Accurate diffraction data integration by the EVAL15 profile prediction method

Application in chemical and biological crystallography

Nauwkeurige diffractie-data integratie door de EVAL15 profielvoorspellingsmethode

Toepassing in de chemische en biologische kristallografie

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van rector magnificus, prof. dr. J.C.Stoof, in gevolge het besluit van het college voor promoties in het openbaar te verdedigen op maandag 20 april 2009 des middags te 12.45 uur

door

Xinyi Xian

geboren op 8 december 1976, te Qingdao, China

Promotor:	Prof. dr. P.Gros
Co-promotor:	Dr. L.M.J. Kroon-Batenburg

Aan mijn ouders

Het in dit proefschrift beschreven onderzoek werd gefinancierd door de technologiestichting STW en is uitgevoerd bij de vakgroep Kristal- en Structuurchemie, Bijvoet Centrum, Universiteit Utrecht, Padualaan 8, 3584 CH Utrecht, Nederland.

ISBN: 978-90-9024026-8

Contents

Chapter 1

Introduction

1.1 Crystallography

X-ray crystallography plays an important role in the determination of molecular structures at an atomic level, with applications in e.g. material science, chemical design and structural biology.
Previously unknown electromagnetic waves, having a wavelength of 10^{-8}-10^{-11}m (100-0.1Å), were discovered by Röntgen in 1895, when he was investigating various types of vacuum tube equipments. He referred to these as X-rays. Nowadays, the easiest way to produce X-rays is using sealed tubes. In a sealed tube a cathode emits electrons under high voltage, which collide with the anode at high speed. As a result of the collision, electrons of low orbitals can be freed; electrons from higher energy levels then reoccupy that position by which X-ray photons are emitted. When the transition is from the L-shell to the K-shell, K_α -radiation is emitted and when it is from the M- to the K-shell transition, K_β is emitted. K_α consists of $K_{\alpha1}$ and $K_{\alpha2}$ radiation, with only ~ $0.4 \cdot 10^{-3}$ Å wavelength difference. The electrons slowed down by the strong electric field near the nuclei produce a continuous Bremsstrahlung radiation (Fig.1.1). By using a K_β-filter where metal foils absorb the Bremsstrahlung as well as the K_β-radiation, or by using a monochromator, which is a crystal that is oriented such that it reflects only a narrow wavelength range, the $K_{\alpha1}$ and $K_{\alpha2}$ radiation can be selected.

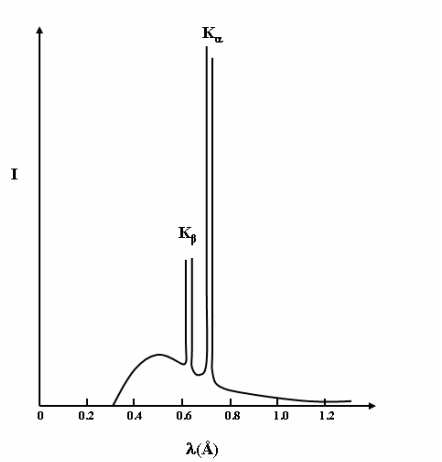


Fig. 1.1 X-ray spectrum of molybdenum showing the K_α and K_β lines superimposed on a continuous Bremsstrahlung.

When X-ray beams hit matter, the electrons, as they oscillate with the electromagnetic field, are responsible for scattering in all directions. Through the regular packing of molecules in a crystal, forming a three-dimensional periodic lattice, X-rays are diffracted in discrete directions. W.L.Bragg described diffraction as reflections against parallel lattice planes, whereby the planes are defined by the Miller indices (h k l) with distance d_{hkl} between the planes. When the path difference of scattered waves in the adjacent planes is an integer number of the incident wavelength λ, the Bragg law

$$2d_{hkl} \sin\theta = n\lambda\,, \qquad \textbf{(1.1)}$$

is fulfilled and constructive interference occurs. The integer n is the reflection order and 2θ is the angle between the incident and the refracted wave. The Miller indices (h k l) have the relation nh=h_1, nk=h_2 and nl=h_3 with the Laue indices h_1, h_2, h_3. Von Laue used these indices in an alternative way of describing the diffraction condition. The scattered waves interfere constructively in certain discrete directions when the Laue diffraction conditions

$$\mathbf{a}\cdot\mathbf{S} = h_1$$
$$\mathbf{b}\cdot\mathbf{S} = h_2$$
$$\mathbf{c}\cdot\mathbf{S} = h_3$$

are satisfied, with **a**, **b** and **c** being the basis vectors of the direct crystal lattice and **S** being the diffraction vector (see below). It is common practice to use h, k, l also for Laue indices. The dimensions of the direct space and reciprocal space have an inverse relationship, described by $\mathbf{a}^* \equiv (\mathbf{b} \times \mathbf{c})/V$, $\mathbf{b}^* \equiv (\mathbf{c} \times \mathbf{a})/V$ and $\mathbf{c}^* \equiv (\mathbf{a} \times \mathbf{b})/V$, where V is the cell volume of the unit cell and $\mathbf{a}^*$, $\mathbf{b}^*$ and $\mathbf{c}^*$ are the basis vectors of the reciprocal crystal lattice. The reciprocal lattice is a mathematical construction, which can be used conveniently in combination with the Ewald sphere to visualize the diffraction process. The Ewald sphere can be constructed in the following way. The incoming beam with wave vector $\mathbf{k_0}$ is elastically scattered by the crystal positioned at the centre of the Ewald sphere with radius 1/λ, in the direction $\mathbf{k_1}$. The origin of the reciprocal lattice is placed at O (Fig.1.2), and when $\mathbf{k_1}$ coincides with a reciprocal lattice point the Bragg condition is fulfilled and constructive interference occurs in that direction. The diffraction vector $\mathbf{S} \equiv \mathbf{k_1} - \mathbf{k_0}$ then coincides with the reciprocal lattice vector $h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*$ and has the length $|S| = 2\sin\theta/\lambda$. The lattice points can be brought in diffraction position either by rotating the crystal or by providing a broad λ-spectrum, thus effectively having a large number of Ewald spheres. The first method is also known as the monochromatic diffraction method and the second as the Laue diffraction method. Sharp peaks will be observed on the detector in directions where the scattered X-rays interfere constructively: together they form a diffraction pattern. From the positions of these spots the geometry of the unit cell can be determined.

The scattered intensity I(hkl) of the reflections is proportional to the square of the structure factor F(hkl). The structure factor is the Fourier transform of the electron density

$$F(hkl) = V \int_x \int_y \int_z \rho(x,y,z) \exp\left[2\pi i(hx + ky + lz)\right] dxdydz\,, \qquad \textbf{(1.2)}$$

where ρ(x,y,z) is the electron density at fractional coordinates (x,y,z) in the unit cell with volume V. The structure factor F(hkl) can be written as a complex number $F = |F|\exp(i\alpha)$, having the amplitude |F| and phase α.

By applying an inverse Fourier transform the electron density distribution in the unit cell can be calculated:

$$\rho(xyz) = \frac{1}{V} \sum_h \sum_k \sum_l |F(hkl)| \exp[(-2\pi i(hx + ky + lz) + i\alpha(hkl)]\,. \qquad \textbf{(1.3)}$$

From a diffraction experiment |F(hkl)| can be derived using $I \sim FF^* \sim |F^2|$. However, the phase angle α(hkl) is lost. Thus, the electron density cannot be immediately calculated and one has to find a way to estimate the phases. This is known as the phase problem.

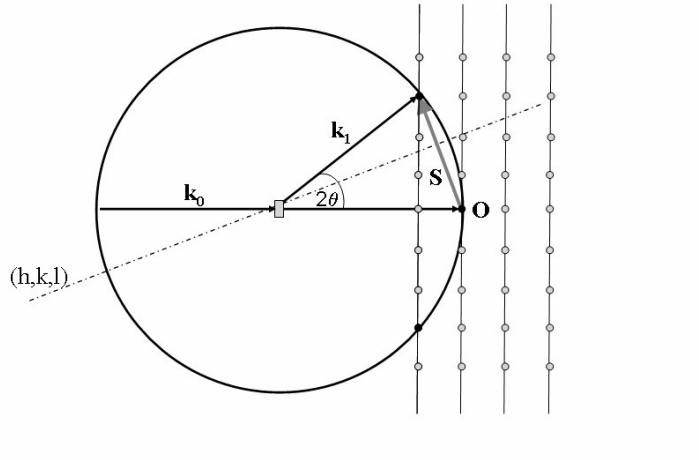


Fig. 1.2 Ewald sphere with radius 1/λ is drawn around the crystal (rectangle). The origin of the reciprocal lattice is at O. k_0 is the incident beam and k_1 the diffracted beam. S denotes the diffraction vector and is perpendicular to the lattice plane (hkl). Lattice points of the reciprocal lattice are light dots, and those in diffraction condition are dark dots.

The lost phase information can be retrieved using additional experimental or structural information.

In the **direct methods**, use is made of the fact that the positivity and atomicity of electron density restrict the possible values of phases, thus allowing initial estimates of some of them. Other phases are estimated from these through the probability relationship

$$\alpha_{-h} + \alpha_{h'} + \alpha_{h-h'} \approx 0\,. \qquad \textbf{(1.4)}$$

The limitation of this method (Uson & Sheldrick, 1999; Hauptman, 1997) is that diffraction data with high resolution ($(d_{hkl}/n)_{max}=\lambda/2\sin\theta_{max}\equiv d_{max}<$ 1.2 Å) are required, which is not often reached in protein crystallography. This method is used routinely for phasing small molecules and small proteins (up to ~1000 atoms) or finding heavy-atom substructures of larger proteins such as in SHELXD (Schneider & Sheldrick, 2002), SHARP (La Fortelle & Bricogne, 1997) and Shake-and-bake (Miller *et al.*, 1994). Through the combination of direct methods for finding the heavy-atom substructure and density modification (Foadi *et al.*, 2000) large protein structures can be solved.

Another phasing method is **molecular replacement** (Rossmann & Blow, 1962), which relies on the additional structural information of a homologous model, with a sequence identity > 25 % to the unknown structure. A Patterson map is created by Fourier transformation of the squared structure factors. A peak in the Patterson map corresponds to a difference vector between two atoms in the unit cell. The Patterson map of the homologous structure is compared to that of the unknown structure by rotation and translation, which gives information about the orientation and location of the unknown molecule in the unit cell. The application of maximum likelihood-based algorithms uses the fact that the probability of a model given the data is proportional

to the probability of the data given the model multiplied by the prior probability of the model. Read has improved this method and allows the use of lower homology models (Read, 2001).

In the **isomorphous replacement** method crystals are soaked in heavy-atom solutions to create isomorphous heavy-atom derivatives. This was first applied in small molecule crystallography (Beevers & Lipson, 1934) and later also in protein crystallography (Kendrew *et al.*, 1958; Perutz, 1956). The amplitude $|F_{PH}|$ of the derivative crystal and the amplitude $|F_P|$ of the native crystal are measured and the isomorphous difference $|F_{PH}|-|F_P|$ is used in the estimation of the heavy-atom $|F_H|$ (Fig. 1.3). Using direct methods or Patterson methods, the heavy-atom positions can be localised and the resulting $\mathbf{F_H}$ is used in the estimation of the protein phases. This method can be distinguished in the single isomorphous replacement method (SIR), using one derivative data set and the multiple isomorphous replacement method (MIR), using multiple derivative data. The disadvantage of this method is that the isomorphism between crystals can be too small to get good phase estimates.

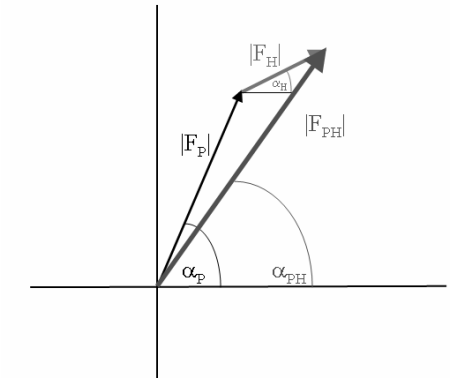


Fig. 1.3 Argand diagram for SIR. $|F_P|$ is the amplitude of the native data and $|F_{PH}|$ of the derivative data.

So far the electrons in an atom have been regarded as free electrons. However, this changes when the X-ray wavelength approaches an absorption edge. The X-ray photon energy is then sufficient to promote an electron to a higher energy shell. The scattering of the atom is then anomalous and can be written as

$$f(\theta,\lambda) = f_0(\theta) + f'(\lambda) + if''(\lambda) \qquad \textbf{(1.5)}$$

where f_0 is the atomic scattering that is dependent on the Bragg angle θ, f′ is the dispersive term that reduces the normal scattering factor and f″ is the absorption term that is 90º advanced in phase. The anomalous effect is in general stronger for heavier atoms than for light atoms. Normally the structure factor of atoms obey Friedel's law:

$$|F(hkl)| = |F(\bar{h},\bar{k},\bar{l})|. \qquad \textbf{(1.6)}$$

However, in the presence of anomalously scattering atoms, the amplitude of the total structure factor $|F^+_{PH}|$[‡] is not equal to $|F^-_{PH}|$[§] and their phases become different (Fig. 1.4). With the additional experimental information of the anomalous or Bijvoet difference the anomalously scattering atoms can be located and hence the initial protein phases can be estimated.
This is done with the multiwavelength **anomalous diffraction method** (**MAD**) (Hendrickson & Ogata, 1997), which allows the phase determination from one single crystal by collecting data at different wavelengths. The anomalous scatterers can e.g. be introduced into the protein by substituting methionine for seleno-methionine. Data are typically collected at the absorption edge (the peak data), at the inflection (inflection data) and at a remote wavelength (remote data). Collecting MAD diffraction data can take as much as eight times longer than the collection of derivative data (Walsh *et al.*, 1999; Rice *et al.*, 2000) and is only possible at synchrontrons, where the X-ray wavelength is tunable.

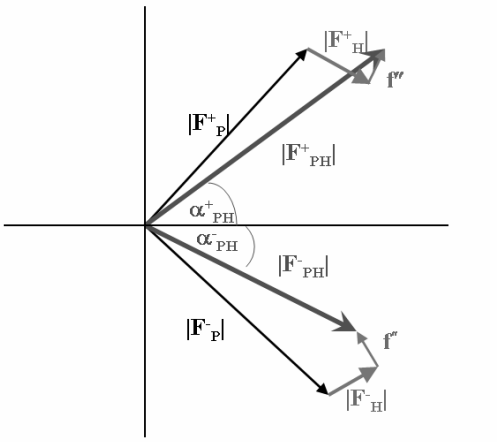


Fig. 1.4 The structure factor F^+_P and F^-_P of structures without anomalous scatterers have the same amplitude. However Friedel's law breaks down in the presence of anomalous scattering atoms. This results in unequal amplitudes of F^+_{PH} with F^-_{PH}. The Bijvoet difference is $\Delta F=|F^+_{PH}|-|F^-_{PH}|$.

A more time-efficient alternative is using single anomalous diffraction (**SAD**) data in conjunction with solvent flattening, which can unambiguously determine the phases (Wang, 1985; Hendrickson & Wuethrich, 1991-1999). There have been developments in direct-methods phasing combined with SAD data, where solvent flattening is not even needed (Hauptman, 1996; Langs *et al.*, 1999). Another advantage of SAD data is that it can also be collected in-house. It was demonstrated that with the anomalous scattering of sulphur at the wavelength of CuK_α, a macromolecular crystal structure could be solved (Dauter *et al.*, 1999).
The experimentally determined phases can be improved through solvent flattening, which removes negative electron density and flattens the solvent region, wherein the

[‡] $F^+_{PH}= F(hkl)_{PH}$
[§] $F^-_{PH}= F(\bar{h},\bar{k},\bar{l})_{PH}$

molecular structure can be more easily interpreted. Other techniques are histogram-matching and non-crystallographic averaging. Density modification is a cyclic procedure: the modified electron-density is back-transformed to give modified phases, these in combination with experimental phases are then used in the calculation of a new map. This cycle continues until convergence is reached. Programs like SHELXE (Sheldrick, 2008), DM (Cowtan & Zhang, 1999), CNS (Brunger *et al.*, 1998) and RESOLVE (Terwilliger, 2003) carry out density modification and produce improved electron-density maps. The amino-acid sequence and known structural characteristics are vital in the process of building the protein molecule. This can be carried out with automated model-building programs like e.g. RESOLVE (Terwilliger, 2004) or ARP/wARP (Morris *et al.*, 2002, 2003), where atoms are repositioned in the electron density map in combination with unrestrained refinement. This requires a high observation-to-parameter ratio and an electron density map with high resolution. The molecular structure can be build by hand in e.g. COOT (Emsley & Cowtan, 2004).
The initial model is refined, which results in a model that gives the best fit between the observed structure factor amplitudes and the calculated amplitudes. The quality of the fit is measured with a R-factor (see below).
Small molecule refinement carried out with SHELXL (Sheldrick, 2008) is based on the principle of least squares. The iterative process minimizes the target function

$$Q = \sum_{hkl} w_{hkl} \left(\left| F_{obs}(hkl) \right| - \left| F_{calc}(hkl) \right| \right)^2 , \qquad \textbf{(1.7)}$$

where w_{hkl} is the weight given to an observation. Difference Fouriers assist in improving the model.
With the formulation of maximum likelihood a statistically valid way to deal with errors and model incompleteness was developed. The refinement program REFMAC (Winn *et al.*, 2001) is entirely based on this formalism and is used for the refinement of protein structures. The observation/parameter ratio of protein data is usually not high. To avoid overfitting of experimental data, geometrical knowledge is also taken into account in the refinement. With the better set of phases, a new model can be fit to the new electron density map and a further round of refinement is carried out. This continues until the correlation between the diffraction data and the model is maximized. Depending on the resolution, the refined parameters are three positional parameters (x,y,z) and one isotropic temperature factor B or entire domains can be modelled during the refinement anisotropically, using TLS-parametrization of translation-, libration- and screw-rotation displacements. Since recently, the refinement of macro-molecules can also be carried out with SHELXL (Sheldrick, 2008), but this requires data with at least 2 Å resolution.

1.2 History of data integration

Accurate integration of reflection intensities plays an essential role in structure determination. The intensities of reflections can be recorded with photographic films, with point detectors or with area detectors, such as CCD-cameras or image plates. There are two methods to obtain estimates of the diffracted intensity I: summation integration or profile fitting.
In small molecule crystallography it has been common practice to use summation integration, which is also referred to as the Background-Peak-Background (BPB) method. On a classic 4-circle diffractometer, all reflections are measured one by one.

An adjustable aperture is placed behind the crystal and in front of the point detector at the place where the reflected ray is expected to pass. The intensity that passes this rectangle aperture is measured. An one-dimensional scan through the reflection is made and the net intensity is found with

$$I_{net} = I_P - k \cdot I_B, \tag{1.8}$$

with k=2 being the ratio between the width of the peak region P and the background region B (the B region is equal to L+R in Fig. 1.5). The photon counts behave Poissonian, therefore the standard deviation $\sigma(N)=N^{1/2}$ and

$$\sigma(I_{net}) = (I_P + k^2 \cdot I_B)^{1/2}. \tag{1.9}$$

A well defined peak region or reflection contour is needed to apply summation integration. A too large reflection contour leads to a lower I/σ ratio and may lead to less reliable atomic positions in the structure refinement. A too small reflection contour leads to intensity loss and may result in an incorrectly refined B-factor. Lehmann& Larsen (1974) chose the reflection border for one-dimensional reflections such that I/σ is maximized. However, Duisenberg has shown that this leads to systematically too small peak regions (Duisenberg *et al.*, 2003). Using an area detector, two-dimensional reflections can be recorded. Pflugrath & Messerschmidt (1993) use standard ellipsoids for the peak region. By summation of diffraction images over a couple degrees of ω-rotation, a reflection is recorded in three-dimensions. The border for a three-dimensional reflection is calculated accurately with an ab inito method in the data integration program EVAL14 (Duisenberg *et al.*, 2003). In this method only a few physical and instrument parameters are needed to generate the reflection contour using the principle of general impacts (Duisenberg *et al.*, 2003). By tracing the X-ray originating from extreme points in the beam source to extreme points in the crystal and using extreme mosaic vectors, a contour is calculated for a reflection at a given position of the detector.

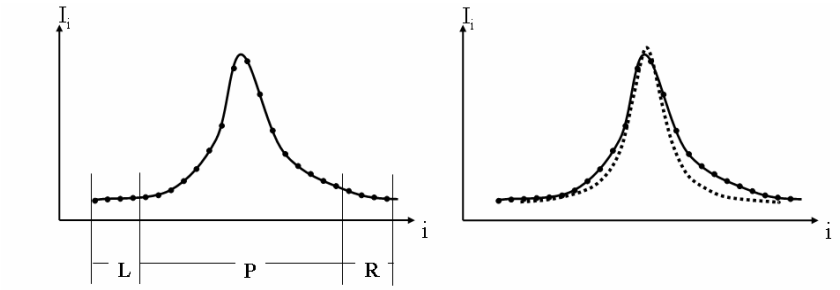


Fig. 1.5 On the left: summation integration of the one-dimensional reflection: regions L and R form the background B. P denotes the peak region and I_{net}=P-kB, with k=2 being the ratio of P and B. On the right: profile fitting of the observed intensity with a standard profile (dotted).

The summation integration method has the disadvantage that intensities from e.g. zingers, the radioactive-decay in the fiber taper and phosphor, also contribute to I_{net} and that the complete peak region must be available to derive the intensity of a reflection.

