78 | 5.38 | 5769.6 | 136.4 | 77.1 | 3.930 | 0.055 | 0.067 |
| 5.38 | 4.70 | 8860.7 | 212.5 | 115.3 | 4.366 | 0.055 | 0.065 |
| 4.70 | 4.27 | 9548.5 | 220.3 | 118.1 | 4.527 | 0.058 | 0.070 |
| 4.27 | 3.97 | 7174.4 | 174.9 | 95.1 | 4.039 | 0.054 | 0.060 |
| 3.97 | 3.73 | 5328.5 | 134.1 | 78.0 | 4.507 | 0.062 | 0.070 |
| 3.73 | 3.55 | 3416.7 | 91.3 | 59.4 | 3.907 | 0.066 | 0.073 |
| 3.55 | 3.39 | 2660.5 | 75.9 | 52.9 | 3.953 | 0.073 | 0.077 |
| 3.39 | 3.26 | 2080.5 | 64.8 | 47.4 | 3.705 | 0.086 | 0.088 |
| 3.26 | 3.15 | 1562.8 | 55.1 | 42.7 | 3.678 | 0.095 | 0.095 |
| 3.15 | 3.05 | 1181.1 | 45.3 | 37.1 | 3.157 | 0.100 | 0.098 |
| 3.05 | 2.96 | 912.0 | 40.5 | 34.7 | 3.066 | 0.119 | 0.112 |
| 2.96 | 2.89 | 733.3 | 36.0 | 31.9 | 3.075 | 0.131 | 0.136 |
| 2.89 | 2.82 | 582.6 | 33.8 | 30.5 | 2.876 | 0.151 | 0.151 |
| 2.82 | 2.75 | 571.6 | 33.7 | 30.8 | 2.674 | 0.154 | 0.157 |
| 2.75 | 2.69 | 426.9 | 32.2 | 30.3 | 2.341 | 0.156 | 0.153 |
| 2.69 | 2.64 | 342.3 | 30.0 | 28.6 | 2.372 | 0.168 | 0.169 |
| 2.64 | 2.59 | 355.1 | 32.1 | 30.6 | 2.411 | 0.174 | 0.167 |
| 2.59 | 2.54 | 250.6 | 29.9 | 29.0 | 2.262 | 0.209 | 0.207 |
| 2.54 | 2.50 | 229.4 | 32.1 | 31.4 | 2.115 | 0.206 | 0.206 |
| All reflections | | 3669.8 | 98.5 | 60.7 | 3.438 | 0.064 | 0.067 |

The display of the overlay of the predicted pattern on the raw data eliminated a large number of possibilities. Further investigation showed the phi motor backlash. The next data collection (when the motor backlash has been fixed) provided much better results presented in the following table.

| *Resolution Shell* | | *Average Intensity* | *Average error* | *Average Stat.* | $\chi^2$ | $R_{linear}$ | $R_{Square}$ |
|---|---|---|---|---|---|---|---|
| 40.00 | 6.78 | 56624.6 | 1161.1 | 382.5 | 0.753 | 0.013 | 0.013 |
| 6.78 | 5.38 | 16480.2 | 354.6 | 145.4 | 1.222 | 0.018 | 0.019 |
| 5.38 | 4.70 | 25422.5 | 546.2 | 228.0 | 1.110 | 0.018 | 0.019 |
| 4.70 | 4.27 | 29563.1 | 623.9 | 245.6 | 1.501 | 0.021 | 0.022 |
| 4.27 | 3.97 | 23532.1 | 536.8 | 226.8 | 1.415 | 0.023 | 0.023 |
| 3.97 | 3.73 | 17731.4 | 422.8 | 204.8 | 1.555 | 0.025 | 0.027 |
| 3.73 | 3.55 | 11496.3 | 289.9 | 169.1 | 1.471 | 0.031 | 0.034 |
| 3.55 | 3.39 | 8843.1 | 242.7 | 158.2 | 1.683 | 0.035 | 0.034 |
| 3.39 | 3.26 | 7353.4 | 228.2 | 158.7 | 1.708 | 0.043 | 0.045 |
| 3.26 | 3.15 | 5450.0 | 188.5 | 143.7 | 1.635 | 0.048 | 0.046 |
| 3.15 | 3.05 | 4372.0 | 173.7 | 140.5 | 1.779 | 0.058 | 0.054 |
| 3.05 | 2.96 | 3152.6 | 157.2 | 135.9 | 1.771 | 0.075 | 0.069 |
| 2.96 | 2.89 | 2769.5 | 153.7 | 137.7 | 1.667 | 0.080 | 0.075 |
| 2.89 | 2.82 | 2314.0 | 152.1 | 139.3 | 1.775 | 0.095 | 0.084 |
| 2.82 | 2.75 | 2227.2 | 153.0 | 141.2 | 1.928 | 0.111 | 0.103 |
| 2.75 | 2.69 | 1668.9 | 139.5 | 131.9 | 2.025 | 0.123 | 0.113 |
| 2.69 | 2.64 | 1480.3 | 137.3 | 130.7 | 1.856 | 0.122 | 0.111 |
| 2.64 | 2.59 | 1434.8 | 135.8 | 129.1 | 1.976 | 0.125 | 0.111 |
| 2.59 | 2.54 | 978.0 | 136.2 | 132.7 | 1.780 | 0.164 | 0.153 |
| 2.54 | 2.50 | 960.2 | 137.1 | 134.0 | 1.778 | 0.170 | 0.164 |
| All reflections | | 11558.9 | 309.4 | 172.2 | 1.582 | 0.027 | 0.020 |