Nowadays, profile fitting is the standard technique to integrate reflection intensities. The net intensity is derived from a least-squares fit of a standard model profile with the observed profile, which determines the scale factor J. The intensity I_P is calculated with:

$$I_P = J\sum_i P_i\,, \qquad \textbf{(1.10)}$$

where P_i is the profile value of the model for the i-th pixel (Fig. 1.5). To derive the standard profile, two assumptions are made. The first assumption is that neighbouring reflections, whether on the detector or in reciprocal space, have similar profiles. The second is that positions of peak profiles should be precisely predicted (Pflugrath, 1999). Since the profiles vary across the detector plate due to geometrical deformations, standard profiles have to be learned from nearby neighbour reflections and are created by superimposing observed reflection profiles with high I/σ's. Due to the mapping of pixels onto each other the reference profile is always slightly broader than that of an observed one. If the position of the reflection peaks would not be accurate, it would lead to a significant broadening. Reflections can be recorded completely within a rotation-angle increment (fullies) or only partially (partials). It is assumed that the reflection profiles of fullies and partials are the same (Otwinowski & Minor, 1997; Pflugrath, 1999). The learning process of profiles encounters problems, when neighbour reflections are weak or overlapping. As high I/σ reflections are needed, the integration is hampered at high resolution regions. In the case of multiple lattices or long cell axes, reflections can occasionally overlap, implying that the learned profile may not match the observed profile. $K_{\alpha1}$,$K_{\alpha2}$-splitting would be possible to learn, but this demands an even higher accuracy of the prediction of peak position. Moreover, even a small θ-difference on the detector can change the $K_{\alpha1}$,$K_{\alpha2}$-splitted profile significantly. Due to these reasons, profile learning methods are hard to apply in case of $K_{\alpha1}$,$K_{\alpha2}$-splitting. When reflection profiles differ due to anisotropic mosaicity or lattice distortion, the first assumption of profile fitting (similar profile shapes) is not adhered to. Profile fitting has a couple of advantages: effects of zingers do not influence the net intensity (or are very much reduced), complete reflections are not needed to derive the intensity and the I/σ ratio is improved (especially for weak reflections). Diamond used profiles of strong reflections as standard profiles to integrate one-dimensional reflections and estimated that the standard deviations can be lowered by as much as a factor $2^{1/2}$ (Diamond, 1969). Other profile methods for one-dimensional profiles are described by Clegg (1981) and by Oatley & French (1982). Clegg also treats $K_{\alpha1}$,$K_{\alpha2}$-splitting. The data integration program DENZO (Otwinowski & Minor, 1997) and MOSFLM (Leslie, 1999) integrate two-dimensional profiles. Intensity estimates of reflections that lie on different images are then summed after the integration step in a post-refinement/scaling step. Three-dimensional profiles are integrated in programs like d*TREK (Pflugrath, 1999), SAINT (Bruker AXS, Madison, WI) and XDS (Kabsch, 1988). Kabsch overcame the necessity of having strong neighbouring reflections: by transformation to the reciprocal space. However, as reflection profiles are a convolution of many parameters, we believe that strict transformation to reciprocal lattice is impossible. These programs have in common that only diffraction patterns arising from a single crystal lattice can be described.

Due to the principle of general impacts, EVAL14 is very flexible and can even handle anisotropic mosaicity, mica effects and treat $K_{\alpha1}$,$K_{\alpha2}$-splitting. With the indexing

program DIRAX (Duisenberg, 1992), which is a part of the EVAL-software suit, reflections arising from multiple crystal lattices can be indexed and integrated straightforwardly.

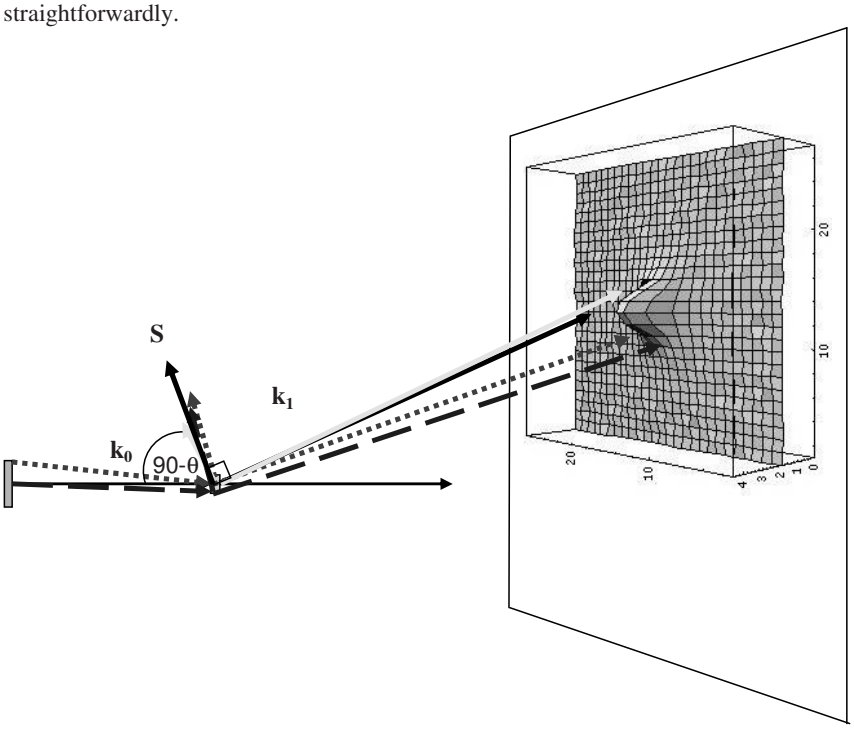


Fig. 1.6 The principle of general impacts is illustrated with a selection of traced rays. One originates from the centre of the focus (left rectangle) to the centre of the crystal (right rectangle). When the angle between the diffraction vector S and the incoming beam is 90-θ, the diffraction condition is met and the resulting out-ray is detected as the central impact of this reflection (black solid arrow). Another ray originates from an extreme point of the focus (short dashed arrow). A third X-ray hits an extreme point of the crystal (long dashed arrow). The light coloured arrow illustrates a different mosaic vector. The collections of all possible impacts form the reflection on the detector. When one selects only the extreme impacts, the reflections border is generated.

In this thesis we introduce the data integration method EVAL15 (Schreurs *et al.*, 2009), that combines the advantages of EVAL14, with the more accurate integration of weak intensities by using profile fitting. This program does not need neighbour reflections to derive a standard profile. With a few physical crystal and instrument parameters, like mosaicity, crystal shape, beam divergence and wavelength, a three-dimensional standard profile is predicted. The simulation of profiles is done by generating impacts through X-ray tracing from random sample points of the already mentioned physical parameters (Fig. 1.6). This standard profile is then used in a profile fitting algorithm using Singular Value Decomposition (SVD) (Press *et al.*, 1986).

1.3 Data quality indicators

The quality of the crystal structure determination is directly dependent on the quality of the underlying diffraction data (Weiss, 2001).
During the data collection, the crystal is rotated and the measured diffraction intensity is affected by physical factors from the experiment, e.g. minimal fluctuations in the beam flux, the position of the primary beam relative to the crystal rotation axis or to the detector. Therefore, the intensities have to be corrected through a scaling procedure. In SADABS (Sheldrick, 1996) the intensities of equivalent or repeatedly measured reflections are used in the scaling process and from these σ's of the intensities are estimated. The signal-to-noise ratio I/σ obtained after scaling is a good measure of the precision of the intensity.
By comparing the intensity of a reflection with the averaged intensity of the set of equivalents, a merged R factor is derived. The most frequently reported R factor is

$$R_{merge} = \frac{\sum_{hkl}\sum_{i} | I_i(hkl) - \overline{I(hkl)} |}{\sum_{hkl}\sum_{i} I_i(hkl)}, \qquad \textbf{(1.11)}$$

which is intrinsically dependent on the redundancy of a data. Lower redundancy yields a low R_{merge}, but less accurate data. The redundancy-independent-merging R factor $R_{r.i.m.}$ (Weiss *et al.*, 1998) is calculated with

$$R_{r.i.m.} = \frac{\sum_{hkl}[N/(N-1)]^{1/2}\sum_{i} | I_i(hkl) - \overline{I(hkl)} |}{\sum_{hkl}\sum_{i} I_i(hkl)}. \qquad \textbf{(1.12)}$$

Since this R-factor takes into account how many times (N) a reflection is measured, the precision of measurements is redundancy independent.
Weiss introduced the precision-indicating-merging R-factor $R_{p.i.m.}$ (Weiss & Hilgenfeld, 1997; Weiss *et al.*, 1998), which calculates the precision of averaged measurements. Since averaged intensities or amplitudes are normally used in the structure determination and refinement,

$$R_{p.i.m.} = \frac{\sum_{hkl}[1/(N-1)]^{1/2}\sum_{i} | I_i(hkl) - \overline{I(hkl)} |}{\sum_{hkl}\sum_{i} I_i(hkl)}, \qquad \textbf{(1.13)}$$

should give the best prediction about the performance of the data set in structure determination. All these R-factors give information about the equivalence of equivalent reflections. Systematic errors in the integration of intensities cannot be measured with these quality indicators.
After merging of the data, the quality can be assessed by looking at the resolution, completeness, signal-to-noise ratio I/σ, the redundancy and the Wilson plot. Highly redundant data have intrinsically higher quality than data consisting of single measurements. One quality indicator is the maximum resolution d_{max}, which is related to the resolution in the electron density map. However, limited completeness leads to a lower effective resolution in the map.

For protein data the following additional quality indicators can be used. The phasing capacity can be determined by looking at the anomalous signal-to-noise ratio $\Delta F/\sigma_{\Delta F}$ or the anomalous signal $\Delta F/F$ (Hendrickson & Teeter, 1981). After location of the anomalous scatterer-sites with e.g. the program SHELXD or SOLVE (Terwilliger, 2003), the correlation coefficient CC_E between E_o and E_c, is calculated with:

$$CC_E = 100\frac{(\sum wE_oE_c\sum w-\sum wE_o\sum wE_c)}{\sqrt{[\sum wE_{o^2}\sum w-(\sum wE_o)^2][\sum wE_{c^2}\sum w-(\sum wE_c)^2]}}, \quad \textbf{(1.14)}$$

(Schneider & Sheldrick, 2002), where E_c is calculated from the located sites and E_o is derived from the observed F_A's. The success rate of correctly located sites is also a good phasing quality indicator.
The phase errors of the initial phases and the refined phases can be compared as well as the reciprocal map correlation CC_{map} (Rice *et al.*, 2000), which is calculated with

$$CC_{map} = \frac{\sum_{hkl} fom_i\left|F_i^{hkl}\right| fom_j\left|F_j^{hkl}\right|\cos(\phi_i-\phi_j)}{\sqrt{\sum_{hkl}\left|fom_iF_i^{hkl}\right|^2\sum_{hkl}\left|fom_jF_j^{hkl}\right|^2}}, \quad \textbf{(1.15)}$$

where fom denotes the figure of merit, |F| the structure factor amplitude, ϕ_i is the phase calculated from the initial model and ϕ_j that of the refined model.
For both small molecule and protein data the quality of refinement of a structure is given by the R-value

$$R = \frac{\sum_{hkl}\left\|F_{obs}(hkl)\right|-k\left|F_{calc}(hkl)\right\|}{\sum_{hkl}\left|F_{obs}(hkl)\right|}. \quad \textbf{(1.16)}$$

In small molecule refinement the following criteria are used. The maximum and minimum rest-density $\Delta\rho_{max/min}$ of the difference Fourier should be close to zero, because small molecule crystal data mostly have a very high observation/data ratio. The refinement program SHELXL refines against F^2. Besides the conventional R-value, a weighted R-value $wR_2=\{\Sigma[w(F_o^2-F_c^2)^2/\{\Sigma[w(F_o^2)^2]\}^{1/2}$ is used. The weight w is defined as $1/[\sigma^2(F_o^2)+(aP)^2+bP]$, where $P=(F_o^2+2F_c^2)/3$. a and b are introduced, because structure factor calculation is an approximation based on atomic scattering factors and therefore F_c will contain errors. However, if a and b are large, $\sigma(F_{obs})$ is apparently not accurate enough. In addition, the "goodness of fit" $S=\{\Sigma[w(F_o^2-F_c^2)^2/(n-p)]\}^{1/2}$ should be close to one, where n=number of reflections and p=number of parameters refined.
For protein data, the quality of refinement can be assessed with the value of the figure of merit (fom), indicating the correctness of the phases and the root mean square deviation (rms) of the bond length, bond-angles etc. from the ideal value. Due to the low observation/parameter ratio of protein data an additional R-value, R_{free} is calculated with ~5%-10% of the data, that is kept apart from the refinement in order to do the cross-validation. A last criterion to judge the data is by looking directly at

the $2F_o$-F_c electron density maps, which reveal how well the refined structure and the electron density coincide.

1.4 Scope of the thesis

The new diffraction-data integration method EVAL15 is presented in this thesis and its application both on standard and overlap diffraction data is investigated.
In chapter 2 the method of EVAL15 is described. Reflection profiles are predicted based on the principle of general impacts (Duisenberg *et al.*, 2003), using only a small number of physical parameters. The reflection intensity is then derived by least-squares fit of the predicted profile to the observed profile. In chapter 3 the application of this method on standard diffraction data is investigated. The quality of EVAL15 data is assessed for a set of standard diffraction experiments both for small molecule and protein crystals and compared to that of EVAL14 (Duisenberg *et al.*, 2003), in particular to investigate if the EVAL15 profile method results in improved quality of the weaker data. The potential of EVAL15 dealing with complicated overlapping reflection data due to multiple lattices or a long cell axis is shown in chapter 4. In chapter 5 reflection profiles are studied in detail, which can reveal peculiar features of the crystal structure or unexpected instrumental characteristics.
Finally a summary of the results and a general conclusion of the work described in this thesis are presented.

References

Beevers, C. A. & Lipson, H. (1934). *Proc.R.Soc.London Ser.A* **146**, 570-582.
Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905-921.
Clegg, W. (1981). *Acta Cryst.* A**37**, 22-28.
Cowtan, K. D. & Zhang, K. Y. (1999). *Prog. Biophys. Mol. Biol.* **72**, 245-270.
Dauter, Z., Dauter, M., de La Fortelle, E., Bricogne, G. & Sheldrick, G. M. (1999). *J. Mol.Biol.* **289**, 83-92.
Diamond, R. (1969). *Acta Cryst.* A**25**, 43-55.
Duisenberg, A. J. M. (1992). *J. Appl.Cryst.* **25**, 92-96.
Duisenberg, A. J. M., Kroon-Batenburg, L. M. J. & Schreurs, A. M. M. (2003). *J. Appl.Cryst.* **36**, 220-229.
Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126-2132.
Foadi, J., Woolfson, M. M., Dodson, E. J., Wilson, K. S., Jia-xing, Y. & Chao-de, Z. (2000). *Acta Cryst.* D**56**, 1137-1147.
Hauptman, H. A. (1996). *Acta Cryst.* A**52**, 490-496.
Hauptman, H. A. (1997). *Curr. Opin. Struct. Biol.* **7**, 672-680.
Hendrickson, W. A. & Ogata, C. M. (1997). *Methods Enzymol.* **276**, 494-522.
Hendrickson, W. A. & Teeter, M. M. (1981). *Nature* **290**, 107-113.
Hendrickson, W. A. & Wuethrich, K. (1991-1999). *Macromolecular Structures.* London: Current Biology Publications.
Kabsch, W. (1988). *J. Appl.Cryst.* **21**, 916-924.
Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H. & Phillips, D. C. (1958). *Nature* **181**, 662-666.
La Fortelle, E. d. & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472-494.
Langs, D. A., Blessing, R. H. & Guo, D. (1999). *Acta Cryst.* A**55**, 755-760.
Lehmann, M. S. & Larsen, F. K. (1974). *Acta Cryst.* A**30**, 580-584.
Leslie, A. G. W. (1999). *Acta Cryst.* D**55**, 1696-1702.
Miller, R., Gallo, S. M., Khalak, H. G. & Weeks, C. M. (1994). *J.Appl.Cryst.* **27**, 613-621.
Morris, R. J., Perrakis, A. & Lamzin, V. S. (2002). *Acta Cryst.* D**58**, 968-975.
Morris, R. J., Perrakis, A. & Lamzin, V. S. (2003). *Methods Enzymol.* **374**, 229-244.

Oatley, S. & French, S. (1982). *Acta Cryst.* A**38**, 537-549.
Otwinowski, Z. & Minor, W. (1997). *Macromolecular Crystallography, Pt A* **276**, 307-326.
Perutz, M. F. (1956). *Acta Cryst.* **9**, 867-873.
Pflugrath, J. W. (1999). *Acta Cryst.* D**55**, 1718-1725.
Plugrath, J. W. & Messerschmidt, A. (1993), *MADNES Reference Manual*,v.2.4, Nonius Delft B.V. Delft, The Netherlands.
Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1986). *Numerical Recipes*. Cambridge: Cambridge University Press.
Read, R. J. (2001). *Acta Cryst.* D**57**, 1373-1382.
Rice, L. M., Earnest, T. N. & Brunger, A. T. (2000). *Acta Cryst.* D**56**, 1413-1420.
Rossmann, M. G. & Blow, D. M. (1962). *Acta Cryst.* **15**, 24-31.
Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst.* D**58**, 1772-1779.
Schreurs, A. M. M., Xian, X. & Kroon-Batenburg, L. M. J. (2009). *J. Appl.Cryst.*
Sheldrick, G. M. (1996). *SADABS.* University of Goettingen, Germany.
Sheldrick, G. M. (2008). *Acta Cryst.* A**64**, 112-122.
Terwilliger, T. C. (2003). *Methods Enzymol.* **374**, 22-37.
Terwilliger, T. C. (2004). *J. Synchrotron Rad.* **11**, 49-52.
Uson, I. & Sheldrick, G. M. (1999). *Curr. Opin. Struct. Biol.* **9**, 643-648.
Walsh, M. A., Evans, G., Sanishvili, R., Dementieva, I. & Joachimiak, A. (1999). *Acta Cryst.*D**55**, 1726-1732.
Wang, B. C. (1985). *Methods Enzymol.* **115**, 90-112.
Weiss, M. S. (2001). *J.Appl.Cryst.* **34**, 130-135.
Weiss, M. S. & Hilgenfeld, R. (1997). *J.Appl.Cryst.* **30**, 203-205.
Weiss, M. S., Metzner, H. J. & Hilgenfeld, R. (1998). *FEBS Lett.* **423**, 291-296.
Winn, M. D., Isupov, M. N. & Murshudov, G. N. (2001). *Acta Cryst.* D**57**, 122-133.

Chapter 2

EVAL15: a diffraction data integration method based on *ab initio* predicted profiles*

Antoine Schreurs†, Xinyi Xian† & Loes Kroon-Batenburg

* Submitted to the Journal of Applied Crystallography
† These authors contributed equally to this work

Abstract

A novel diffraction data integration method is presented, EVAL15, based upon *ab initio* calculation of 3-dimensional (x,y,ω) reflection profiles from a few physical crystal and instrument parameters. Net intensities are obtained by least squares fitting the observed profile with the calculated standard using Singular Value Decomposition. This paper shows that profiles can be predicted satisfactorily and that accurate intensities are obtained. The detailed profile analysis has the additional advantage that specific physical properties of the crystal are revealed. The EVAL15 method is particularly useful in circumstances where other programs fail, such as regions of reciprocal space with weak scattering, crystals with anisotropic shape or anisotropic mosaicity, $K_{\alpha1}$,$K_{\alpha2}$ peak splitting, interference from close neighbours, twin lattices or satellite reflections of modulated structures, all of which may frustrate the customary profile learning- and fitting procedures. EVAL15 straightforwardly allows the deconvolution of overlapping reflections.

2.1 Introduction

Several software packages have been developed for the integration of diffraction data from area detectors. Compared to photon-counting point detectors some extra problems have to be solved to get accurate intensities (see Zhurov *et al.* (2008) for an error analysis of data of point detectors, CCD's and image plates). The advantages are: faster data collection and a complete picture of reciprocal space. All sorts of phenomena related to specific crystal properties can be visible at a glance like twinning, aperiodic structure, disorder and thermal diffuse scattering, and also unwanted effects like the formation of ice at cryo-temperatures. The developments in detector technology and integration software have been triggered by macromolecular crystallography, where a large number of reflections can be collected simultaneously, many of which usually have a low signal. In small molecule crystallography it was common practice to acquire reflection data through summation-integration. However, for weak reflections a better standard deviation can be achieved by profile fitting (Diamond, 1969; Ford, 1974). This involves a least-squares fit of the observed pixel intensities in a reflection peak to a learned standard profile. The profile learning process relies on two main assumptions (Pflugrath, 1999). Profiles of strong reflections are superimposed to construct an averaged standard profile. As the reflection profile varies with the position on the detector due to geometrical deformations, it is assumed that only standard profiles learned from spatially nearby reflections give an adequate description. The second assumption is that the reflection positions are predicted accurately. Uncertainties in reflection centroids lead to artificially broad profiles and to wrong profile fits. Profile learning and fitting can be carried out in two dimensions on a single image, like with Denzo/HKL2000 and MOSFLM (Otwinowski & Minor, 1997; Leslie, 1999) or in a complete three-dimensional reflection box like with XDS, d*TREK and SAINT and Crysalis (Kabsch, 1988; Pflugrath, 1999; Bruker, 1998; Oxford Diffraction, 2008).
The need for yet an other integration program lies in the fact that each of the existing ones lacks one of the following properties. 1) Profile fitting in regions of reciprocal space where all reflections are weak: profile learning needs high I/σ reflections, usually non-existent at high-resolution. 2) Proper treatment of reflections with $K_{\alpha1}$,$K_{\alpha2}$-splitting, which is a prerequisite for high-resolution studies. 3) Use of twin matrices. 4) Deconvolution of overlapping reflections.

Kabsch (1988) developed an elegant procedure to get uniform three-dimensional profiles for all reflections by transformation to the undistorted reciprocal space, thereby overcoming the need for strong nearby reflections in profile learning. However, even then the results are better if the standard profiles are learned separately from different regions on the detector if suitable reflections are available. As reflection profiles are a convolution of broadening effects, such as crystal size and shape, mosaicity, focus dimensions and beam divergence, wavelength dispersion, experimental geometry, lattice distortions and internal structure of the crystal, detector point spread and spatial distortion, an exact transformation to reciprocal space is impossible. This insight led us to predict accurate reflection profiles by taking into account all these effects explicitly and apply this standard profile in a least-squares fit. EVAL15 is based on general impacts as introduced in EVAL14 (Duisenberg *et al.*, 2003). In that program an *ab initio* reflection boundary is calculated, within which summation-integration is performed. The method is widely used in chemical crystallography, in particular the version implemented in COLLECT (Nonius, 1999). Building on that experience EVAL15 calculates a complete standard reflection profile from general impacts. We will discuss the method and algorithm of EVAL15, the details of its implementation and the quality of the profiles. In separate papers the EVAL15 data quality and the performance in the deconvolution of overlap will be addressed.

2.2 The EVAL15 Method

In this section all steps in the EVAL15 data integration method are explained. General impacts are generated by sampling from distributions of physical parameters. These impacts have to be convoluted with a detector point spread function in order to get a realistic predicted profile. For each individual reflection such a profile is used in a least-squares minimization using SVD to get the integrated intensity and its standard deviation. Contributions to the standard deviations are discussed.

2.2.1 General impacts

The concept is explained in detail by Duisenberg *et al.* (2003), here we give only the principle.
Consider a diffraction experiment with one rotation axis and an area detector. The reflection normal $\mathbf{S}_0$ for reflection *hkl* in the zero position of the goniometer is:

$$\mathbf{S}_0 = \begin{pmatrix} S_{0x} \\ S_{0y} \\ S_{0z} \end{pmatrix} = \begin{pmatrix} a_x^* & b_x^* & c_x^* \\ a_y^* & b_y^* & c_y^* \\ a_z^* & b_z^* & c_z^* \end{pmatrix} \begin{pmatrix} h \\ k \\ l \end{pmatrix}. \qquad \textbf{(2.1)}$$

The matrix containing the reciprocal cell axes in the laboratory axis system is called the **R**-matrix. If $\mathbf{S}_0$ can be rotated over some angle ω to a position $\mathbf{S}_\omega$ such that the angle between $\mathbf{S}_\omega$ and the primary beam equals 90º–θ, then and only then *hkl* will reflect. The diffracted ray departs from the crystal along

$$\mathbf{r} = \mathbf{S}_\omega - \mathbf{X}/\lambda\,, \qquad \textbf{(2.2)}$$

where **X** is a unit vector along the primary beam pointing to the focus centre. Eq.(2) follows from $\mathbf{S}_\omega$ bisecting $\angle$ (**X**,**r**). We denote the 'central impact' co-ordinates by

(x,y,ω), with x,y the impact position on the detector plane and ω the rotation angle at which *hkl* is brought to reflection. This point in (x,y,ω) space represents the complete reflection that would be obtained from a point source, point crystal with no mosaicity and pure monochromatic radiation.
In practice a reflection results from radiation of different wavelengths, coming from different parts of the focus and scattered by different parts of the crystal having different orientations of the mosaic blocks. Each combination of these parameters may yield a general impact with co-ordinates (x,y,ω), as follows.
Consider one point **K** of the crystal, one mosaic vector $\mathbf{S_m}$, one possible focal point **F** and one wavelength λ, then this combination will reflect if, by ω rotation, the angle between $\mathbf{S_{m,\omega}}$ and $\mathbf{F}-\mathbf{K}_\omega$ can be made 90º–θ. The outgoing direction **r** for a general impact is given by:

$$\mathbf{r} = \mathbf{S_{m,\omega}} - \left[(\mathbf{F}-\mathbf{K}_\omega)/\left|\mathbf{F}-\mathbf{K}_\omega\right|\right]/\lambda\,. \qquad \textbf{(2.3)}$$

The origin of r and $S_{m,\omega}$ is not (0,0,0), but K_ω. The formula follows from $S_{m,\omega}$ being the bisector of $\angle((\mathbf{F}-\mathbf{K}_\omega),\mathbf{r})$ the incoming and reflected ray, respectively. The subscript $_\omega$ denotes ω-rotated vectors.

2.2.2 Modelling the profile

For each reflection EVAL15 general impacts are calculated for randomly selected $(\mathbf{F},\mathbf{K},\mathbf{S_m},\lambda)$ combinations chosen from the sets of all focal points, crystal points, mosaic vectors and wavelengths, respectively. A sufficiently large number of selections from realistic distributions will generate a true reflection profile eventually. (See Appendix A for sampling methods of the various distributions.) The simulated profile is used as a standard profile in a least-squares fit. As the simulation is carried out for each individual reflection, specific reflection geometries are automatically accounted for.
The focus is modelled by a rectangular surface with realistic dimensions (e.g. 0.3x0.3 mm^2) consisting of a grid of point sources that scatter in all directions. It is assumed that from each point source a ray can hit any point in the crystal. By changing the distance of this virtual focus to the crystal, the divergence of the beam can be changed. A small distance corresponds to a larger divergence. The points on the focus are sampled homogeneously, although a Gaussian distribution of intensities around the focal centre could be more realistic when certain optical elements are used. This procedure delivers a collection of vectors **F**.
The crystal shape can be described by face indexing, or, alternatively, by one of five basic shapes built into the program (a pie, a box, a sphere approximated by a dodecahedron or a icosahedron or a cylinder based on an octagon). The crystal is treated as a fine rectangular grid and each of these grid points can be selected, thus obtaining a collection of vectors **K** defined relative to the origin.
Three distributions can be used to describe the mosaic spread. Random polar angles are sampled according to a block shaped function within the range given by the mosaicity μ or by a Gaussian or a Lorentzian distribution of width σ_m. For the latter two $3\sigma_m$ corresponds to the mosaicity μ. Each vector $\mathbf{S_m}$ is then obtained by rotating $\mathbf{S_0}$ over these polar angles followed by a rotation over a random azimuthal angle (in the range 0-2π).
The wavelength of the rays is described by a spectrum built from several Gaussian or Lorentzians, each having a central $\boldsymbol{\lambda}$-value and a width σ_λ, each of which has a defined

relative integrated ratio. Characteristic radiation from a sealed tube is described by a pair $K_{\alpha1}$,$K_{\alpha2}$ with an intensity ratio 2:1. Algorithms for sampling the various distributions are explained in Appendix A The produced rays hit the detector at impact positions (x,y) and will each be collected in one associated pixel.
Position sensitive detectors convert X-ray photons into an electronic signal. This process involves several steps like absorption of the photons by a phosphor layer, photon storage or conversion to visible light photons, laser read-out (in case of image plates) or light transportation through a fibre optic taper to a CCD-chip and analog-to-digital conversion (Arndt, 1986). This cascade causes the X-ray signal to spread out over several pixels, although it hits the detector at a single point. The main source of the point-spread is usually the phosphor layer (Bourgeois *et al.* 1994) and its broadening effect increases with layer thickness and with incidence angle of the impact. We found that, when simulating impacts for realistic dimensions of the crystal, focus and mosaic spread, the resulting profiles were too narrow when the point-spread was neglected. We have introduced a two-dimensional pseudo-Lorentzian as the point spread function and took care that the integral over space to infinity, in terms of polar coordinates measured from the centre, converges to 1.0. A symmetric function is currently implemented in EVAL15 (Fig. 2.1):

$$\mathbf{PSF}(x,y) = \frac{\gamma}{4\pi\left[\left(x^2+y^2\right)+\left(\frac{1}{2}\gamma\right)^2\right]^{1\frac{1}{2}}}, \qquad \textbf{(2.4)}$$

where γ denotes the width of the function.

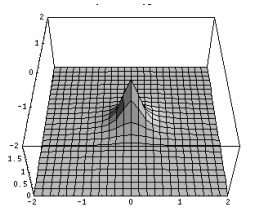


Fig. 2.1 2D pseudo-Lorentzian representing the point spread, here corresponding to 2x2 pixels. Graphics is made using logarithmic function values with Mathematica (Mathematica, 1988-2005).

See Appendix B for details of the implementation. Every simulated impact is spread out over the reflection box using this PSF. We determined, by comparing with many observed reflections, that γ≈0.6 pixels gave realistic profiles on our Nonius KappaCCD detector. This corresponds to a FWHM, FW@1% and FW@0.1% of the PSF of 50 µm, 300 µm and 650 µm respectively, where one pixel is 110 µm. Fig. 2.2 shows the effect of including the point spread for a strong reflection.

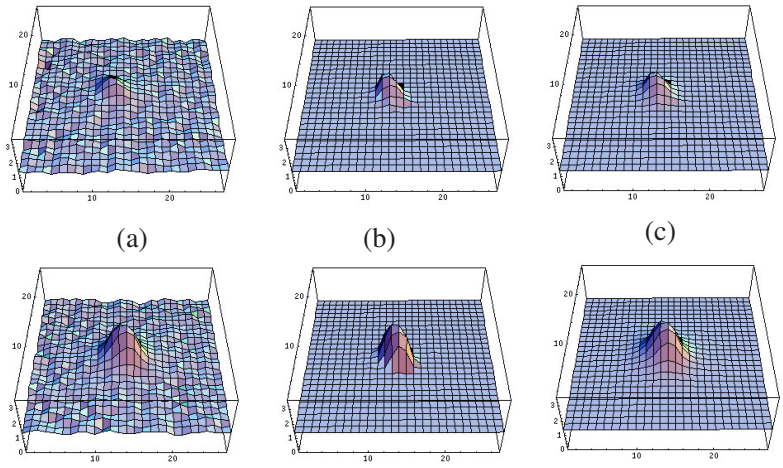


Fig.2.2 Profiles in two consecutive frames (top and bottom): a) observed profile, b) simulated profile and c) simulated profile + point spread. Graphics is made using logarithmic function values with Mathematica.

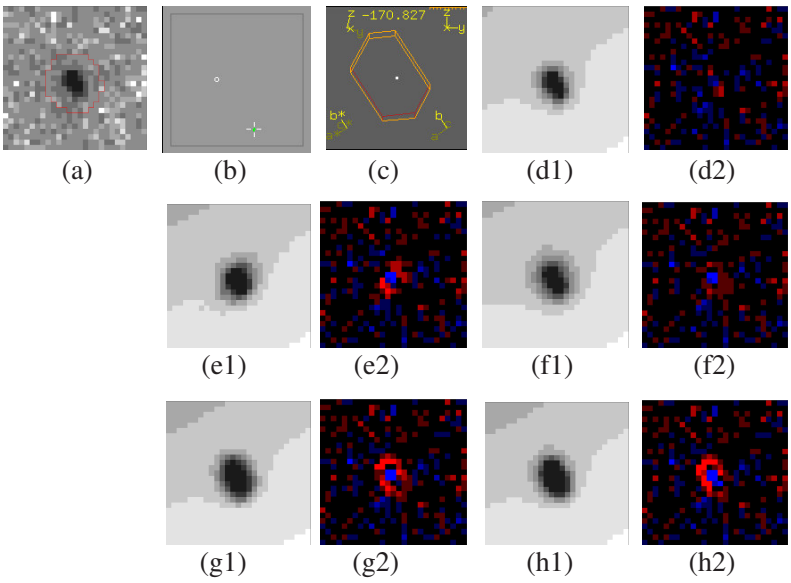


Fig. 2.3 Effect of parameter choice on reflection profiles. (a) Observed profile of a reflection with I/σ~50; (b) position on the detector; (c) face indexed crystal model viewed from the direction of the X-ray beam; (d1) model profile with optimal parameters: mosaicity μ=0.2°, pointspread γ=0.5 pixels, focus distance = 150 mm, resulting in fom_{peak} =1.05; (d2) difference (a) -(d1); (e1) μ=1.0° (fom_{peak} =1.43); (e2) difference (a)-(e1); (f1) γ=1.0 pixels (fom_{peak} =1.20); (f2) difference (a)-(f1); (g1) focus distance = 50 mm,(fom_{peak} =2.27); (g2) difference (a)-(g1); (h1) crystal size multiplied by a factor 2.0 (fom_{peak} =2.51); (h2) difference (a)-(h2). The observed and model profiles are coloured on a grey scale using logarithmic intensities; the difference profiles are coloured blue and red for positive and negative differences on a linear scale by Δ/σ units. For the definition of fompeak see section 2.2.3.

Fig. 2.3 demonstrates the effect of a change in one of the parameters determining the distributions (**F**,**K**,$\mathbf{S_m}$,λ) and the point spread width γ. To predict a reflection profile it is necessary to find good parameters for each of the instrumental constants and crystal properties. The dimensions of the focus and the wavelength spectrum (e.g. $K_{\alpha 1}$, $K_{\alpha 2}$ for a home source or a monochromatic wavelength for synchrotron radiation, each provided with a small dispersion) are more or less known beforehand. The divergence of the primary beam depends on the instrumental set-up. The focus distance and the point-spread of the detector have to be determined once for a particular instrument. We often have a microscope image of the crystal or even a face-indexed description, obtained with COLLECT (Nonius, 1999), that we can use in EVAL15. This leaves only the mosaicity of the crystal as the parameter to be established.

2.2.3 Parameter optimization

The predicted profile is taken as a normalized standard profile, used in minimization of the residual:

$$\chi^2 = \sum_{i=1}^{N} w_i \left[\rho_i - JP_i - \sum_{m}^{M} J_m P_{im} - ax_i - by_i - c \right]^2 , \qquad \textbf{(2.5)}$$

where N is the total number of pixels in the reflection box, ρ_i is the observed photon count, P_i is the normalized predicted profile value at pixel i, x_i and y_i are the horizontal and vertical pixel coordinates, J is the scale factor between the standard and the observed profile such that the integrated intensity $I=\Sigma_i JP_i$, and a, b and c define a plane describing the local background. The weights w_i are the inverse of σ_i^2 (Leslie, 1999). Assuming a Poisson distribution of counting errors, the standard deviation $\sigma_i = \rho_i^{1/2}$. M neighbouring reflections in the reflection box have their own profile P_m and scale factor J_m, some of which may be significantly overlapping the main reflection. In this procedure overlapping neighbour reflections are automatically deconvoluted from the main reflection. A similar approach for including overlapping reflections was followed by Bourgeois *et al.* (1998). The parameters can be found by solving an overdetermined set of normal equations following from Eq. (5): $\mathbf{(A^TA).c=AT.t,}$ where **A** is a $Nx(M+4)$ x N matrix given by

$$\mathbf{A} = \begin{pmatrix} \frac{P_1}{\sigma_1} & \frac{P_{11}}{\sigma_1} & \cdots & \frac{x_1}{\sigma_1} & \frac{y_1}{\sigma_1} & \frac{1}{\sigma_1} \\ \frac{P_2}{\sigma_2} & \frac{P_{21}}{\sigma_2} & & & & \\ \cdot & & & & & \\ \cdot & & & & & \\ \frac{P_N}{\sigma_N} & \frac{P_{N1}}{\sigma_N} & \cdots & & & \end{pmatrix} \qquad \textbf{(2.6)}$$

and **c**, a vector of dimension *(M+4)*, represents the fitting coefficients and **t**, a vector of dimension *N*, contains the observations:

$$\mathbf{c} = \begin{pmatrix} J \\ J_1 \\ . \\ a \\ b \\ c \end{pmatrix}, \quad t = \begin{pmatrix} \frac{\rho_1}{\sigma_1} \\ . \\ . \\ . \\ . \\ \frac{\rho_N}{\sigma_N} \end{pmatrix}. \tag{2.7}$$

The normal matrix $\mathbf{A^T A}$ can become singular or almost singular when the standard profiles P and P_m are almost linearly dependent. A general approach to solve the numerical instability and to choose a reasonable solution is Singular Value Decomposition (Nash, 1990). In SVD the matrix **A** is decomposed into $\mathbf{U.W.V^T}$, whereby W is a diagonal matrix made up of the square roots of the eigenvalues of the normal matrix. The best solution to the normal equations is $\mathbf{c=V.[diag\ (1/W_j)].U^T.t}$. It follows that the coefficients $\mathbf{c}_j$ are given by:

$$c_j = \sum_{i=1}^{M} \left(\frac{\mathbf{U}_i \bullet t}{W_i} \right) V_{ji} \ . \tag{2.8}$$

If the singular value W_i is (close to) zero the normal matrix is singular. The corresponding $1/W_i$ has to be set to zero (Press *et al.*, 1986). This algorithm also provides a variance-covariance matrix with

$$\sigma^2(c_j) = \sum_{i=1}^{M} \left(\frac{V_{ji}}{W_i} \right)^2 \ , \ \mathrm{cov}(c_j, c_k) = \sum_{i=1}^{M} \left(\frac{V_{ji} V_{ki}}{W_i^2} \right). \tag{2.9}$$

The variance of the main reflection is then given by:

$$\sigma_I^2 = \sigma_J^2 \left(\sum_{i=1}^{N} P_i \right)^2 . \tag{2.10}$$

In this procedure we automatically obtain the intensities and variances of neighbouring (overlapping) reflections in the box too. Even in case the overlap with neighbouring reflections is insignificant, the profiles of the neighbours are still important to calculate a proper background. The covariance of the main reflection and an overlapping neighbour tells us if we can reliably split the two or that we should sum the intensities of the overlapping reflections. If the value of χ^2 in Eq. (5) is large then the standard profile does not give a good fit. In EVAL15 we use the following figure-of-merit to indicate the quality of the fit:

$$fom_{box} = \left[\frac{\sum_{i=1}^{N} w_i (\rho_i - \rho_i^{calc})^2}{N - N_p} \right]^{1/2}, \qquad \textbf{(2.11)}$$

where N_p is the number of fitting parameters, usually *(M+4),* and $\rho_i^{calc} = JP_i + \Sigma_m J_m P_{im} + ax_i + by_i + c$. In a similar way we calculate fom_{peak} and fom_{bg}, where the summation runs over the pixels in the peak and those in the background, respectively. For this purpose the peak area is defined by those pixels that receive at least 0.3% of the total number of calculated impacts.
All papers describing profile-fitting algorithms conclude that the reflection positions should be known accurately, both for the learning and for the fitting procedure. EVAL15 will optimise the position (horizontal, vertical and rotational impact coordinates) of the reflection by minimizing fom_{box} using the Simplex method (Nelder & Mead, 1965; Press *et al.*, 1986). This can only be done reliably for reflections that are strong enough. If the impact positions of a collection of reflections with similar ω-values are larger than is acceptable, it is advisable to carry out a post-refinement and start the procedure all over.

2.2.4 Standard deviations and gain of the detector

Every detector converts the X-ray photons into an electronic signal that is read out and stored in an image file. The detective quantum efficiency (DQE) is a measure of the efficiency with which photons are detected and of the noise performance of the detector. It is defined as the signal-to-noise ratio of the output signal divided by that of and the input signal. For an ideal detector this ratio would be 1.0. In practice many factors reduce this number, like phosphor absorption efficiency, window transmission, phosphor noise factor, read-out noise, dark current and detector gain (Phillips *et al.*, 2002). The definition of *gain* varies in the literature. Here we will use *gain* as the number of ADU's (analogue-to-digital units) per X-ray photon. In EVAL15 all pixel intensities are divided by the *gain*, if this number is available from the header of the image files; otherwise it can be input manually. This manipulation obviously has no effect on the relative intensities. After dividing by *gain*, the best estimate of the standard deviation for each pixel-intensity is then obtained using Poisson statistics. It is obvious that because the DQE is lower than 1.0 the true I/σ will be smaller than what results from Poisson statistics. Below we discuss the relevance of applying the correct gain value. The background intensity of a reflection box is represented as a plane with parameters *a*, *b* and *c* (see above). Noise causes deviations between fitted and observed background pixel-intensities that are measured by fom_{bg}. These deviations are expected to follow a Gaussian distribution and fom_{bg} should be near 1.0 if the correct value for *gain* is used such that Poisson statistics applies. If fom_{bg} shows a large deviation from 1.0 in the various reflection boxes, this could be an indication for a wrong *gain* value. This reasoning assumes that ADC and/or dark current are effectively removed from the background so that it only consists of X-ray scattering. Popov & Bourenkov (2003) elaborate on the various contributions to standard deviations from summation integration. These can be described by a second order polynomial in *I*. The zeroth order coefficient is related to the incoherent background scattering, dark current and read-out noise. In EVAL15 the noise originating from

dark current and read-out is included by the parameter *bgnoise* that is estimated from dark images. The first order term in Popov's approach is the standard deviation due to Poisson counting statistics of the integrated intensity. Both effects are introduced as weights in the least-squares fit by using $\sigma_i=[(\rho_i+bgnoise^2]^{1/2}$ resulting in the standard deviation σ_I (Eq. (10)) of the integrated intensity *I*. The second order term is the contribution of instrument errors. Both systematic errors and errors in the profile will unavoidably lead to misfits of profiles especially at large I/σ. If the fom_{peak} values is larger than 1.0 the deviation between model and observation is larger than expected and obviously the model is not completely correct; somehow this should be expressed in the estimated standard deviation of the intensity. Two approaches seem justified to adapt the standard deviations σ_I of the integrated intensities. *1)* Multiply σ by fom_{peak}. A similar approach is followed by Leslie (1999). *2)* Leave it to the scaling program, in our case SADABS (Sheldrick, 1996) to find an error model for the standard deviations from the internal root-mean-square deviations $\sigma_{int}=[\Sigma_i(I_i-<I>)^2/(N-1)]^{1/2}$ of equivalent reflections. Both options are implemented in EVAL15.

2.3 Results and discussion

2.3.1 The EVAL15 graphical display

After finding the **R**-matrix with DIRAX (Duisenberg, 1992) and refinement of variables determining the reflection positions with PEAKREF (Schreurs, 1999), integration boxes are extracted, one for each separate reflection (of typically 27 pixels x 27 pixels x 5 frames), from the images using the 'datcol' procedure in VIEW (Schreurs, 1998) or with the help of the GUI in COLLECT. The size of the boxes should normally be sufficiently large to contain the complete reflection and a fair portion of background, though this is less critical in EVAL15 as it can also integrate incomplete reflections. Fig. 2.4 shows the graphical display of EVAL15 for one reflection. The top left panel shows successive observed ω-slices. In the second row the resulting profile from a sample of 10,000 impacts calculated using Eq. (3) is shown. In the third row the point-spread function is applied; then the scale factor for the profile and the background parameters are determined and in the fourth row the resulting model is displayed. Finally the difference between observation and model is shown in a red/blue colour scale. The right side of the window contains information on the position of the reflection: the resolution, θ and relative duration as well as the central horizontal, vertical and rotational impact coordinates. The difference between the original and final impact is shown on the right panel (impact vs. finalimp). The type of distributions (Lorentzian, Gaussian or block) and the numerical values for relevant parameters are shown as well. The values *I*, σ and I/σ corrected for Lorentz (Milch & Minor, 1974) and polarisation and the *fom's* can be found a few lines lower.

The lower left part of the window shows EVAL14-contours. In the case shown, the shape of the crystal was obtained through face indexing and it is shown by default in the orientation at diffracting position seen from the X-ray source.

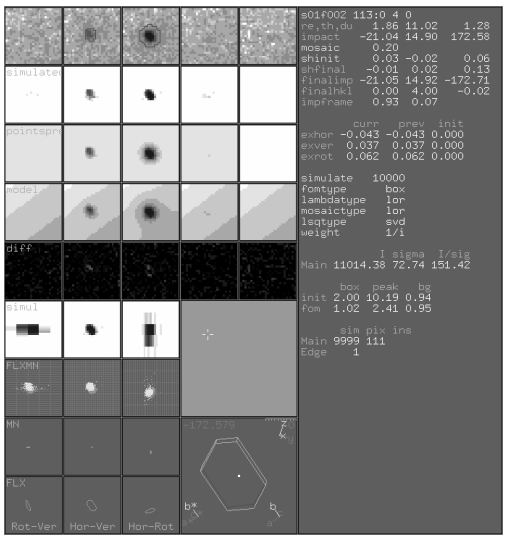


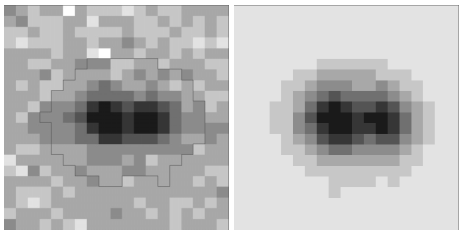
Fig.2.4 The EVAL15 graphical display

2.3.2 Predicted profiles for notoriously difficult cases

In this section we demonstrate the performance of EVAL15 for reflection data that present most integration software packages with significant difficulties.