The dinitions of values used in both tables are as follows:

$R_{linear} = \Sigma ( ABS(I - <I>)) / \Sigma (I) )$

$R_{Square} = \Sigma SUM ( (I - <I>)^2) / \Sigma (I^2) )$

$\chi^2 = \Sigma ( (I - <I>)^2) / (\sigma^2) ) * N / (N-1)$

The other factors affecting data quality can be detected on similar way. For instance, data reduction problems would result in location of reflection masks not corresponding to the positions of the Bragg peaks, severe absorption would result in non-uniform (non-radially symmetric) diffuse background level.

The new imaging plate systems like DIP systems or RAXIS-IV have very high dynamic range. For the combination of CCD detector with strong synchrotron radiation source, although the detector saturation is a critical issue. Incorrect handling of detector saturation in data acquisition hardware/software results in flat-top histogram of the largest pixel values.

Sometimes visual inspection of the calculated diffraction pattern superimposed on the diffraction image can immediately explain simple mistakes in data processing, like using wrong file format - quite often mistake in handling synchrotron data. Graphical feedback is invaluable in getting the confidence that the problem is caused by something else like non-uniform exposure during crystal oscillation. It may be due to spindle motor backlash, shutter malfunction (opening too early or too late), ionization chamber electronics (if used), decay or variation of the X-ray beam intensity (if ionization chamber is not used), variable speed of the spindle motor etc. Unfortunately, problems with lack of uniformity of exposure are best diagnosed by exclusion of other problems that may affect data quality.

Macromolecular crystallography is a highly iterative process. Rarely first crystals provides all the necessary data to solve the biological problem studied. Each step benefits from experience learned in previous steps. To monitor progress *HKL* package provides two tools:

a) Statistics - both weighted ($\chi^2$) and unweighted (R-merge). The sophisticated error model based on multi component system makes the error model realistic.
b) Visualization of the process plays double role: shows that the that all statistics are meaningful and allows to visualize certain parameters when there is no good statistical criteria of success.

## 6 Applications

The methods presented here has been applied to solve large variety of problems, from inorganic molecules with 5Å unit cell to rotavirus of 700 Å diameter crystallized in 700 * 1000 * 1400Å cell [38].

Precision of the data reduction has been tested by many researchers by successful application of the programs to MAD structure determinations, although we may expect that progress in the detector software development will push the MAD limits much further in the coming years. The combination of the techniques described above: cryo-crystallography, high intensity synchrotron radiation beam, modern experimental techniques and new fast detectors and processing software enabling use of small oscillation angles leads to very high quality data and subsequently very high quality of electron density maps as illustrated on Fig.1.

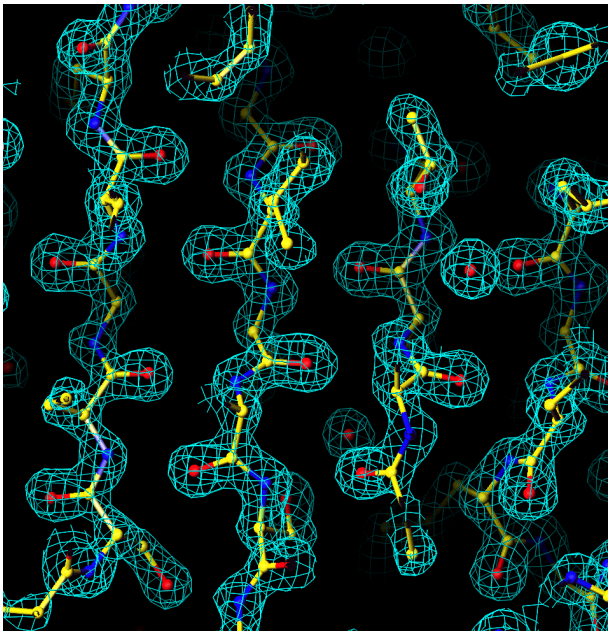

Figure 1. Portion of the $2F_o$-$F_c$ map at 1.4 Å resolution corresponding to several b strands in the C-domain of Soybean L-1 Lipoxygenase.