$K_{\alpha 1}$, $K_{\alpha 2}$ -splitting

Reflections are notably split at higher resolution due to the $K_{\alpha 1}$, $K_{\alpha 2}$ radiation from sealed tubes or rotating anodes. Since EVAL15 uses both wavelengths in the simulation, with a ratio of 2:1, accurate profiles are obtained as can be seen in Fig.2 5.



(a) (b)

Fig. 2.5 $K_{\alpha 1}$, $K_{\alpha 2}$-splitting for a reflection at 0.78 Å resolution. (a) Observed and (b) simulated profile.

Fine slicing

Pflugrath (1999) discusses the possible advantages of fine slicing, i.e. images recorded over a rotation angular range significantly smaller than the effective mosaic

spread.* The advantage could be a lower X-ray background per image, fewer saturated pixels, fewer spatial overlaps and better positional accuracy of the reflection after post-refinement. Inherent disadvantages are that the intensity of a reflection is spread over a larger number of pixels (reducing the signal/noise), read-out noise is accumulated over several images and the process is more demanding in terms of disk space, goniometer hardware, shutter synchronization and the scaling procedure. As EVAL15 integrates 3D reflection boxes the treatment of fine sliced data is straightforward. Fig. 2.6 shows that, similar to what was observed by Pflugrath (1999), the shape of partial reflections can be different from frame to frame. The profile-fitting algorithm in EVAL15 is in no way hampered by these differing shapes; in fact the profiles are predicted accordingly.

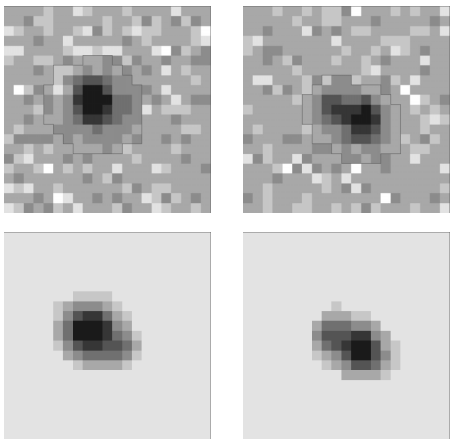


Fig. 2.6 Observed (top) and simulated profile (bottom) of a reflection on two successive frames.

Overlapping reflections

Overlapping reflections due to long cell axes or twin lattices can be deconvoluted even if up to 90% of their pixels overlap (Xian *et al.*, 2009b). The intensity and standard deviation for all reflections in the reflection box are estimated, though we only use that of the main reflection for which the box is made. The neighbour's intensity will be integrated in a separate neighbour reflection box. The profiles of the neighbours are obtained by the same sets of sampled rays that are impacted to the pixels centred near the neighbour predicted position (Fig. 2.7). The indexing programs DIRAX (Duisenberg, 1992) or CELL_NOW (Sheldrick, 2005) are particularly suited to find interfering lattices.

* The rotation range of a reflection is determined by the size and mosaic spread of the crystal, the wavelength dispersion, the beam divergence and the Lorentz factor [
Helliwell, J. R., Ealick, S., Doing, P., Irving, T. & Szebenyi, M. (1993). *Acta Cryst. D* **49**, 120-128.].
The relative duration used in EVAL15 is defined as the Lorentz factor divided by $2\sin\theta$ and thus is the duration relative to a reflection passing through the Ewald sphere in the equatorial plane when the rotation axis is perpendicular to the primary beam.

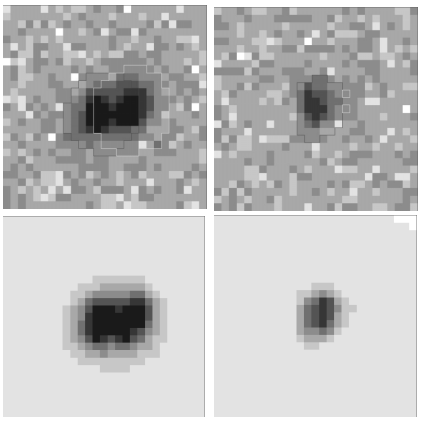


Fig. 2.7 Overlapping reflections from twin lattices can be simulated and their relative intensities are obtained through the SVD algorithm. Two consecutive frames of the observed reflection box (top) are shown. The model profiles (bottom) are made using the fitted relative intensities (I/σ=77.6 for hkl-main = 1 4 7 and I/σ=67.9 for hkl-neighbour = 1 4 -7).

2.4 Standard deviations

The error model used in SADABS for the standard deviations is $\sigma_e=K[\sigma_I^2+(g<I>)^2]^{1/2}$, where g (typically ~0.02 for EVAL data) accounts for (systematic) instrument errors (McCandlish *et al.*, 1975). Using this expression and the fact that we expect σ_{true} to be equal to $fom_{peak}\cdot\sigma_I$ we can write: $fom_{peak}=\sigma_{true}/\sigma_p\approx K[1+(gI/\sigma_I)^2]^{1/2}$. Fig. 2.8 shows that the EVAL15 fom_{peak} and the standard deviation multiplication factor of SADABS are strongly correlated for a test data set (for details see Xian *et al.* (2009a)) and thus fom_{peak} accounts for a large portion of the instrument errors indicated by SADABS. Minimization of fom_{peak} for a selected set of strong reflections (I/σ>20) turns out to be a good guide in finding the optimal profile prediction parameters and reduces the contribution of the profile part to the value of g in SADABS (see a separate paper (Xian *et al.*, 2009a) for a recipe to find the best profile prediction parameters).

The values of *gain* and *bgnoise* may not be known exactly. We have examined the consequence of the choice of these values on the estimation of the standard deviations σ and the intensities. KappaCCD test data of a crystal of an organo-metallic compound were integrated and the gain was initially set to 1.5, the value given in the header. However, from the average fom_{bg} we estimated it to be 1.2. Table 2.1 shows that the SADABS error model parameters change by changing *gain* and *bgnoise*, but the SHELXL refinement results were not significantly different except for the weighting scheme. This suggests that the integrated intensities have been slightly changed.

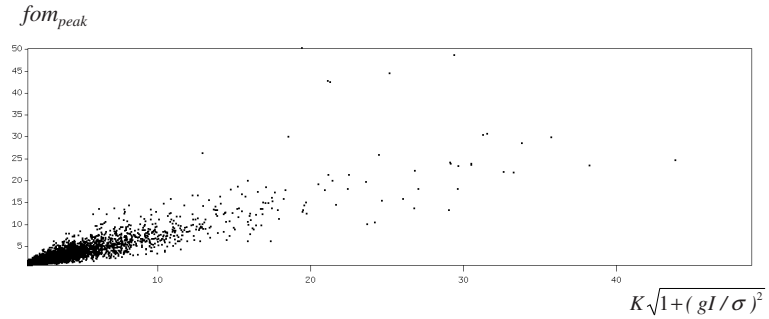


Fig. 2.8 Linear relationship between fompeak and the correction factor for standard deviations obtained with SADABS. Graphics made with ANY (Schreurs, 2007)

We also investigated the effect of multiplying the standard deviations with fom_{peak}. Since every reflection has its own fom_{peak} the I/σ values are sometimes changed considerably (Fig. 2.9). This has little consequence on the refinement. However, the error model parameters and the weights in the refinement become similar for the different *gain* and *bgnoise* values (Table 2.1) and the value for g is near 0.0. It can be concluded that multiplying the standard deviation with fom_{peak} reduces the sensitivity to the values of *gain* and *bgnoise* and gives better estimates of the true standard deviations before going into SADABS.

The choice of the profile model clearly matters for the *fom*-values and the refinement residuals as is seen in Table 2.1. In case the mosaic spread is chosen too small (0.2°) the results are significantly worse.

Most refinement programs (e.g. SHELXL (Sheldrick, 1997), Crystals (Watkin *et al.*, 2000)) establish a weighting scheme for the intensities or structure factors, not only to account for additional experimental errors, but also for model errors. Normal probability plots (Abraham & Keve, 1971) (Fig. 2.10) indicate that the standard deviations are underestimated as was shown earlier by Zhurov et al. (2008) for area detector data. In refinement of data up to a resolution of 0.77 Å usually the model errors are substantial so that a weighting scheme is essential. Using the square root of the refinement weights instead of the σ's, the normal probability plots behave much better. It follows that the estimated standard deviations of EVAL15 as such only contribute little to the weights. However, large values of the parameters in the weighting scheme are an indication that errors in the integrated intensities are substantially larger than what is expected from the σ's. The weights in the refinement for high resolution structures in programs like JANA (Petricek *et al.*, 2000) and XD (Koritsanszky *et al.*, 2003), where the model errors are small, are taken to be $1/\sigma^2$. A correct estimation of the standard deviations would be profitable in such cases. We believe that the use of fom_{peak} in combination with a scaling program like SADABS will give reliable standard deviations.

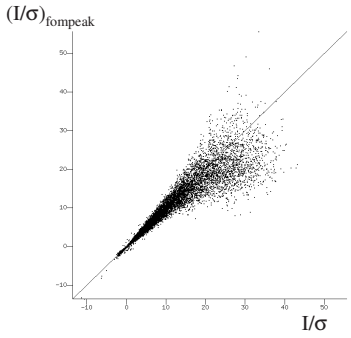


Fig. 2.9 After applying SADABS the I/σ values are changed considerably by multiplying the initial σ with fom_{peak}.

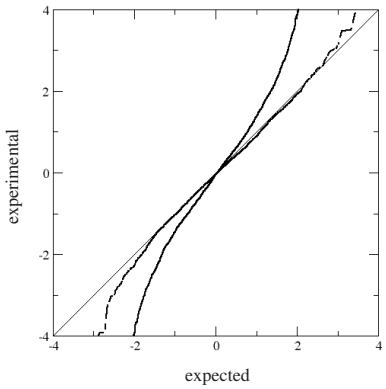


Fig. 2.10 Normal probability plots for $(Fobs^2\text{-}Fcalc^2)/\sigma$ (solid line) and $w^{1/2}(Fobs^2\text{-}Fcalc^2)$ (broken line). The standard deviations were estimated using $\sigma_I * fom_{peak}$.

2.5 Conclusions

In this paper we show that it is possible to make an ab initio prediction of reflection profiles as found in X-ray diffraction area detector data. The EVAL15 profile prediction method needs only a modest number of physically realistic parameters to simulate reflection profiles. We have shown that high-quality profiles are obtained, also in more demanding cases such as fine sliced data, $K_{\alpha 1}$,$K_{\alpha 2}$-splitting and overlapping reflections. Moreover, deviation of the profiles from what is expected on the basis of the physical parameters gives insight into unusual crystal properties or instrumental peculiarities. The simulated profiles are successfully applied in a profile fitting analysis to obtain accurate integrated intensities. It is relatively easy to include additional properties of the crystal (like anisotropic mosaic spread and lattice distortion) or of the instrument (like focussing mirrors or newly developed detectors) into the ray tracing simulation. This fully flexible approach has the potential to solve many difficult diffraction problems. EVAL15 has the possibility to work with

multiple lattices (several R-matrices) and can include overlapping neighbor reflections (from the same or from different lattices) in the least-squares procedure.

Appendix A. Distributions

When certain parameters are required to have specific non-uniform distributions care has to be taken to ensure unbiased samples. This applies e.g. to the wavelength distribution within the spectrum or to the distribution of mosaic orientations. If

$$p(y) = \frac{dx}{dy} p(x) \qquad \textbf{(2.12)}$$

is the required distribution and $p(x)$ represents a uniform distribution of deviate x, it follows that

$$F(y) = \int p(y) dy = x. \qquad \textbf{(2.13)}$$

The transformed deviate $y(x)=F^{-1}(x)$ has the required distribution (Press *et al.*, 1986). For instance a 1D-Gaussian distribution can be obtained from $y=erf^{-1}(x)$. Press et al. also describe how a 2D-Gaussian distribution can be obtained. Selection of random points inside a circle and transformation of their Cartesian coordinates $(v1,v2)$ to random polar coordinates gives the uniform deviate $R=v_1^2+v_2^2$, the radius from the centre of the distribution. Then $y_1 = \sqrt{-2ln(R)}v_1/R$ is sampled according to a Gaussian distribution and represents the radial coordinate along a 1D-section through the 2D-Gaussian distribution. The azimuthal coordinate results from rotation over a random angle between 0 and 2π or from the second coordinate $y_2 = \sqrt{-2ln(R)}v_2/R$.
In a similar way a 2D-Lorentzian distribution can be obtained. In this case taking

$$y_1 = \sqrt{\left|\left(\frac{1}{2\sqrt{R}}\right)^2 - 1/4\right|}\,\frac{v_1}{R} \qquad \textbf{(2.14)}$$

and proceeding in a similar way gives a 2D-Lorentzian distribution (Fig. 2.11).

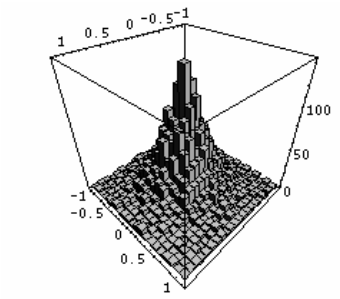


Fig. 2.11 A 2D-Lorentzian distribution obtained from uniform sampling of random coordinates (v1,v2) and transformation to the corresponding coordinates (y1,y2.) A histogram of (y1,y2) is plotted. The standard deviation σ was chosen to be 1.0. Graphics is made using logarithmic function values with Mathematica.

Appendix B. Point-spread function

The point-spread of a detector can be conveniently described by a pseudo 2D-Lorentzian. It is based on a normal 1D-Lorentzian where the variable is replaced by two Cartesian variables. Integration of this function to infinity only converges when the determinant of the Jacobian matrix for transformation of the Cartesian to polar coordinates is included, which is $1/\sqrt{x^2+y^2}$. In addition, we have included a factor $(½\gamma)^2$ to avoid the denominator from becoming zero. The resulting point-spread function is:

$$\mathbf{PSF}(x,y) = \frac{\gamma}{4\pi\left[\left(x^2+y^2\right)+\left(\frac{1}{2}\gamma\right)^2\right]^{1½}} \quad , \qquad \textbf{(2.15)}$$

x and y are the distances in horizontal and vertical directions measured from the centre of the impacted pixel. The intensity accumulated in a pixel is thus spread over neighbouring pixels at (x,y) away from its centre. It is wrong to assume that the contribution to a target pixel depends only on the distance of its centre to the centre of the source pixel. In fact this function has to be evaluated as an integral over the surface of the whole target pixel. The integral is given by

$$\frac{1}{2\pi}\tan^{-1}\left[\frac{\frac{2}{\gamma}xy}{\sqrt{\left(\frac{1}{2}\gamma\right)^2+x^2+y^2}}\right] \qquad \textbf{(2.16)}$$

and the four corners of the pixel are taken as the integration limits.

Table 2.1 SHELXL refinement of data of α-Tris(2,4-pentanedionato-κ^2-O,O')cobalt(III) (Chrzanowski *et al.*, 2007) using different estimations of σ.

mosaic spread	gain/ bgnoise	<fom_{peak}>	<fom_{bg}>	σ_{EVAL15}	K	g	R_1	wR_2	S	Δρ	weights a/b
0.7	1.5 3.0	1.2319	0.6860	σ_p	1.18	0.0318	0.0316 0.0506	0.0787	1.041	0.33 -0.60	0.0315 0.87
0.7	1.5 3.0			σ_p* fom_{peak}	1.50	0.0003	0.0328 0.0510	0.0813	1.032	0.32 -0.62	0.0292 1.29
0.7	1.5 1.87	1.3418	0.7962	σ_p	1.32	0.0246	0.0318 0.0503	0.0785	1.028	0.36 -0.59	0.0299 1.13
0.7	1.5 1.87			σ_p* fom_{peak}	1.45	0.0003	0.0324 0.0503	0.0804	1.032	0.34 -0.60	0.0286 1.30
0.7	1.5 0.67	1.434	0.894	σ_p	1.43	0.0214	0.0335 0.0507	0.0800	1.040	0.35 -0.59	0.0271 1.44
0.7	1.5 0.67			σ_p* fom_{peak}	1.43	0.0003	0.0330 0.0503	0.0803	1.051	0.34 -0.58	0.0265 1.32
0.7	1.2 0.83	1.599	0.994	σ_p	1.54	0.0205	0.0342 0.0511	0.0811	1.049	0.35 -0.59	0.0262 1.54
0.7	1.2 0.83			σ_p* fom_{peak}	1.43	0.0003	0.0331 0.0504	0.0795	1.050	0.31 -0.56	0.02511. 39
0.2	1.2 0.83	1.778	1.008	σ_p	1.53	0.0307	0.0391 0.0562	0.0982	1.046	0.64 -0.66	0.0353 1.79

$R_1=\Sigma||F_{obs}|-|F_{calc}||/\Sigma|F_{obs}|$
$wR_2 = \{\Sigma[w(F_o^2-F_c^2)^2]/\{\Sigma[w(F_o^2)^2]\}\}^{1/2}$
$S= \{\Sigma[w(F_o^2-F_c^2)^2]/(n-p)\}\}^{1/2}$, where n = number of reflections, p = number of refined parameters
Δρ=maximum and minimum difference density
weights $w = 1/[\sigma^2(F_o^2)+(aP)^2+bP]$, where $P = (F_o^2+2F_c^2)/3$

Acknowledgements

The authors thank Dr. Rob Hooft of Bruker AXS for fruitful discussions and Dr. Albert Duisenberg for critically reading the manuscript. Financial support of The Netherlands Technology Foundation STW, project number UPC 6148, is gratefully acknowledged.

References

Abrahams, S. C. & Keve, E. T. (1971). *Acta Cryst.* A**27**, 157-165.
Arndt, U. W. (1986). *J. Appl. Cryst.* **19**, 145-163.
Bourgeois, D., Moy, J. P., Svensson, S. O. & Kvick, A. (1994). *J. Appl. Cryst.* **27**, 868-877.
Bourgeois, D., Nurizzo, D., Kahn, R. & Cambillau, C. (1998). *J. Appl. Cryst.* **31**, 22-35.
Bruker (1998). *SAINT Version 4.0.* Bruker AXS BV.
Chrzanowski, L. S.v, Lutz, M. & Spek, A. L. (2007). *Acta Cryst.* C**63**, m283-m288.
Diamond, R. (1969). *Acta Cryst.* A**25**, 43-55.
Duisenberg, A. J. M. (1992). *J. Appl. Cryst.* **25**, 92-96.

Duisenberg, A. J. M., Kroon-Batenburg, L. M. J. & Schreurs, A. M. M. (2003). *J. Appl. Cryst.* **36**, 220-229.
Ford, G. C. (1974). *J. Appl. Cryst.* **7**, 555-564.
Helliwell, J. R., Ealick, S., Doing, P., Irving, T. & Szebenyi, M. (1993). *Acta Cryst.* D**49**, 120-128.
Kabsch, W. (1988). *J. Appl. Cryst.* **21**, 916-924.
Koritsanszky, T., Howard, S. T., Richter, T., Macchi, P., Volkov, A., Gatti, C., Mallinsson, P. R., Farrugia, L., Su, Z. & Hansen, N. K. (2003). *XD.* Free University of Berlin.
Leslie, A. G. W. (1999). *Acta Cryst.* D**55**, 1696-1702.
Mathematica (1988-2005). *Mathematica.* Version 5.2. Wolfram Research, Inc.
McCandlish, L. E., Stout, G. H. & Andrews, L. C. (1975). *Acta Cryst.* A**31**, 245-249.
Milch, J. R. & Minor, T. C. (1974). *J. Appl. Cryst.* **7**, 502-505.
Nash, J. C. (1990). *Compact Numerical Methods for Computers: Linear Algebra and Function Minimalisation,Ch 3. The Singular-Value Decomposition and Its Use to Solve Least-Squares Problems*, 2nd ed. Bristol, England: Adam Hilger.
Nelder, J. A. & Mead, R. (1965). *Computational Journal* **7**, 308-313.
Nonius (1999). *COLLECT.* Delft, The Netherlands.
Otwinowski, Z. & Minor, W. (1997). *Macromolecular Crystallography, Pt A* **276**, 307-326.
Oxford Diffraction (2008). *Crysalis*. Oxford Diffraction Ltd. UK.
Petricek, V., Dusek, M. & Palatinus, L. (2000). *JANA2000.* Institute of Physics, Czech Academy of Sciences, Prague, Czech Republic.
Pflugrath, J. W. (1999). *Acta Cryst.* D**55**, 1718-1725.
Phillips, W. C., Stewart, A., Stanton, M., Naday, I. & Ingersoll, C. (2002). *J Synchrotron Radiat* **9**, 36-43.
Popov, A. N. & Bourenkov, G. P. (2003). *Acta Cryst.* D**59**, 1145-1153.
Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1986). *Numerical Recipes*. Cambridge: Cambridge University Press.
Schreurs, A. M. M. (1998). *VIEW.* University of Utrecht, The Netherlands.
Schreurs, A. M. M. (1999). *PEAKREF.* University of Utrecht, The Netherlands.
Schreurs, A. M. M. (2007). *ANY.* University of Utrecht, The Netherlands.
Sheldrick, G. M. (1996). *SADABS.* University of Goettingen, Germany.
Sheldrick, G. M. (1997). *SHELXL.* University of Goettingen, Goettingen, Germany.
Sheldrick, G. M. (2005). *CELL_NOW.* University of Goettingen, Goettingen, Germany.
Watkin, D. J., Prout, C. K., Carruthers, J. R., Betteridge, P. W. & Cooper, R. I. (2000). *Crystals.* Chemical Crystallography Laboratory, University of Oxford.
Xian, X., Schreurs, A. M. M. & Kroon-Batenburg, L. M. J. (2009a). *to be submitted*.
Xian, X., Schreurs, A. M. M. & Kroon-Batenburg, L. M. J. (2009b). *to be submitted*.
Zhurov, V. V., Zhurova, E. A. & Pinkerton, A. A. (2008). *J. Appl. Cryst.* **41**, 340-349.

Chapter 3

Accuracy of X-ray diffraction data integrated by the EVAL15 profile prediction method*

Xinyi Xian, Antoine Schreurs & Loes Kroon-Batenburg

* To be submitted to the Journal of Applied Crystallography

Abstract

We recently introduced the new diffraction-data integration method EVAL15 (Schreurs *et al.*, 2009). It uses the principle of general impacts (Duisenberg *et al.*, 2003) to predict reflection profiles, using only a small number of physical parameters. By least-squares fit of the predicted profile to the observed profile, the reflection intensity is derived. EVAL15 has the potential of integrating complicated reflection data, such as due to from anisotropic crystal shape, anisotropic mosaicity, $K_{\alpha 1}$,$K_{\alpha 2}$-splitting and overlap due to multiple lattices or to a long cell axis. The aim of this paper is to assess the quality of EVAL15 data for a set of standard diffraction experiments, both for small molecule and protein crystals. A comparison is made with data integrated with EVAL14 (Duisenberg *et al.*, 2003), in particular to investigate if the EVAL15 profile method results in improved quality of the weaker data. We found that EVAL15 delivers a lower R_{merge}, a higher I/σ (especially in the highest resolution shell) and a lower R-value in the refinement. The EVAL15 data are also more successful in phasing. In a separate paper we will deal with overlap problems (Xian *et al.*, 2009).

3.1 Introduction

In recent years a rapidly increasing number of protein crystal structures was solved, while the research focus has shifted to large multimer complexes. Their crystals tend to be smaller and of lower quality. A consequence is that structure determination relies on the accurate integration of a large number of weak intensities. It has been common practice to use summation integration in small molecule crystallography, as is done in EVAL14. However, weak reflections are integrated more accurately by profile fitting and it is estimated that the standard deviations can be lowered by as much as a factor $2^{1/2}$ (Diamond, 1969). The intensity of a reflection is derived by the least-squares fit of a learned standard profile to the observed profile. This standard profile is obtained by averaging profiles of strong reflections. Due to geometrical deformations, the profiles vary across the detector plate. This implies that standard profiles have to be learned from nearby neighbour reflections. The disadvantage of this procedure is immediately obvious in the case that neighbour reflections are weak or overlapping with other reflections.

Several programs use profile fitting for the integration of crystallographic diffraction data such as the HKL package (Otwinowski & Minor, 1997), MOSFLM (Leslie, 1999), d*TREK (Pflugrath, 1999) and XDS (Kabsch, 1988).

We recently introduced a new data integration method EVAL15 (Schreurs *et al.*, 2009), that does not need neighbour reflections to derive a standard profile. With a few physical crystal and instrument parameters, like mosaicity, crystal shape, beam divergence and wavelength, a three-dimensional standard profile is predicted by simulating general impacts. This standard profile is then used in a profile fitting algorithm using Singular Value Decomposition (SVD) (Press *et al.*, 1986). The principle of general impacts (Duisenberg *et al.*, 2003) was first applied in EVAL14 in order to predict reflection boundaries that can be used in summation integration.

In this paper the quality of EVAL15 data is evaluated with four data sets. We compare scaling, merging, refinement results and in some cases also phasing. The quality of weak reflections is compared with that of EVAL14.

3.2 Theory

In EVAL15 three-dimensional reflection profiles are simulated by producing 10000 generated impacts by tracing X-rays from different focus, crystal and mosaic sample points and for different wavelength values of the radiation. The simulated profile is then used as a standard profile in a least-squares minimization. A least-squares minimization of

$$\chi_{lsq}^{\ 2} = \sum_{i=1}^{N} \left[\frac{\rho_i - JP_i - \sum_{m}^{M} J_m P_{im} - ax_i - by_i - c}{\sigma_i} \right]^2 \qquad \textbf{(3.1)}$$

leads to a set of linear algebraic equation that are solved by SVD. N is the total number of pixels in the reflection box, ρ_i the observed photon count, P_i is the value of the simulated profile at pixel i, J is the scale factor for the main reflection, J_m are the scale factors for possible neighbour reflections m, x_i and y_i are the horizontal and vertical pixel coordinates, a, b, and c are the parameters describing a planar background in the reflection box , $\sigma_i = (\rho_i^{calc})^{1/2}$ and

$$\rho_i^{calc} = JP_i + \sum_{m}^{M} J_m P_{im} + ax_i + by_i + c\,. \qquad \textbf{(3.2)}$$

The intensity of the main reflection is obtained by

$$I = \sum_{i=1}^{N} JP_i \ . \qquad \textbf{(3.3)}$$

The SVD algorithm also delivers a variance-covariance matrix, from which the standard deviation in I can be obtained. For the integration of the intensity the accuracy of the predicted reflection position is vital (Pflugrath, 1999), because even sub-pixel differences can produce a significant difference in the calculated intensity. The predicted reflection position is shifted such that it coincides as well as possible with the observed reflection. This is done by minimizing $fom_{box,}$ a figure-of-merit measuring the quality of the profile fit:

$$fom_{box} = \left[\frac{\sum_{i=1}^{N} w_i (\rho_i - \rho_i^{calc})^2}{N - N_p} \right]^{1/2} , \qquad \textbf{(3.4)}$$

where the weights w_i are chosen to be $1/\sigma_i^2$. The number of fitted parameters N_p is usually 4, i.e. the scaling parameter J and three background parameters. The area of the main reflection is defined by those pixels, which receive at least 0.3% of the impacts. Similarly to fom_{box} we calculate fom_{peak} where the summation runs over pixels in the peak area. After integration of the data, SADABS (Sheldrick, 1996) was used to scale and to estimate the correct σ_c's, by $\sigma_c = K (\sigma_p^2 + (gI)^2)^{1/2}$, where K and g are derived by minimizing

$$\chi_{\text{int}}^{\ 2} = \left\langle \frac{N\sum_{i}^{N}\left|I_i - \langle I\rangle\right|^2}{(N-1)\sum_{i}^{N}\sigma_i^2} \right\rangle . \qquad \textbf{(3.5)}$$

σ_p are largely determined by Poisson statistics and come out of the least-squares fit, and g is generally seen as a measure of the instrument error (Popov & Bourenkov, 2003).

3.3 Methods

Four standard data sets were integrated. Two data sets of a small molecule crystal were collected at different temperatures. The other two were anomalous diffraction data of protein crystals.
The images were processed with the EVAL-software suite consisting of DIRAX (Duisenberg, 1992) for indexing, PEAKREF (Schreurs, 1999) for refinement of unit cell dimensions and detector, crystal and goniometer offsets and VIEW (Schreurs, 1998) for viewing and generation of reflection boxes at predicted reflection positions (typically 27x27 pixels and 5 oscillation increments). The integration is performed by EVAL15 (using the profile prediction method) or alternatively by EVAL14 (by applying summation integration using accurate contours) and ANY was used for statistical analysis and graphical display of the data. Subsequently, the data are scaled and absorption correction is applied by SADABS (Sheldrick, 1996). At that stage outliers can be rejected. The data statistics after scaling is calculated with Xprep (Bruker AXS, Madison, USA). Anomalous structure factor differences were calculated by either XPREP or TRUNCATE (French & Wilson, 1978). Phases were obtained by either SOLVE/RESOLVE (Terwilliger, 2003a) or SHELXD/E (Sheldrick, 2008). ARP/wARP (Perrakis *et al.*, 2001; Perrakis *et al.*, 1999; Morris *et al.*, 2002) or RESOLVE (Terwilliger, 2003b) were used for automatic model building. The structures were refined against the data by SHELXL (Sheldrick & Schneider, 1997; Sheldrick, 2008) or REFMAC (Murshudov *et al.*, 1997).

3.4 Results and Discussion

3.4.1 Derivation of the profile prediction parameters

Data sets of crystalline α-Tris (2,4-pentanedionato-κ^2O,O')cobalt(III) at 240 K (data set I) and 150 K (data set II) were collected in-house on a Nonius KappaCCD, with rotating anode MoK_α radiation (Chrzanowski *et al.*, 2007). The crystal has space group $P2_1/c$ and the cell dimensions are slightly smaller at lower temperature (Table 3.1). The asymmetric unit contains 4 molecules. The reflection profile is a convolution of broadening effects due to the crystal size, mosaic spread, point spread, beam divergence and wavelength distribution. Parameters describing the crystal shape were derived by accurate measurement with the video microscope (COLLECT (BrukerAXS, 1999)) and the wavelength distribution was assumed to be known. This leaves three unknown parameters to be determined: the point spread, the beam divergence and the mosaic spread. In EVAL15 a pseudo-Lorentzian point spread function with a width γ is implemented (Schreurs *et al.*, 2009). The beam divergence is modelled by putting a virtual focus (usually 0.3x0.3 mm^2) at a given distance: the

focus distance. In the program the corresponding parameter *focus dist* is given in mm. The mosaicity μ is modelled by a Gaussian or a Lorentzian distribution, where 99% of the distribution is within μ=3σ, or a block-shaped distribution where values between -μ and μ are equally probable. We started the integration with an initial profile model for the data at 240 K (Table 3.1). We used the default value for *focus dist* of 500 mm and optimised the γ value of the point spread by minimizing fom_{box} and fom_{peak} interactively for a set of randomly chosen reflections. However, we found that the parameters *focus dist* and γ are strongly correlated in the sense that a large focus distance (small divergence) can be compensated by a large point spread and vice versa. We sought an efficient way of optimising the model profiles. Strong reflections (I/σ>100) are the most sensitive to changes in the parameters. We anticipated that the reflection shape at low θ and low duration would be mainly determined by the point spread, as the divergence is still small and the sensitivity to the mosaicity is small. Reflections are spread over a large rotation range (ω) at high duration, i.e. they occur at several successive images. The duration is very sensitive to the mosaic spread; therefore we intended to optimise this parameter on high duration reflections. Table 3.2 lists the results of optimisations for strong reflections with low durations. At several values of γ we searched for the value of *focus dist* that gave the lowest fom_{box} and fom_{peak}. Combination of a small point spread γ with a small *focus dist* is clearly favoured. Therefore, we have chosen to perform a new integration with γ=0.5 and *focus dist*=150.

Optimisation of reflection profiles of strong reflections at high duration pointed to a mosaicity μ of 0.1-0.3° with a Lorentzian distribution for data set (I) (Table 3.3). However, for data set (II) some reflections favoured μ=0.2°, while others needed 1.1-1.2°. Figs. 3.1-3.3 show the observed profiles for data set (I) and of data set (II). The latter was measured after cooling down to 150 K. Reflection 5 –3 0 has a similar profile for both (Fig. 3.1) and these can be modelled satisfactorily with a Lorentzian type mosaicity of μ=0.2°. However, reflections 1 –1 2 and -2 –1 4 clearly have a much broader profile in (II), and it turned out that a larger mosaicity with block-type mosaic spread fits much better (Figs. 3.2-3.3 and Table 3.3). The value of the mosaicity turns out to depend on the angle between the c-axis of the crystal and the (vertical) rotation axis such that it is larger when the angle is closer to 0°. This means that during cooling, the crystal cracked into several fragments that are rotated with respect to each other around the c-axis, which corresponds to the axis perpendicular to the crystal plate. In future we will describe such an anisotropic mosaic spread by a superposition of two distributions. For now we have settled on the compromise of a Lorentzian distribution with μ=0.7° for (II) and μ = 0.2° for (I).

Changing the model parameters has a modest consequence for weak reflections. The largest change is observed in data (II), where the difference between the initial and the improved model is the largest. However, the I/σ of the strong reflections are significantly larger with the improved model (Fig. 3.4). This implies that strong reflections are the most sensitive to changes in model parameters. The lower g and K as well as the lower R-value after scaling confirm that the models of both data (I) and (II) are indeed improved (Table 3.4). Note that the g-value of the improved model of (II) is larger than the g of (I). This is caused by the compromise we made with respect to the mosaic-model and may also be caused by instrument errors related to fine slicing. In Figure 3.5 an example of the improved model is shown in the EVAL15-display for data (II). This reflection has an I/σ larger than 1000, therefore it has a large fom_{peak} (24.62), whereas the average fom_{peak} in the data set is 1.599. Due to

instrument and profile errors fom_{peak} is always large for strong reflections. This error is taken into account by SADABS through the parameter g. EVAL15 produces better data in both cases judged from the scaling (Table 3.4), merging (Table 3.5) and refinement results (Table 3.6). The improvement compared to EVAL14 of the weak data in EVAL15 is evident from the lower R_{merge} and the higher I/σ's in the highest shell.

Table 3.1 Crystal and measurement parameters

	(I)	(II)	(III)	(IV)
cell dimensions:				
a (Å)	13.81	13.69	23.11	94.47
b (Å)	7.43	7.39	42.92	99.61
c (Å)	16.15	16.06	64.01	104.39
α (°)	90	90	90	90
β (°)	98.43	98.49	90	90
γ (°)	90	90	90	90
spacegroup	$P2_{1/c}$	$P2_{1/c}$	$P2_12_12_1$	$I222$
resolution (Å)	0.77	0.77	1.8	1.58
rotation (°) /frame	1	1	1	0.2
dist(mm)	40	40	125	180
λ (Å)	0.71073	0.71073	0.9793/0.9795/0.9763	1.54178
temperature (K)	240	150	100	100
profile parameters:				
crystal dimensions (mm)	face indexed crystal[i]	face indexed crystal[i]	0.05x0.1x0.1	0.2x0.2x0.2
mosaicity μ (°)	0.2/0.2[ii]	0.2/0.7[ii]	0.9	0.6
point spread γ(pixels)	0.8/0.5[ii]	0.8/0.5[ii]	0.879	1.0
focus dist/width/length (mm)	500/0.3/0.3/ 150/0.3/0.3[ii]	500/0.3/0.3/ 150/0.3/0.3[ii]	1000/0.1/0.1	150/0.3/0.3
$\lambda/\sigma_\lambda/w_\lambda$	0.71073/0.0001/2 0.71359/0.0001/1	0.71073/0.0001/2 0.71359/0.0001/1	0.9793/0.00052/1.0[iii] 0.9795/0.00052/1.0 0.9763/0.00052/1.0	1.54178/0.001/1.0

(I), (II)= α-Tris (2,4-pentanedionato-κ^2O,O')cobalt(III),
(III)= Staphylococcal complement inhibitor (SCIN), (IV)=Triplet of glucose-isomerase
[i]measured accurately with a video microscope: shortest dimension 0.2 mm, longest dimension 0.4 mm
[ii]initial model/improved model
[iii]the wavelength distribution of the peak/inflection/remote data

Table 3.2 Optimization of instrumental parameters with data (I)

low duration and low θ[i], high I/σ, Lorentzian type mosiaicity of μ=0.2°								
hkl (1 1 0), I/σ 560								
point spread γ	0.4	0.5	0.6	0.8	*focus dist*	120	500	280
focus dist[ii]	110	120	130	170	point spread γ[ii]	0.5	0.8	0.8
fom_{box}	1.87	1.83	1.97	2.44	fom_{box}	1.83	2.86	2.74
fom_{peak}	10.19	9.69	10.15	12.1	fom_{peak}	9.69	14.87	13.5
hkl (0 0 2), I/σ 630								
point spread γ	0.4	0.5	0.6	0.7				
focus dist[ii]	100	100	110	130				
fom_{box}	2.64	2.45	2.52	2.66				
fom_{peak}	16.88	14.74	14.53	14.59				
hkl (-2 0 0), I/σ 650								
point spread γ	0.4	0.5	0.6		*focus dist*	500		
focus dist[ii]	110	110	110		point spread γ[ii]	0.8		
fom_{box}	2.60	2.67	3.04		fom_{box}	5.01		
fom_{peak}	11.89	12.13	12.79		fom_{peak}	20.67		
hkl (3 0 -2), I/σ 290								
point spread γ	0.4	0.5						
focus dist[ii]	120	150						
fom_{box}	1.14	1.14						
fom_{peak}	4.84	4.53						

[i]duration is lower than 1.5 and θ is smaller than 5°, [ii]optimized parameter

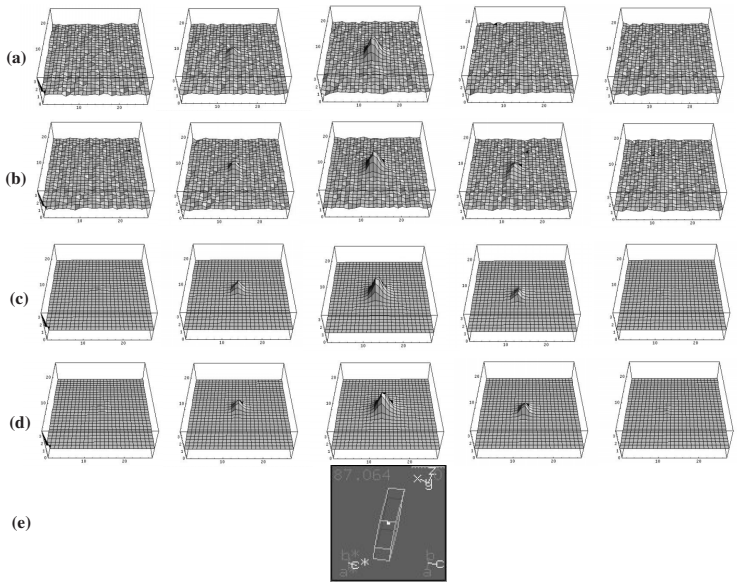


Fig. 3.1 Reflection profiles of logarithmic intensities drawn with Mathematica: (a) Observed reflection 5 –3 0 of data (I), and (b) of data (II), (c),(d) modelled with a Lorentzian type mosaic distribution with μ=0.2° 2° for (I) and (II) respectively and (e) the crystal orientation in reflecting position, seen from the primary beam.

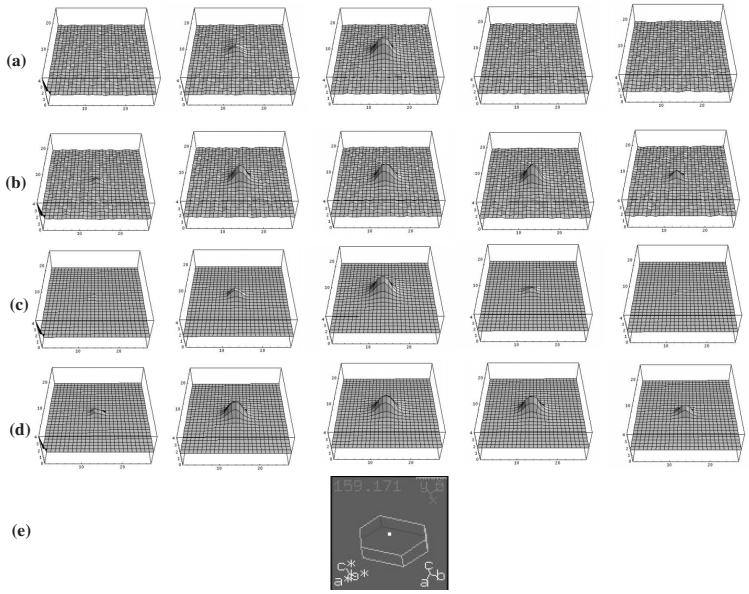


Fig. 3.2 Reflection profiles of logarithmic intensities drawn with Mathematica: (a) Observed reflection –2 –1 4 of data (I), and (b) of data (II), (c) modelled with a Lorentzian type mosaic distribution with μ=0. and (d) with a block-type mosaic distribution with μ=1.1°. (e) The crystal orientation in reflecting position.

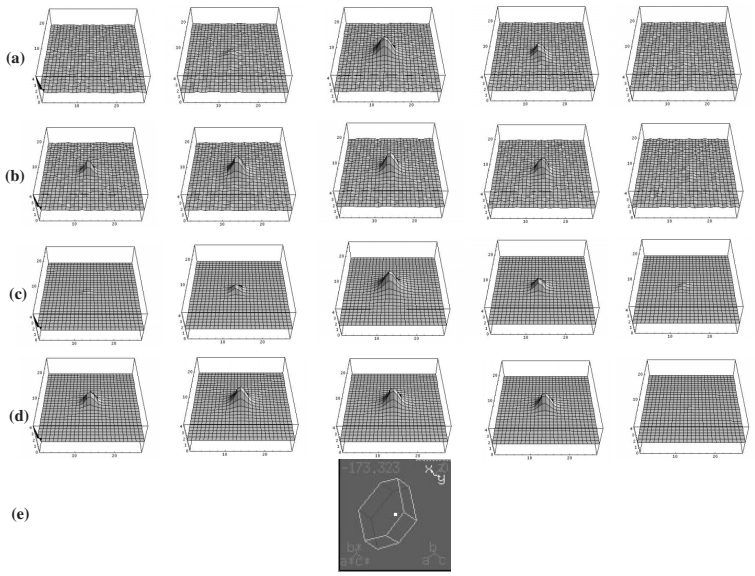


Fig. 3.3 Reflection profiles of logarithmic intensities drawn with Mathematica: (a) Observed reflection 1 -1 2 of data (I), and (b) of data (II), (c) modelled with a Lorentzian type mosaic distribution with μ=0.2° and (d) with a block-type mosaic distribution with μ=0.7°. (e) The crystal orientation in reflecting position.

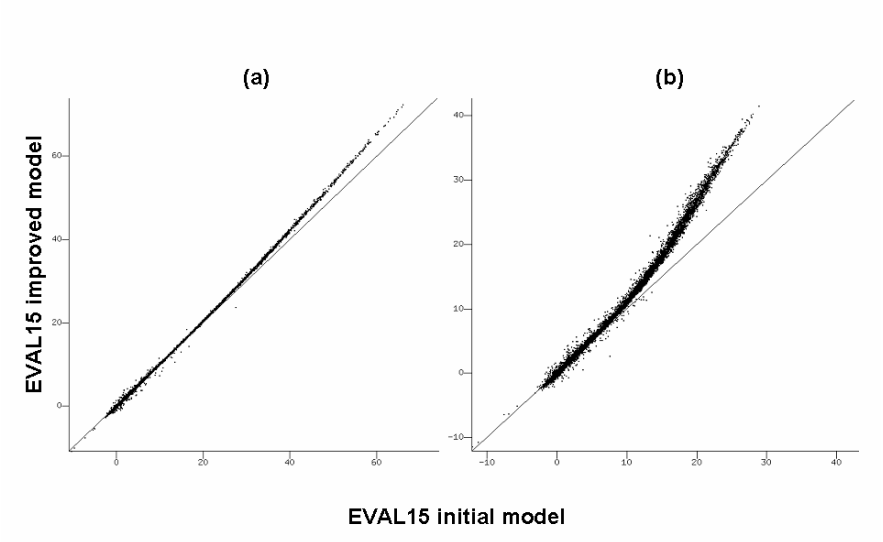


Fig. 3.4 Scaled unmerged I/σ of the improved model versus the initial model for (a) data (I) and (b) data (II).