The electron density map presented here was produced from large challenging structure of L-1 Soybean Lipoxygenase, a single chain protein of 839 amino acids [39-40]. Data were collected on A-1 station at CHESS with 5.5cm CCD detector, processed on-line and completely scaled on site. The structure has been solved to 1.4 Å.

**Acknowledgments**

*Authors would like to thank W. Majewski, J.Raynor, K.Lewinski and Z.Derewenda for helpful discussions*

## References

[1] Herzenberg, A.; Lau H. S. M. (1967) "Anomalous Scattering and the Phase Problem", Acta Cryst. **22**:24 - 28

[2] Murthy HMK, Hendrickson WA, Orme-Johnson WH, Merritt EA, Phizackerley RP: (1988) “Crystal Structure of Clostridium Acidinurici Ferredoxin at 5 Å Resolution Based on Measurements of Anomalous X-ray Scattering at Multiple Wavelengths, J.Biol.Chem **263**: 18430-18436

[3] Guss JM, Merritt EA, Phizackerley RP, Hedman B, Murata M., Hodgson KO, Freeman HC (1988) “Phase determination by Multiple-wavelength X-ray Diffraction: Crystal Structure of a Basic ‘Blue’ Copper Protein from Cucumbers”, Science **241**:806-811

[4] Hendrickson WA, Pahler A., Smith JL, Satow Y, Merrit EA, Phizackerley RP (1989) “Crystal Structure of Core Streptavidin Determinated from Multiwavelength Anomalous Diffraction of Synchrotron Radiation” Proc. Nat. Acad Sci USA **86**: 2190-2194

[5] Smith J.L. (1991) “Determination of Three-dimensional Structure by Multiwavelength Anomalous Diffraction” Current Opinion in Structural Biology 1:1002-1011

[6] Hendrickson WA (1991) “Determination of Macromolecular Structures from Anomalous Diffraction of Sychrotron Radiation, Science **254**: 51 - 58

[7] Watenpaugh K.D., “Macromolecular Crystallography ar Cryogenic Temperatures” Current Opinion in Structural Biology **1**:1012-1015

[8] Hope, H. (1988) “Cryo-crystallography of biological macromolecules; a generally applicable method” in Acta crystallogr. **B 44**, 22 - 26.

[9] Henderson, Richard (1990) “Cryo-protection of protein crystals against radiation damage in electron and X-ray diffraction” in Proc. R. Soc. London 241, p. 6 - 8

[10] Leslie, A. “Autoindexing of rotation diffraction images and parameter refinement” Data Collection and Processing, proceedings of the CCP4 Study Weekend, 29-30 January 1993, Compiled by L Sawyer, N. Isaac, S. Bailey pp 44-51

[11] Higashi, T.J. 1990. “Auto-Indexing of Oscillation Images” J. Appl. Crystallography. **23**:253-257

[12] Evans, P. 1993. “Data Reduction: Data Collection and Processing” Proceedings of the CCP4 Study Weekend, 29-30 January 1993, Compiled by L Sawyer, N. Isaac, S. Bailey pp 114-123

[13] Leslie, A.G.W. (1987), “Profile Fitting” in proceedings of the Daresbury Study Weekend at Daresbury Laboratory, 23-24 January 1987, Compiled by Helliwell, J.R., Machin, P.A. and Papiz, M.Z., pp 39-50

[14] Greenhough, T. (1987), “Partials and Partiality” in proceedings of the Daresbury Study Weekend at Daresbury Laboratory, 23-24 January 1987, Compiled by Helliwell, J.R., Machin, P.A. and Papiz, M.Z., pp 51-57

[15] Kabsch, W. 1988. “Automatic Indexing of Rotation Diffraction Patterns” J. Appl. Crystallography. **21**:67-81

[16] Kabsch, W. 1988. “Evaluation of Single-Crystal X-ray Diffraction Data from Position Sensitive Detectors” J. Appl. Crystallography. **21**:916-924

[17] Kabsch, W. “Recent Extension of the Data-Processing Program XDS” Data Collection and Processing, proceedings of the CCP4 Study Weekend, 29-30 January 1993, Compiled by L Sawyer, N. Isaac, S. Bailey pp 63-70

[18] Rossmann, M. G., A. G. W. Leslie, S. S. Abdel-Meguid, T. Tsukihara. 1979. Processing and post-refinement of oscillation camera data. J. Appl. Crystallogr. **12**:570-581.