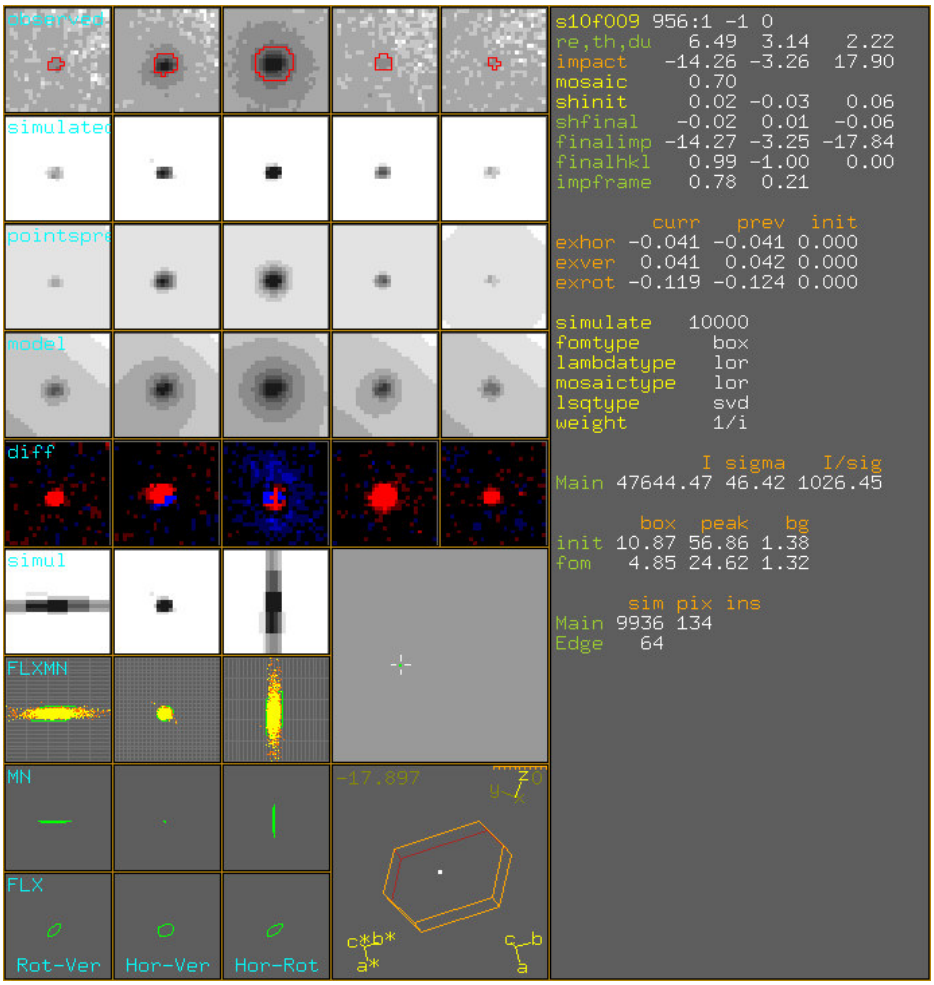


Fig. 3.5 EVAL15-display of the improved model of data (II). The main reflection is shown within the red contour, defined by the area of pixels that receive at least 0.3% of impacts. The final position of this reflection on the detector after refinement of the predicted position is displayed [finalimp]. fom_{box} shows that the difference of observed and calculated intensity is reduced by the refinement (compare the value at [init] to that at [fom]). The resolution in Å [re], relative duration [du], mosaicity value in ° [mosaic], the shift in the horizontal, vertical directions in mm and in the rotational direction in ° are shown at [shfinal]. 10000 impacts for the main reflection are generated, 64 impacts fall outside of the box [Edge]. The rest of the impacts fall into 134 pixels. The intensity [I] and σ [sigma] of the main reflection is shown. Left upper part of the display shows the observed reflection in successive ω-slices [obs]. Below that the simulated profiles, those including point spread and the scaled model including background are shown. The difference after the least-squares fit is also shown [diff], where the surplus of calculated pixel intensity is coloured in red and that of the observed is coloured in blue.

Table 3.3 Optimization of the mosaic value for data (I) and (II)

high duration*, high I/σ, *focus dist* of 150 mm, point spread γ of 0.5 pixels					
(I)					
hkl	(2 0 -2)	(-1 1 -2)	(4 -1 3)	(-3 0 6)	(-2 0 4)
I/σ	470	630	250	330	600
mosaicity μ (Lorentzian type)	0.1	0.3	0.3	0.2	0.2
fom_{box}	1.33	2.42	1.2	1.26	2.33
fom_{peak}	4.89	10.23	4.47	4.89	11.92
(II)					
hkl	(-2 -1 4)	(1 -1 2)	(5 -3 0)		
I/σ	470	500	320		
mosaicity μ (Lorentzian type)	1.2	1.1	0.2		
fom_{box}	5.18	3.51	1.06		
fom_{peak}	24.68	16.54	3.31		
mosaicity μ (block type)	1.1	0.7	0.2		
fom_{box}	2.92	3.25	1.13		
fom_{peak}	3.64	15.86	4		

*larger than ~3

3.4.2 Synchrotron protein MAD data

Diffraction data (data set III) of the Se-Met protein Staphylococcal complement inhibitor (SCIN) (Rooijakkers & Milder *et al.*, 2007) were collected on an ADSC detector at the ESRF Beam line ID14-EH4 in Grenoble, France. Multi-wavelength anomalous dispersion (MAD) data were obtained at λ= 0.9793 Å (peak), λ= 0.9795 Å (inflection) and λ= 0.9763 Å (remote). The inflection and remote data comprise 100 images and the peak data 200 images (recorded with the inverse beam method). There is no clear indication of radiation damage.
One SCIN molecule is contained in the asymmetric unit and the solvent content is 25%. The merging statistics of the three wavelengths is shown in Table 3.5. The average fom_{peak} of the optimal profile (profile parameters are shown in Table 3.1) is 1.4903. The instances of high fom_{peak} are mostly paired with high fom_{bg}, which occur in case of overflows or background scattering such as of ice. The following analyses were performed to compare the EVAL15 and EVAL14 data:
First, the statistics of the peak, inflection and remote data are compared. Secondly, the phasing is examined, by investigation of the anomalous signal, success rate of location of the heavy atom (HA)-sites, anomalous correlation coefficients, map correlation, and the difference between the experimental and the refined phases. Thirdly, the results of the automatic structure building in RESOLVE are compared. Finally, the quality of refinement of the complete model was assessed.

Statistics

The error model obtained with SADABS gives a lower K-value and a higher g-value for EVAL15, implying that the estimation of σ in EVAL15 is better for weak than for strong data (Table 3.4). Note that we made no effort to find the correct gain. A consequence is that g and K are larger than for (I) and (II). We have shown that an incorrect value of gain does not affect the quality of data after scaling with SADABS (Schreurs *et al.*, 2009). On average the I/σ from EVAL15 is higher for weak and lower for strong reflections (Fig. 3.6). It is a well known fact that the intensity of strong reflections are determined more accurately by summation integration (Bourgeois *et al.*, 1998). Close inspection learns that the intensities of weak

reflections are usually larger in EVAL14, but also their standard deviations. The reason for this lies in the summation integration method: intensities of random effects like zingers and cosmic rays also contribute to the net intensity and can make a significant contribution to weak reflections. We will see how this difference affects the success of phasing and the quality of the refined structure. The merging statistics in Table 3.5 shows that EVAL15 has a lower R_{merge} (especially the last shell), a higher number of reflections and consequently higher completeness and redundancy. The R-factor between EVAL15 and EVAL14 peak data is 0.034.

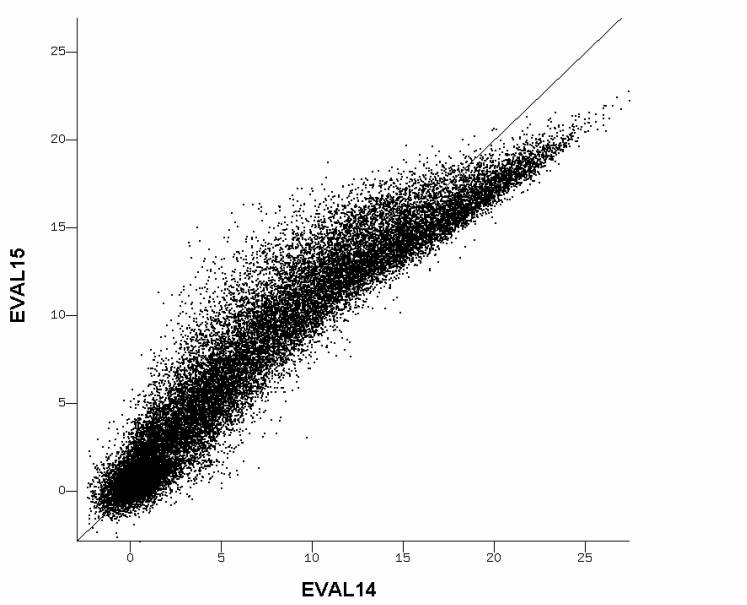


Fig. 3.6 Linear plot of the scaled unmerged peak I/σ of the EVAL15 data against the EVAL14 data.

Phasing quality

The anomalous signal-to-noise ratio $\Delta F/\sigma_{\Delta F}$ of the peak data as a function of the resolution is shown in Figure 3.7. The highest ratio for both data sets occurs at about 6Å, and the behaviour is similar for EVAL14 and EVAL15. We already know that the σ's are larger for weak high resolution data in EVAL14, so it must be concluded that the ΔF's are also larger in EVAL14. This is confirmed by the larger anomalous signal ΔF/F at higher resolution (Fig. 3.8). Up to a resolution of about 3.2 Å, <ΔF>/<F> is close to the expected ratio of

$$\frac{\langle |\Delta F| \rangle}{\langle F \rangle} = \frac{\sqrt{2N_A}\,\delta f''_A}{\sqrt{N_P}\,Z_{eff}} \quad \text{(Hendrickson \& Teeter, 1981),} \qquad \textbf{(3.6)}$$

which for SCIN is ~ 6%, with N_A = 2 Se atoms, N_P = 624 atoms, Z_{eff} = 6.7 electrons and $\delta f''_A$ = 3.85 electrons at λ = 0.9793 Å. Then at higher resolution it rises, due to an increasing uncertainty in the estimation of ΔF and F (Dauter *et al.*, 1999). We found

that in EVAL14 large ΔF's sometimes occur for weak reflections. These erroneously large Δ F's could later hamper the HA-location.

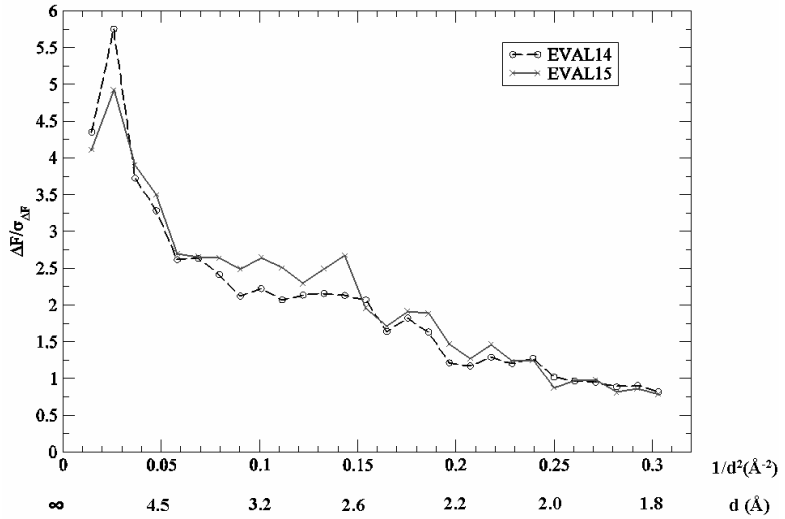


Fig. 3.7 The anomalous signal-to-noise ratio $\Delta F/\sigma_{\Delta F}$ of the peak data versus resolution.

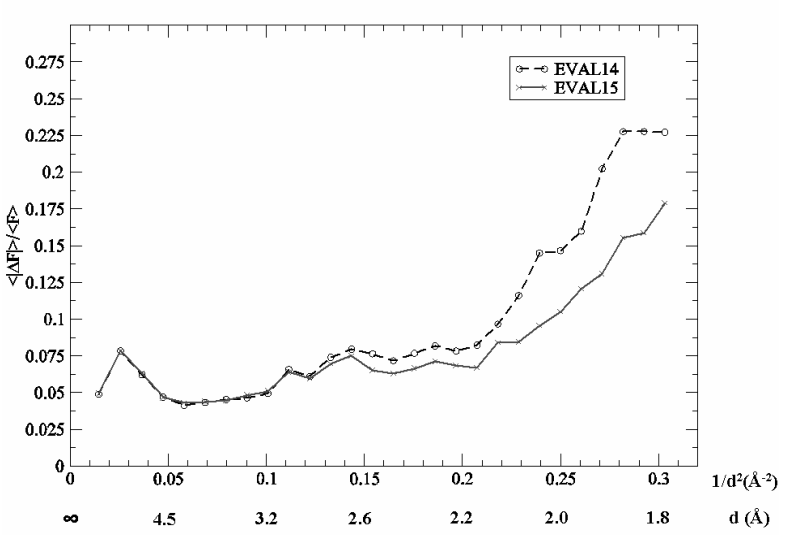


Fig. 3.8 <ΔF>/<F> as a function of resolution. The expected <ΔF>/<F> ratio is ~ 0.06 for SCIN. Up to 3.16 Å the experimental value is close to expected, beyond that the ratio increases due to errors in the estimation of ΔF and F.

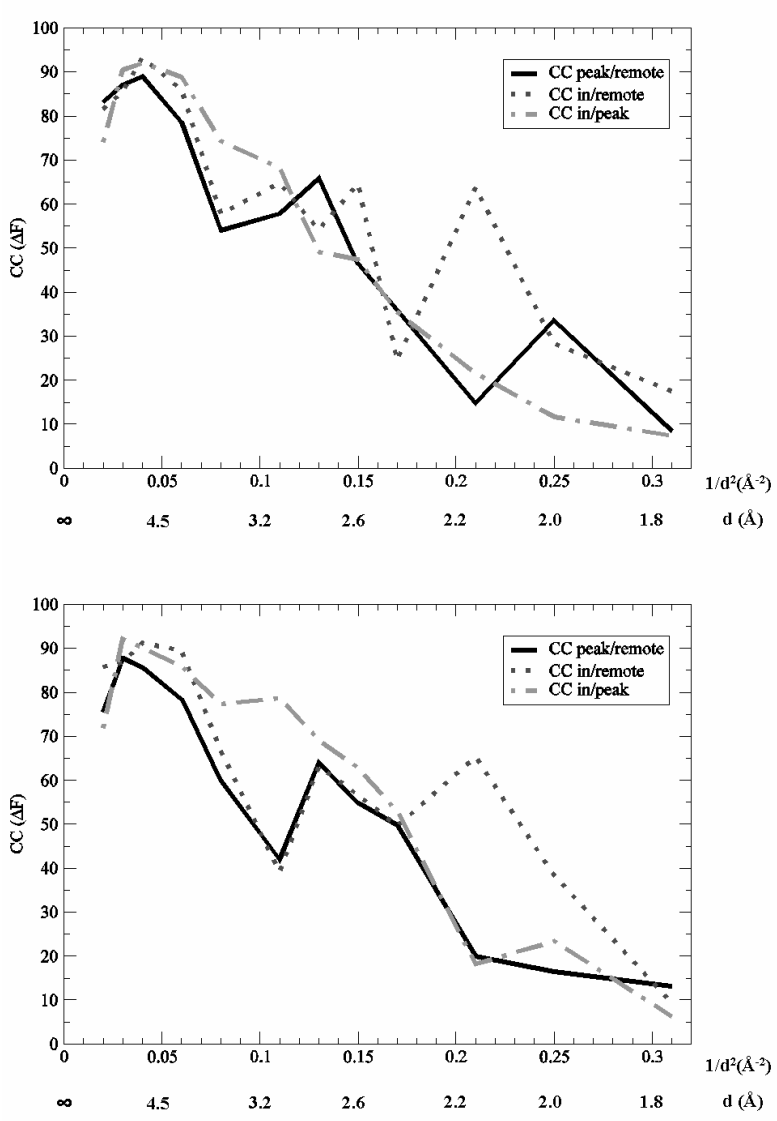


Fig. 3.9 The correlation between the anomalous signal for the peak, remote and inflection data of EVAL14 (upper) and EVAL15 (lower) versus resolution.

The resolution limit up to which a meaningful anomalous signal is available can be obtained from correlation plots of peak, inflection, and remote data. Schneider & Sheldrick found that once the correlation falls below 25 %, the anomalous differences are too unreliable to estimate phases (Schneider & Sheldrick, 2002). Correlation plots

of the signed anomalous differences (Fig. 3.9) show that the data are reliable to about 2.3 Å for both EVAL14 and EVAL15. We have calculated the success rate of finding the correct Se-sites from the MAD-data using SHELXD. The evaluation of the correct HA-sites was based on the PATFOM's and examination of coordinates. It is clear that EVAL15 is more successful, independent of the resolution cut-off (Table 3.7).
The increase of the success rate in EVAL14, by truncation at 2.3 Å, may be due to a smaller amount of erroneously large ΔF's, which is confirmed by comparison of the EVAL14 MAD Patterson maps, where the lower resolution map shows much less noise (Fig. 3.10). However, a resolution cut-off does not improve the success rate of the EVAL15 data (Table 3.7). This implies that no large errors in ΔF occur in EVAL15, as a consequence of the more accurate integration of weak reflections.

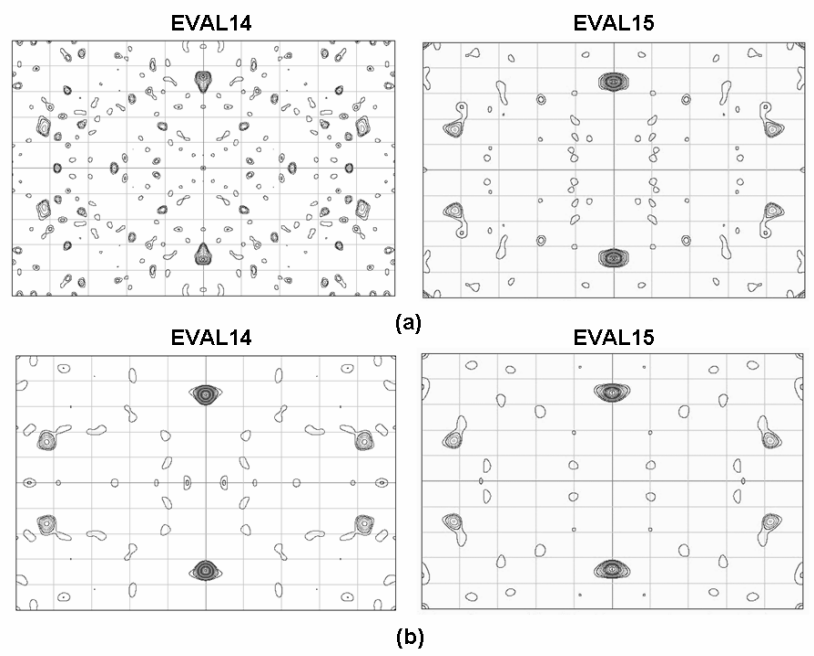


Fig. 3.10 Anomalous Patterson maps of MAD data (a) up to 1.8 Å and (b) up to 2.3 Å. The cross section is at x = 0.5 and the contour interval = 1.0 σ.

After location of the Se-sites, the correlation coefficient CC_E between E_o and E_c, is calculated with:

$$CC_E = 100\frac{(\sum wE_oE_c\sum w-\sum wE_o\sum wE_c)}{\sqrt{[\sum wE_{o^2}\sum w-(\sum wE_o)^2][\sum wE_{c^2}\sum w-(\sum wE_c)^2]}}, \qquad \textbf{(3.7)}$$

(Schneider & Sheldrick, 2002), where E_c is calculated from the Se-sites found with SHELXD and E_o is derived from the observed F_A's. The CC_E of the MAD data is 20.1% for EVAL15 and 16.7% for EVAL14. In RESOLVE the structure was built automatically and EVAL15 performs slightly better (Table 3.8).

We proceed by comparing the initial phases after RESOLVE with the final phases of the refined and complete model (Rooijakkers & Milder *et al.*, 2007) (Fig. 3.11). The phase errors are rather large due to the low solvent content, especially for the EVAL14 data. However, RESOLVE had no problems in building the model automatically. The larger phase error of EVAL14 causes the slightly worse results in RESOLVE_BUILD and corresponds to a lower reciprocal map correlation CC_{map} (Rice *et al.*, 2000) (Fig. 3.12), which is calculated with

$$CC_{map} = \frac{\sum_{hkl} fom_i \left|F_i^{hkl}\right| fom_j \left|F_j^{hkl}\right| \cos(\phi_i - \phi_j)}{\sqrt{\sum_{hkl} \left|fom_i F_i^{hkl}\right|^2 \sum_{hkl} \left|fom_j F_j^{hkl}\right|^2}} \quad . \qquad \textbf{(3.8)}$$

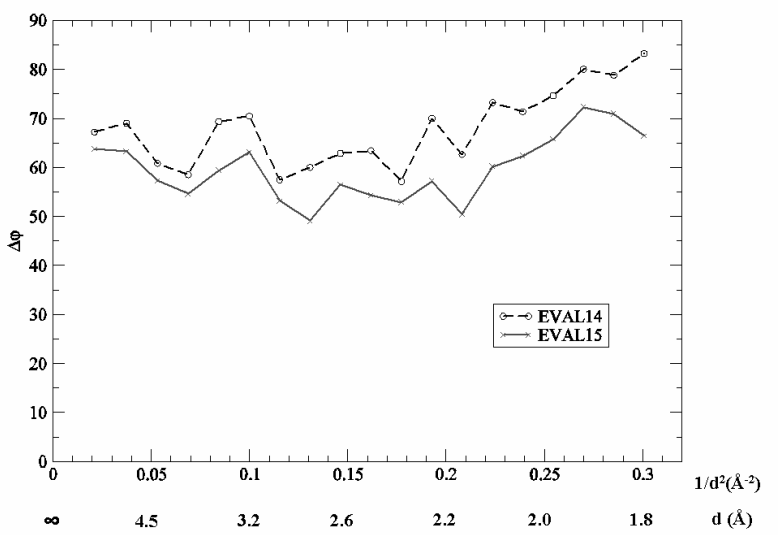


Fig. 3.11 The differences between the initial phases and the calculated phases derived from the final model versus resolution.

Table 3.4 Scale-statistics of SADABS

	(I) 240K			**(II) 150K**		
	EVAL14	EVAL15[i]	EVAL15[ii]	EVAL14	EVAL15[i]	EVAL15[ii]
g	0.0161	0.0177	0.0162	0.029	0.0424	0.0285
K	1.45	1.474	1.459	1.435	1.509	1.498
R	0.0452	0.0329	0.0319	0.0439	0.0414	0.0343
	(V)					
	peak		**inflection**		**remote**	
	EVAL14	EVAL15	EVAL14	EVAL15	EVAL14	EVAL15
g	0.0376	0.0603	0.0282	0.0576	0.0332	0.0605
K	1.703,1.755	1.376,1.391	1.945	1.472	1.89	1.492
R	0.0573	0.0516	0.0456	0.0441	0.0472	0.0471
	(VI)					
	EVAL14	EVAL15				
g	0.0391	0.022				
K	1.391	1.841				
R	0.0491	0.0480				

[i]initial model, [ii]improved model

Table 3.5 Data statistics after scaling and merging

	R_{merge} (Friedels merged)[iii]	R_{merge} (Friedels not merged)[iii]	completeness	I/σ[iii]	total data	unique data[iii]	redundancy
(I) 240K	resolution 0.77 (0.90-0.77)Å						
EVAL14	0.045 (0.326)	-	99.9	18.20 (3.44)	19558	4054 (1470)	4.82
EVAL15[i]	0.033 (0.173)	-	99.8	20.55 (5.39)	19273	4053 (1469)	4.75
EVAL15[ii]	0.032 (0.176)	-	99.9	21.32 (5.50)	19301	4055 (1472)	4.76
(II) 150K	resolution 0.77 (0.90-0.77)Å						
EVAL14	0.044 (0.171)	-	99.8	18.13 (5.98)	19264	3988 (1404)	4.82
EVAL15[i]	0.043 (0.123)	-	99.8	16.15 (7.11)	18699	3985 (1400)	4.68
EVAL15[ii]	0.037 (0.115)	-	99.8	19.82 (8.00)	18903	3985 (1400)	4.73
(III)	Resolution 1.8(1.9-1.8)Å						
Peak							
EVAL14	0.083 (0.587)	0.057 (0.597)	99.4	18.83 (3.85)	40704	6301 (891)	6.42
EVAL15	0.075 (0.390)	0.052 (0.355)	99.5	20.28 (5.52)	44165	6310 (889)	6.96
Inflection							
EVAL14	0.060 (0.695)	0.038 (0.471)	98.4	13.48 (1.97)	20104	6241 (880)	3.17
EVAL15	0.056 (0.423)	0.044 (0.355)	99.3	13.72 (2.90)	22291	6294 (893)	3.52
Remote							
EVAL14	0.065 (0.839)	0.047 (0.771)	97.8	12.04 (1.60)	19118	6204 (873)	3.01
EVAL15	0.062 (0.484)	0.047 (0.413)	99.2	12.32 (2.44)	21675	6290 (893)	3.42
(IV)	resolution 1.58(1.70-1.58)Å						
EVAL14	0.054 (0.419)	0.047 (0.371)	87.5	16.29 (2.13)	301371	58880 (9232)	5.12
EVAL15	0.044 (0.359)	0.039 (0.311)	89.2	18.74 (2.46)	313006	59916 (9446)	5.22

[i]initial model, [ii]improved model ,[iii]number in parenthesis refer to the highest resolution shell

Table 3.6 Refinement results

	R(strong/all)	#(strong/all)	R_{free} (strong/all)[iii]	#(strong/all)	$\Delta\rho_{max/min}$(Å^{-3})	S	wR_2	WGHT(a b)	K_1
(I) 240K									
EVAL14	0.034/0.058	2578/3762	-	-	0.31/-0.42	1.031	0.0786	(0.031 0.827)	0.957
EVAL15[i]	0.037/0.059	2910/3696	-	-	0.24/-0.43	1.105	0.0922	(0.036 0.893)	0.919
EVAL15[ii]	0.036/0.059	2815/3763	-	-	0.24/-0.37	1.091	0.0914	(0.035 0.738)	0.942
(II) 150K									
EVAL14	0.040/0.079	2812/3698	-	-	0.28/-0.33	1.035	0.0916	(0.037 0.064)	0.956
EVAL15[i]	0.053/0.063	2808/3762	-	-	0.61/-0.81	1.057	0.1199	(0.050 2.516)	0.958
EVAL15[ii]	0.036/0.059	2974/3696	-	-	0.25/-0.38	1.078	0.0889	(0.033 0.857)	0.951
(III) peak	R	data#	R_{free}	#R_{free}	fom	rms(Bond Å)	rms(angle °)	rms(chiral)	B(Å^2)
EVAL14	0.205	5973	0.24	407	0.851	0.006	0.791	0.063	22.408
EVAL15	0.201	5985	0.234	409	0.847	0.007	0.786	0.061	25.258
(IV)	R(strong/all)	#(strong/all)	R_{free}(strong/all)	#(strong/all)	$\Delta\rho_{max/min}$(Å^{-3})	S	wR_2	WGHT(a b)	K_1
EVAL14	0.165/0.187	41112/50562	0.188/0.216	2114/2615	0.42/-0.31	2.251	0.4477	0.20 0.00	0.986
EVAL15	0.181/0.204	42223/51438	0.207/0.238	2165/2676	0.46/-0.31	2.604	0.4957	0.20 0.00	0.894

[i]initial model, [ii]improved model

[iii] R_{free}(5% data)

$R=\Sigma||F_{obs}|-|F_{calc}||/\Sigma|F_{obs}|$

$\Delta\rho_{max/min}$=maximal and minimal rest-density

S = Goodness of fit = $\{\Sigma[w(F_o^2-F_c^2)^2/(n-p)]\}^{1/2}$, where n = number of reflections, p = number of parameters refined

$wR_2 = \{\Sigma[w(F_o^2-F_c^2)^2/\{\Sigma[w(F_o^2)^2]\}^{1/2}$

WGHT = weighting scheme: $w = 1/[\sigma^2(F_o^2)+(aP)^2+bP]$, where $P = (F_o^2+2F_c^2)/3$

$K_1=F_{obs}/F_{calc}$ (this is the ratio in the highetst resolution shell)

fom= Figure of merit

rms=root mean square deviation

B= temperature factor

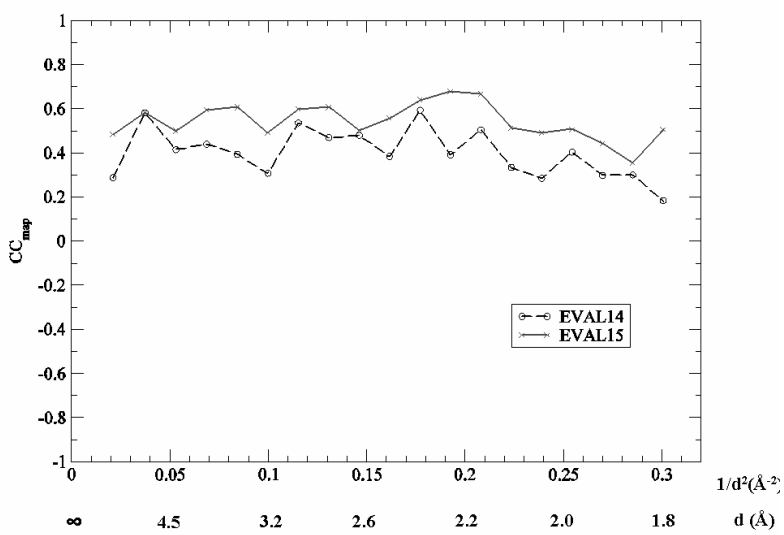


Fig. 3.12 The reciprocal map correlation of the initial electron density map after density modification with that of the map computed from the final refined model versus resolution.

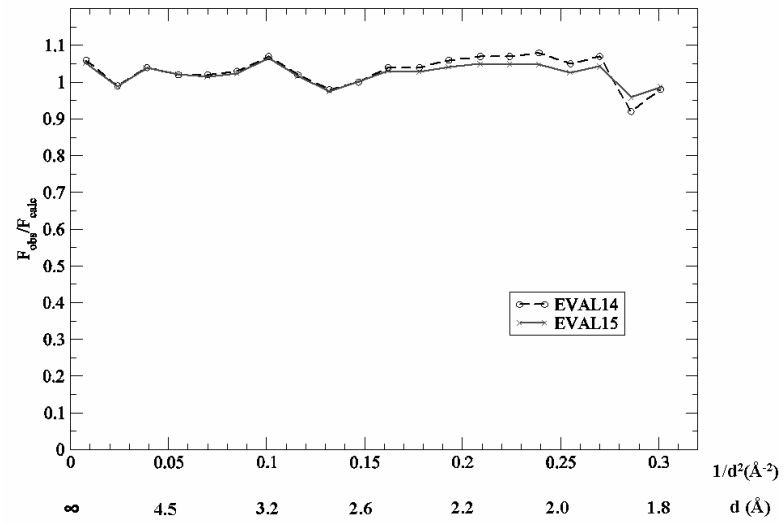


Fig. 3.13 The ratio of Fobs/Fcalc using constant bulk solvent in non-protein region versus resolution. The ratio is close to one, but for the higher resolution shells it deviates more, especially for the EVAL14 data.

Refinement

Up to now the data quality of EVAL15 is better than that of EVAL14, not only in terms of data statistics, but also in terms of anomalous signal and success rate of HA-location. Finally, we examined the refinement results of REFMAC.
The final model has R and R_{free} of 0.20 and 0.23 for EVAL15 and 0.20 and 0.24 for EVAL14. The B-factor of EVAL14 is lower than for EVAL15, caused by the larger intensities of the weak data. Plotting Fobs/Fcalc over the resolution range reveals that the ratio of EVAL15 is closer to one than that of EVAL14 (Fig. 3.13) at higher resolution, which confirms the higher accuracy of the weak data in EVAL15. It can be concluded that with the EVAL15 approach, the data quality is improved in terms of data collection statistics, phasing quality and refinement. The improvement in the phasing quality is remarkable.

3.4.3 In-house protein SAD data

X-ray data of Glucose isomerase (data set IV, kindly provided by Madhumati Sevvana), crystallized as a three component twin, were collected on a Bruker Cu-rotating anode with Osmic focussing mirrors and a SMART 6000 4K detector up to 1.58 Å (Sevvana, 2006). A total of 9362 images were collected in low-, medium- and high-resolution passes. The low-resolution images were used for the determination of the unit cell. The data were integrated successively by EVAL15 and EVAL14, using an artificially smaller contour for the latter in order to be able to split overlapping reflections. The deconvolution of the overlapping reflections carried out by EVAL15 will be explained in detail in another paper. As the overlap percentage is small, we can use the data set of one of the three components to assess the quality of the EVAL15 integration.
Glucose isomerase has 385 amino acids, of which 7 methionines and one cysteine. The merging statistics (Table 3.5) shows that EVAL15 has a higher completeness and larger I/σ, larger number of unique data and redundancy and a lower R_{merge} than EVAL14. The anomalous scattering is mainly caused by two manganese ions, but also the sulphur atoms contribute to the anomalous signal. The anomalous signal-to-noise ratio $\Delta F/\sigma_{\Delta F}$ of the data as a function of the resolution is shown in Figure 3.14, where the signal is slightly higher in the EVAL15 data.

Table 3.7 success rate of HA-location*

(III)		
	EVAL14	EVAL15
mad (2.3 Å)	38/251	62/251
mad (1.8 Å)	57/601	353/601
(IV)		
	EVAL14	EVAL15
sad (3.5 Å)	58/165	105/165

*Number of correctly located HA-sites/total number of trials

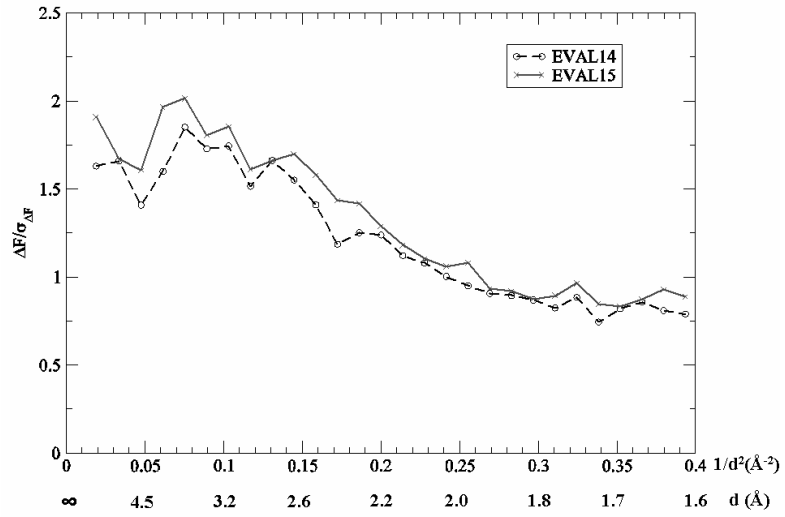


Fig. 3.14 The anomalous signal-to-noise ratio $\Delta F/\sigma_{\Delta F}$ of the glucose isomerase data versus resolution.

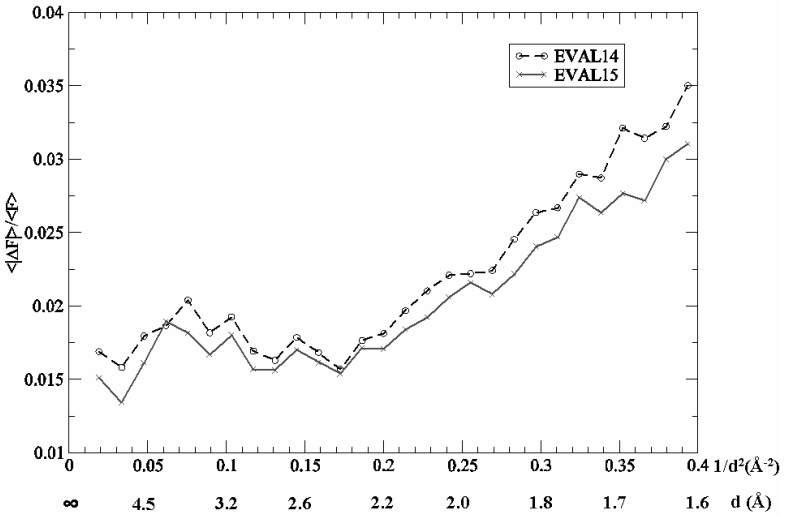


Fig. 3.15 <ΔF>/<F> as a function of resolution.

Up to a resolution of about 2.1 Å, <ΔF>/<F> is close to the expected ratio of ~ 2% with $\delta f''_A$ = 2.74 electrons at λ = 1.54178 Å for Glucose isomerase (Fig. 3.15). Again it rises at higher resolution. However, the anomalous signal-to-noise ratio is rather weak beyond 3.5 Å, which led us to phase with data truncated at 3.5 Å. The anomalous Patterson maps of EVAL14 and EVAL15 are very similar at this resolution (not shown here) and the success rate of correctly located HA-sites is twice

as large with EVAL15 (Table 3.7). The correlation coefficient CC_E between E_o and E_c is 32.51% for EVAL14 and 37.14% for EVAL15. A slight difference between the anomalous electron density maps can be seen in Figure 3.16.

Table 3.8 Automatic model building statistics

(III)		
Resolve_build	EVAL14	EVAL15
Resolution	1.8 Å	1.8 Å
R	0.36	0.33
Rfree	0.36	0.35
map correlation	0.28	0.32
total residue	75	72
accepted residue	33	31
weak residue	42	41
good residue	58	72
(IV)		
ARP/wARP	EVAL14	EVAL15
Resolution	1.91 Å	1.91 Å
Number of residues	23	342
Number of chains	3	9
Correctness of the model (%)	75.2	99.4
sequence coverage(%)	17	97
Total data	38039	38043
completeness data(%)	98.74	98.75

We obtained a map with SHELXE after 500 cycles, using data up to 1.58 Å and obtained a pseudo free CC of 59.03% for EVAL14 and 74.35% for EVAL15. The map of EVAL15 is much better in terms of connectivity (Fig. 3.17) than that of EVAL14. However, up to a resolution of 3.5 Å, both maps are quite similar (not shown here). The intensities of EVAL15 and EVAL14 are also similar up to that resolution, but often differ between 1.8 and 1.58 Å (Fig. 3.18). The large difference in the maps to 1.58 Å can only be explained by the more accurate weak EVAL15 data. The automatic structure building in ARP/wARP (Table 3.8) showed that it was not possible to build a model using the EVAL14 data. However, with EVAL15, this was no problem. The structure (Sevvana, 2006) was refined isotropically using SHELXL. The result is slightly better for EVAL14 (Table 3.6). However, as finding sufficiently good phases is the more critical step, our results prove the quality of the EVAL15 data integration.

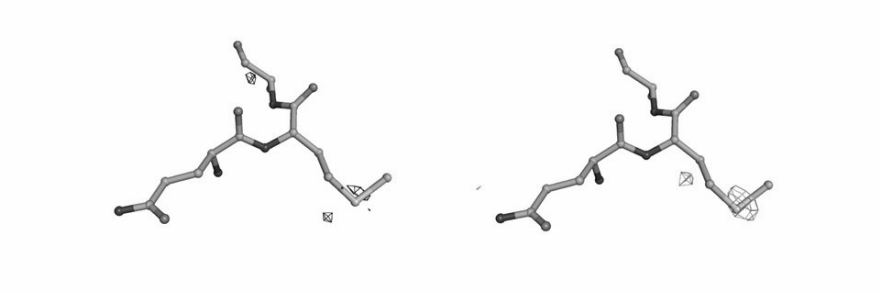


Fig. 3.16 On the left is the anomalous electron density map of EVAL14 and on the right that of EVAL15: a S-atom shows a clearer density in the EVAL15 map than in the EVAL14 map.

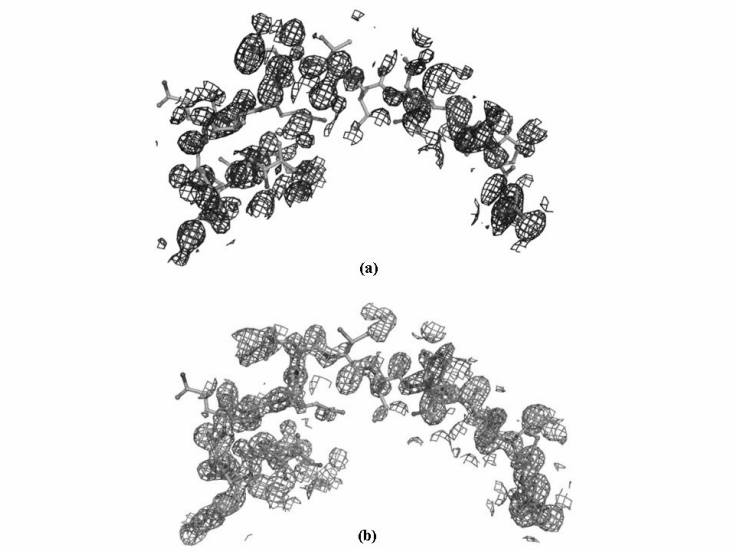


Fig. 3.17 Electron density maps obtained after 500 cycles of SHELXE with the (a) EVAL14 and (b) EVAL15 data. The structure was built in ARP/wARP with the EVAL15 data.

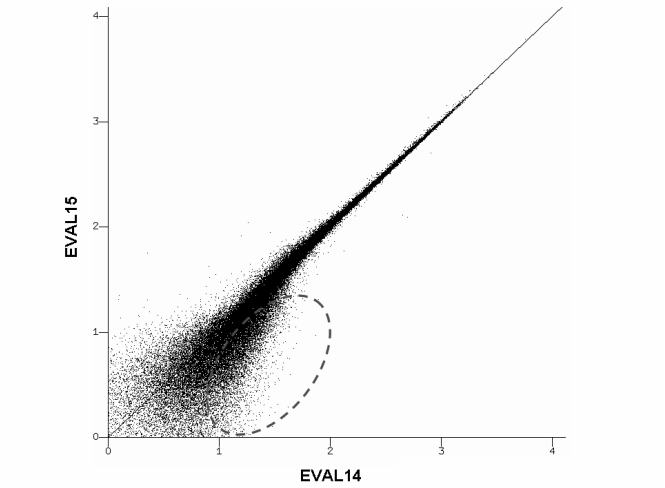


Fig. 3.18 Merged intensities of EVAL15 versus EVAL14. The data differ mostly between 1.8 and 1.58Å (shown in the circle).

3.5 Concluding remarks

The physical profile prediction parameters can be derived straightforwardly with a selection of high I/σ reflections: the point spread γ and *focus dist* are optimised on low duration reflections and the mosaic spread μ on high duration reflections.

We invariably find for EVAL15 data lower R_{merge} for all data sets, especially at high resolution. The I/σ values are larger, notably so at high resolution. Since high resolution shells contain a large number of weak reflections, we conclude that the quality of these weak reflections is superior in EVAL15. Moreover, the more accurately integrated weak intensities contribute significantly to the phasing capacity and improve the electron density map.

Acknowledgement

We thank Madhumati Sevvana, Martin Lutz, Lars von Chrzanowski and Fin Milder for kindly providing the data and structures. We would like to thank the Netherlands Technology Foundation STW for financial support of project UPC 6148.

References

Bourgeois, D., Nurizzo, D., Kahn, R. & Cambillau, C. (1998). *J. Appl. Cryst.* **31**, 22-35.
Chrzanowski, L. S. v., Lutz, M. & Spek, A. L. (2007). *Acta Cryst.* C**63**, m283-m288.
Dauter, Z., Dauter, M., de La Fortelle, E., Bricogne, G. & Sheldrick, G. M. (1999). *J. Mol. Biol.* **289**, 83-92.
Diamond, R. (1969). *Acta Cryst.* A**5**, 43-55.
Duisenberg, A. J. M. (1992). *J. Appl. Cryst.* **25**, 92-96.
Duisenberg, A. J. M., Kroon-Batenburg, L. M. J. & Schreurs, A. M. M. (2003). *J. Appl. Cryst.* **36**, 220-229.
French, S. & Wilson, K. (1978). *Acta Cryst.* A**34**, 517-525.
Hendrickson, W. A. & Teeter, M. M. (1981). *Nature* **290**, 107-113.
Kabsch, W. (1988). *J. Appl. Cryst.* **21**, 916-924.
Leslie, A. G. W. (1999). *Acta Cryst.* D**55**, 1696-1702.
Morris, R. J., Perrakis, A. & Lamzin, V. S. (2002). *Acta Cryst.* D**58**, 968-975.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240-255.
Otwinowski, Z. & Minor, W. (1997). *Macromolecular Crystallography, Pt A* **276**, 307-326.
Perrakis, A., Harkiolaki, M., Wilson, K. S. & Lamzin, V. S. (2001). *Acta Cryst.* D**57**, 1445-1450.
Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458-463.
Pflugrath, J. W. (1999). *Acta Cryst.* D**55**, 1718-1725.
Popov, A. N. & Bourenkov, G. P. (2003). *Acta Cryst.* D**59**, 1145-1153.
Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1986). *Numerical Recipes*. Cambridge: Cambridge University Press.
Rice, L. M., Earnest, T. N. & Brunger, A. T. (2000). *Acta Cryst.* D**56**, 1413-1420.
Rooijakkers, S. H. M., Milder, F. J., Bardoel, B. W., Ruyken, M., van Strijp, J. A. G. & Gros, P. (2007). *J. Immunol.* **179**, 2989-2998.
Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst.* D**58**, 1772-1779.
Schreurs, A. M. M. (1998). *VIEW.* University of Utrecht, The Netherlands.
Schreurs, A. M. M. (1999). *PEAKREF.* University of Utrecht, The Netherlands.
Schreurs, A. M. M. (2007). *ANY.* University of Utrecht, The Netherlands.
Schreurs, A. M. M., Xian, X. & Kroon-Batenburg, L. M. J. (2009). *J. Appl.Cryst.*.
Sevvana, M. (2006). PhD thesis, Georg-August-Universität Göttingen.
Sheldrick, G. (2008). *Acta Cryst.* A**64**, 112-122.
Sheldrick, G. M. (1996). *SADABS.* University of Goettingen, Germany.
Sheldrick, G. M. (1997). *SHELXL.* University of Goettingen, Goettingen, Germany.
Sheldrick, G. M. & Schneider, T. R. (1997). *Methods Enzymol.* **277**, 319-343.
Terwilliger, T. C. (2003a). *Methods Enzymol.* **374**, 22-37.
Terwilliger, T. C. (2003b). *Acta Cryst.* D**59**, 38-44.
Xian, X., Schreurs, A. M. M. & Kroon-Batenburg, L. M. J. (2009). *to be submitted*.