[19] Rossmann, M. G. 1979. Processing oscillation diffraction data for very large unit cells with an automatic convolution technique and profile fitting. J. Appl. Crystallogr. **12**:225-238.

[20] Rossmann, M. G., J. W. Erickson. 1983. Oscillation photography of radiation-sensitive crystals using a synchrotron source. J. Appl. Crystallogr. **16**:629-636.

[21] Otwinowski, Z. “Oscillation data reduction program”. Data Collection and Processing, proceedings of the CCP4 Study Weekend, 29-30 January 1993, Compiled by L Sawyer, N. Isaac, S. Bailey pp 56-62

[22] Messerschmidt, A., Pflugrath, J.W. “Crystal Orientation and X-ray Pattern Prediction Routines for Area-Detector Diffraction Systems” J. Appl. Crystallography. 1987 **20**:306-315

[23] Evans, P. 1993. “Data Reduction: Data Collection and Processing” Proceedings of the CCP4 Study Weekend, 29-30 January 1993, Compiled by L Sawyer, N. Isaac, S. Bailey pp 114-123

[24] Xuong N., Sullivan D., Nielsen C., Hamlin R., 1985. “Use of the Multiwire Area Detector Diffractometer as a National Resource for Protein Crystallography” Acta Crystallogr. **41B**:267-269

[25] Howard, A.J., Gilliland,G.L., Finzel, B.C., Poulos, T.L., Ohlendorf, D.H. and Salemme, F.R. 1987. “The Use of an Imaging Proportional Counter in

Macromolecular Crystallography" J. Appl. Crystallography. **20**:383-387

[26] Bricogne, G. (1987) "Proposal for EEC Cooperative Programming Workshop on Position-Sensitive Detector Software" in proceedings of the Daresbury Study Weekend at Daresbury Laboratory, 23-24 January 1987, Compiled by Helliwell, J.R., Machin, P.A. and Papiz, M.Z., pp 107-119

[27] EEC Cooperative Workshop on Position-Sensitive Detector Software (Phase I and II), LURE, Paris, 16 May - 7 June 1986. (Phase III), LURE, Paris, 12-19 November 1986

[28] Minor W. 1990 "Graphic Workstation - Crystallographer's Basic Tool" American Crystallographic Association Abstracts, p31

[29] Leslie, A. "Autoindexing of rotation diffraction images and parameter refinement" Data Collection and Processing, proceedings of the CCP4 Study Weekend, 29-30 January 1993, Compiled by L Sawyer, N. Isaac, S. Bailey pp 44-51

[30] Kim, S. "Auto-Indexing Oscillation Photographs" J. Appl. Crystallography. 1989 **22**:53-60

[31] Higashi, T.J. 1990. "Auto-Indexing of Oscillation Images" J. Appl. Crystallography. **23**:253-257

[32] Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T, Numerical Recipes: The Art of Scientific Computing, Cambridge University Press, 1989

[33] Rossmann, M. G., A. G. W. Leslie, S. S. Abdel-Meguid, T. Tsukihara. 1979. Processing and post-refinement of oscillation camera data. J. Appl. Crystallogr. **12**:570-581.

[34] Evans, P.R. (1987), "Postrefinement of Oscillation Camera Data" in proceedings of the Daresbury Study Weekend at Daresbury Laboratory, 23-24 January 1987, Compiled by Helliwell, J.R., Machin, P.A. and Papiz, M.Z., pp 58-66

[35] Diamond, R. (1974) Profile Analysis in Single Crystal Diffractometry" Acta Crystallographica, **A25**:43-55

[36] Ford, G. (1974) "Intensity Determination by Profile Fitting Applied to Precessing Photograph" J. Appl. Crystallography. 7:555-564

[37] Fox, G.C., Holmes, K.C. 1966. "An Alternative Method of Solving the Layer Scaling Equation of Hamilton, Rollet and Sparks" Acta Cryst. **20**:886-891

[38] Harrison, S. & Temple, B. Personal communication

[39] Minor W., Steczko J., Bolin J.T., Otwinowski Z., Axelrod B. (1993) "Crystallographic Determination of the Active Site Iron and its Ligands in Soybean Lipoxygenase L-1" Biochemistry 32:6320:6323

[40] Minor W.,. Steczko J., Stec B., Otwinowski Z., Bolin J.T., Walter R Axelrod B.,. (1996) "The Crystal Structure of L-1 Isozyme of Soybean Lipoxygenase at 1.4 Å resolution" Biochemistry *in press*.