Chapter 4

Treatment of spatially overlapping reflections with EVAL15*

Xinyi Xian, Antoine Schreurs & Loes Kroon-Batenburg

* To be submitted to the Journal of Applied Crystallography

Abstract

EVAL15, a recently introduced diffraction data integration method (Schreurs *et al.*, 2009) uses the principle of general impacts (Duisenberg *et al.*, 2003) and only a small number of physical parameters to predict reflection profiles. Reflection intensities are integrated by least squares fit of the predicted profile to the observed profile using Singular Value Decomposition. With this method, overlapping reflections either due to a long cell axis or due to multiple lattices can be deconvoluted straightforwardly. In this paper we assess the quality of deconvoluted data. The deconvolution contributes to a higher completeness, redundancy and improvement in the electron density map.

4.1 Introduction

In the past decade methods and software for the integration of diffraction data have become straightforward and highly automated. However, for complicated diffraction patterns, such as arising from twin lattices, long cell axes, anisotropic mosaicity, lattice distortion, diffuse scattering and aperiodic structures, challenging tasks still remain. In this paper the capabilities of EVAL15 in deconvoluting overlapping reflections will be assessed.

Several programs are available for the integration of crystallographic diffraction data such as the HKL package (Otwinowski & Minor, 1997), MOSFLM (Leslie, 1999), d*TREK (Pflugrath, 1999). All these programs have in common that only diffraction patterns arising from a single crystal lattice can be described and that the integration of intensities is done by least-squares minimization of the difference of standard and observed profiles. The standard profile is usually learned from neighbouring reflections. When being confronted with spatially overlapping spots, every package handles the situation differently. Such spots are either treated as if they were not overlapping by reducing the integration area (MOSFLM, DENZO) or are removed from the data (Untangle; (Buts *et al.*, 2004)). These solutions have the disadvantage of adding possibly erroneous data or losing essential information. Up to now only XDS (Kabsch, 1988) and PROW (Bourgeois, 1999; Bourgeois *et al.*, 1998) can be used to deconvolute data containing overlapping reflections. XDS handles the integration of overlapping reflections as follows: reflection profiles are transformed to the reciprocal lattice and thus all geometrical distortions are removed. Pixel intensities are assigned to the nearest reflection, though it could belong to two reflections simultaneously. In PROW, the profile fitting method for overlapping and/or weak reflections is combined with summation integration for strong reflections. The fitting area is optimised independently for each spot, such that I/σ is maximized. The deconvolution of overlap is done including neighbouring reflections in the least-squares matrix. In the case of a strong spot overlapping with a weak one, the contour level for the weak spot will be underestimated and the profile fitting area is not optimized.

We recently presented the new data integration program EVAL15 that cannot only deconvolute overlapping reflections by least-squares minimization using Singular Value Decomposition (SVD) (Press *et al.*, 1986), but simultaneously deal with multiple lattices in a straightforward way. The integration of intensities is done by using standard profiles that are predicted, rather than learned. The prediction is made by a ray-tracing simulation based on only a few parameters like mosaicity, crystal shape, beam divergence and wavelength. This has the advantage that any reflection profile can be predicted whether it is overlapping or not. The quality of deconvoluted reflections is compared with that of non-overlapping reflections using four data sets

containing spatially overlapping spots, either due to multiple lattices or due to long cell axes. The comparison of data quality encompasses also EVAL14 that uses summation integration (Duisenberg *et al.*, 2003).

4.2 Methods

Four data sets with spatially overlapping reflections were integrated. Two data sets were of protein molecules crystallized in a large unit cell and two of metal-organic compounds with a small unit cell. Processing of the images was carried out with the EVAL-software suite consisting of DIRAX (Duisenberg, 1992) for indexing, PEAKREF (Schreurs, 1999) for refinement of unit cell dimensions and detector, crystal and goniometer offsets and VIEW (Schreurs, 1998) for viewing of the images and generation of reflection boxes at predicted reflection positions (typically 27x27 pixels and 5 oscillation increments). EVAL15, or alternatively EVAL14, is used for data integration and ANY was used for statistical analysis and graphical display of the data. Subsequently, the data are scaled and absorption correction is applied by SADABS (Sheldrick, 1996), whereby also an error model for the standard deviation is determined. At that stage outliers can be rejected. Data statistics after scaling is obtained with Xprep (Bruker AXS, Madison, USA). The structures were refined against the data using SHELXL (Sheldrick & Schneider, 1997; Sheldrick, 2008) or REFMAC (Murshudov *et al.*, 1997). PHASER (McCoy *et al.*, 2005) was used for molecular replacement, ARP/wARP (Perrakis *et al.*, 2001; Perrakis *et al.*, 1999; Morris *et al.*, 2002) for automatic model building.

4.3 Results

4.3.1 Large unit cell Triplet (I)

X-ray data of Glucose isomerase were collected on a Bruker Cu-rotating anode with Osmic focussing mirrors and a SMART 6000 4K detector up to 1.58 Å (kindly provided by Madhumati Sevvana (2006)). The compound crystallized as a three component twin (Fig. 4.1). In order to get improved separation of the coincidentally overlapping reflections due to the three crystal lattices, fine sliced data of 0.2° per frame were recorded and this was done at a large crystal-detector distance (Table 4.1). A total of 9362 images were collected in low-, medium- and high-resolution passes. The low-resolution images were used for the determination of the unit cell. 10 peaks were searched per frame for 100 consecutive frames in different sectors, some 60-80º apart, for 4 different ω-scans. From this list of peaks a random selection of 1000 were used for indexing in DIRAX. In this program reciprocal lattice vectors are constructed from the peak positions, from which the smallest unit cell that fits a maximum number of reciprocal lattice vectors is found. In this case only 464 reciprocal lattice vectors fitted a single unit cell. With the remaining reciprocal lattice vectors two other unit cells, with approximately the same dimensions, but different orientations, could be found. In PEAKREF the lattice dimension of the three unit cells are constrained to the same values and by minimizing the difference between the observed and calculated impact positions with the simplex method (Press *et al.*, 1986), the cell orientations, as well as the detector and crystal position (horizontal, vertical and rotational direction) are refined.

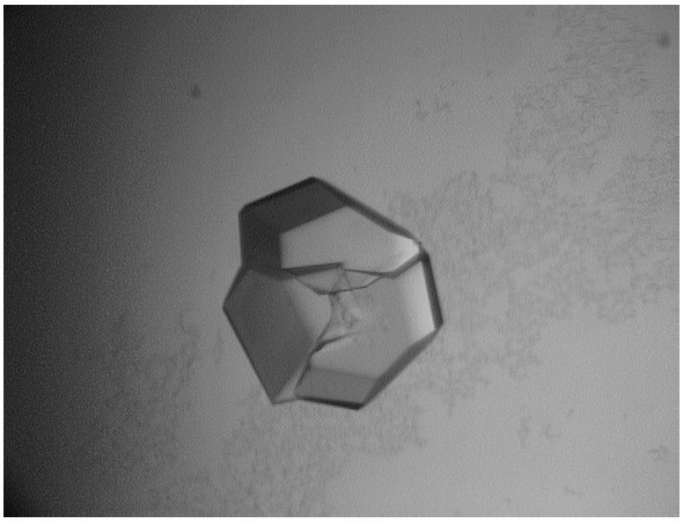


Fig. 4.1 Glucose isomerase is crystallized as a twin with three components with relative rotations of ~120º around arbitrary axes (Picture kindly provided by Madhumati Sevvana).

The three refined unit cells have orientations related by ~ 120º rotations around three arbitrary axes. Therefore, the overlap of the three lattices is coincidental. The refined unit cell dimensions and detector positions result in correct and complete predictions (Fig. 4.2). The prediction for the different lattices is shown as white (component 1), blue (component 2) and green reflection boxes (component 3). For the integration, these 2-dimensional reflection boxes are collected over several images producing 3-dimensional reflection boxes.
In EVAL15 reflection profiles are simulated by producing 10,000 generated impacts from different random selections of focus sample points, crystal sample points, mosaic orientations and wavelength values of the radiation (Schreurs *et al.*, 2009), and these are used in a least-squares minimization with the observed profiles. By generating new impacts from the same sample points combined with the neighbour reciprocal lattice vector, the neighbour profile is obtained.
A least-squares minimization of

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{\rho_i - JP_i - \sum_{m}^{M} J_m P_{im} - ax_i - by_i - c}{\sigma_i} \right]^2 \qquad \textbf{(4.1)}$$

leads to a set of linear algebraic equation that are solved by Singular Value Decomposition. N is the total number of pixels in the reflection box, ρ_i is the observed photon count, P_i is the value of the simulated profile at pixel i, J the scale factor for the main reflection, J_m is the scale factor for the m^{th} neighbour reflection, x_i and y_i are the horizontal and vertical pixel coordinates and a, b, and c are the parameters describing a planar background in the reflection box. Ideally

$$\sigma_i = \sqrt{\rho_i^{calc}} \quad \text{and} \quad \rho_i^{calc} = JP_i + \sum_m^M J_m P_{im} + ax_i + by_i + c\,. \qquad \textbf{(4.2), (4.3)}$$

The intensity of the main reflection is given by

$$I = \sum_{i=1}^{N} JP_i\ . \qquad \textbf{(4.4)}$$

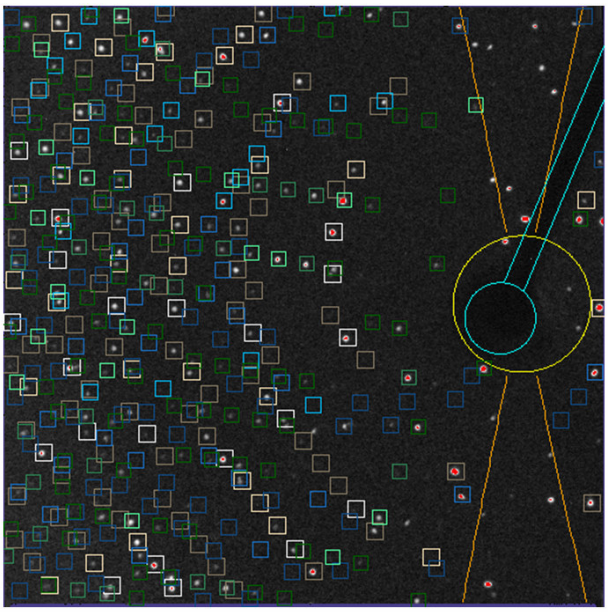


Fig. 4.2 The position of reflections is predicted in VIEW using the three differently oriented unit cells: white reflection boxes indicate component 1, blue component 2 and green component 3. The three colours can have different hues, which indicate differences in rotation angle ω with respect to the current image and represents the duration of the reflection. The beam-stop and the limitation by maximum duration are shown; in these areas the reflections will not be predicted. The 2-D boxes are collected to form 3-D boxes, wherein the reflections will be integrated.

By this procedure the reflection is automatically deconvoluted from the M neighbours. The algorithm also provides standard deviations, the intensities of the neighbours and the variance-covariance matrix, which is instrumental in deciding whether the reflections can be reliably split; if this is not the case, they are summed

(Schreurs *et al.*, 2009). However, in the current data set the latter situation occurs rarely, so that these reflections will not be used in further analysis.
The area of the main reflection is defined by those pixels, which receive at least 0.3% of the impacts (red contour in Fig. 4.3). If the generated impacts of the neighbour reflection also fall in a pixel belonging to the main reflection, then this pixel is an overlap pixel. The overlap fraction is defined as the number of overlapping pixels divided by the number of pixels in the main reflection. When this fraction is smaller than 0.1, the reflection is labelled 'single'. The set of single reflections are referred to as the S-data. The overlap-deconvoluted reflections are referred to as the O-data and the total data set is referred to as the A-data.
During the integration, the intensities of main reflections are stored, while those of the deconvoluted overlapping neighbour reflections are not. By using the three orientation matrices successively as main lattices, the intensity of these neighbour reflections can also be integrated. Nevertheless, the EVAL15-display (Fig. 4.3) does show the intensity of the neighbour below that of the main reflection. For the integration of the intensity the accuracy of the predicted reflection position is vital (Pflugrath, 1999), because even sub-pixel differences can produce a significant difference in the calculated intensity. The predicted reflection position is shifted such that it coincides as well as possible with the observed reflections. This is done by minimizing $fom_{box,}$ a figure-of-merit indicating the quality of the fit:

$$fom_{box} = \left[\frac{\sum_{i=1}^{N} w_i (\rho_i - \rho_i^{calc})^2}{N - N_p} \right]^{1/2}, \qquad \textbf{(4.5)}$$

where the weights are chosen to be $w_i = 1/\sigma_i^2$ (with standard deviation $\sigma_i = (\rho_i^{calc})^{1/2}$). The number of fitted parameters N_p is usually 4, i.e. the scaling parameter J and three background parameters, when no neighbours are in the reflection box.
The quality of the S- and O-data is compared using component 1. The total percentage of overlap-split reflections is only 9% (Table 4.1) and the maximum overlap fraction is 0.69.
In this data the overlap fraction does not increase with resolution (Fig. 4.4). The R_{merge} is low for both the S- and O-data (Table 4.2). The lower R_{merge}, in case the Friedel pairs are not merged, suggests the presence of anomalous scattering. The normal probability plot of $\delta(\text{observed})=(I(S)-I(O))/(\sigma(S)^2+\sigma(O)^2)^{1/2}$ shows no specific deviation between the data, confirming that the quality of the O-data is similar to that of the S-data (Fig. 4.5). The merged I/σ's of the O-data are lower because of lower redundancy, whereas the unmerged I/σ's are similar (Fig. 4.6). Comparison of the three different components (Table 4.2) reveals that component 1 and 3 have a lower R_{merge} than component 2, as well as higher average I/σ's. The phasing step using anomalous scattering of manganese, magnesium and sulphur atoms was straightforward and the structure could be built almost completely with ARP/wARP. A comparison of the phasing will not be discussed here, as the number of reflections in the O-data is not large enough to make a significant contribution to this. For the refinement only the first component is used. The structure (kindly provided by Sevvana) was refined isotropically using SHELXL against A-1 and S-1. The results are similar for both data sets (Table 4.3). R_{free} was calculated separately for 5% of the data.

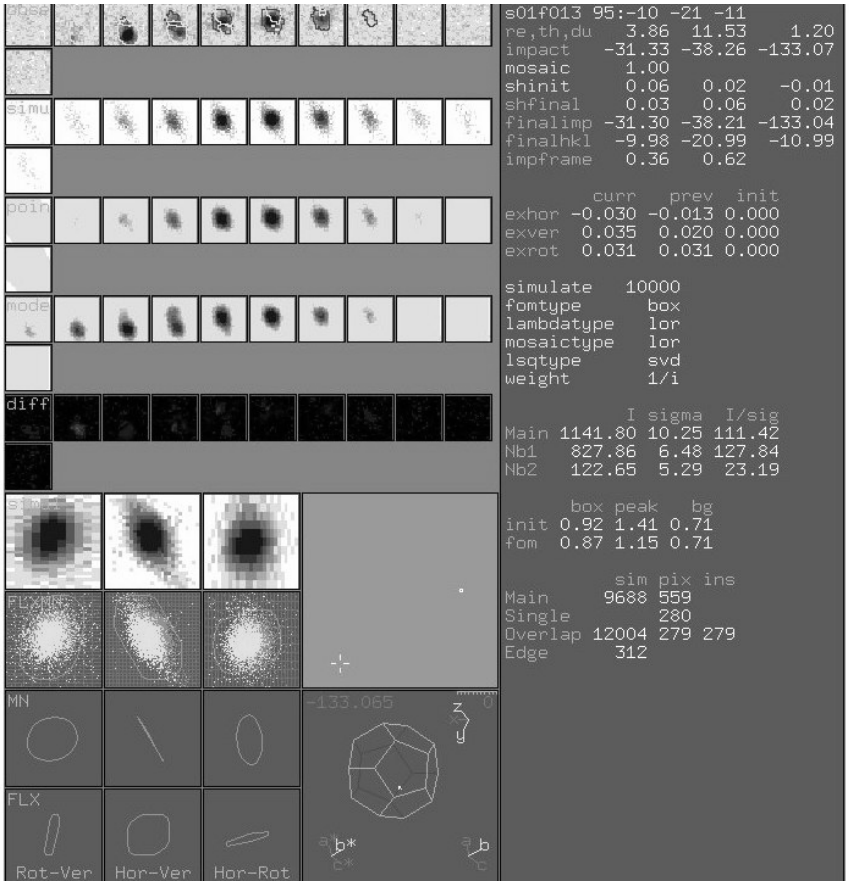


Fig. 4.3 EVAL15-display shows the main reflection (within the dark contour) with its two neighbour reflections (in the light contours). The final position of this reflection on the detector after refinement of the predicted position is displayed [finalimp]. The improvement of fom_{box} after refinement is shown (compare the number at [init] and at [fom]). The resolution in Å [re], relative duration [du], mosaicity value in ° [mosaic], shifts in the horizontal, vertical directions in mm and in the rotational direction in ° are shown [shfinal]. 10,000 impacts for the main reflection are generated, 312 impacts fall outside of the box [Edge]. The rest of the impacts fall into 559 pixels, of which 279 pixels are in the overlap area and 280 pixels are not, so the overlap fraction is 0.49. The intensity [I] and σ [sigma]of the main reflection and the two neighbours is shown. Left upper part of the display shows the observed reflection in successive ω-slices [obs]. Below that the simulated profiles, those including pointspread and the scaled model including background are shown.

Table 4.1 Crystal and measurement parameters

	(I)	(II)	(III)	(IV)
cell dimensions:				
a (Å)	94.47	9.02	67.63	7.40
b (Å)	99.61	9.02	113.28	8.02
c (Å)	104.39	66.48	273.52	8.29
α (°)	90	90	90	89.44
β (°)	90	90	90	77.53
γ (°)	90	90	90	74.66
spacegroup	$I222$	$P4_12_12$	$P2_12_12_1$	$P\bar{1}$
Resolution (Å)	1.58	0.77	2.3	0.77
rotation (°)/frame	0.2	0.3	1	1
crystal-detector distance(mm)	180	100	210	40
maximal overlapfraction	0.69	0.32	0.93	0.69
Overlap %	9 %	41 %	80 %	22 %
λ (Å)	1.54178	0.71073	0.87260	0.71073
temperature (K)	100	150	100	150
profile parameters:				
crystal dimensions (mm)	0.2x0.2x0.2	0.15x0.09x0.21[ii]	0.15x0.15x0.15	0.06x0.15x0.21[ii]
mosaicity (°)	0.6/0.3/0.6[i]	0.3	0.25	0.35
pointspread γ (pixels)	1.0	0.5	0.8	0.5
focus dist/width/length (mm)	150/0.3/0.3	150/0.3/0.3	1000/0.1/0.1	150x0.3x0.3
λ/ σ_λ / w_λ	1.54178/0.001/1.0	0.71073/0.0001/2 0.71359/0.0001/1	0.87260/0.001/1	0.71073/0.0001/2 0.71359/0.0001/1

(I)=Triplet of glucose-isomerase, (II)= $Mn(OTF)_2(Py(ProOH)_2$), where $Py(ProOH)_2$ is 2,6-Bis[[(s)-2-(hydroxymethyl)-1-pyrrolidinyl]methyl]pyridine and T is the triflate F_3CSO_3]
(III)=Hemagglutinin-esterase, (IV)= twin of chloro(triethanolaminato)zinc(II)
[i]the mosaicity is different for the three components (component1/2/3)
[ii]measured accurately by a video microscope

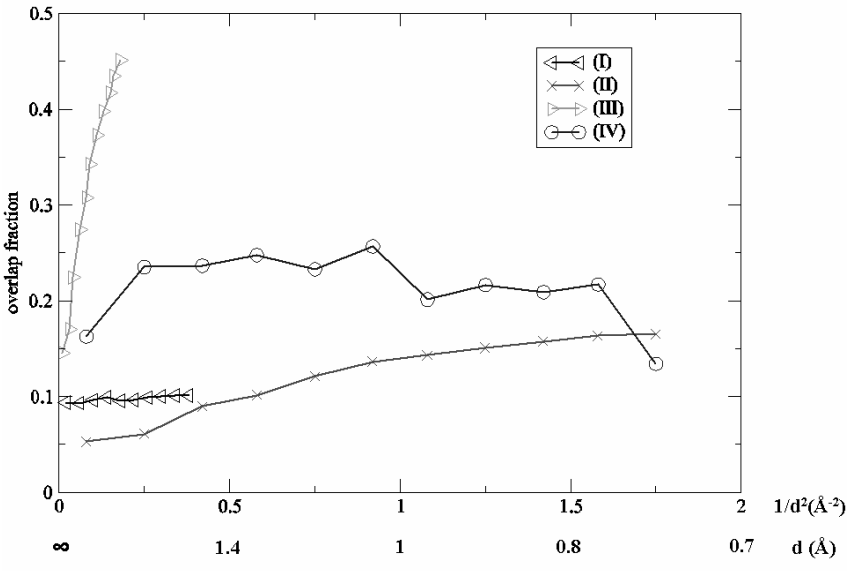


Fig. 4.4 Overlap fraction is plotted against resolution for (I) = large unit cell Triplet, (II) = small unit cell with one long cell axis, (III) = large unit cell with one long cell axis and (IV) = small unit cell twin.

Table 4.2 Data statistics after scaling and merging

	R_{merge} (Friedels merged)*	R_{merge} (Friedels not merged)*	completeness	I/σ^i	total # data	unique data[i]	redundancy
(I)	resolution 1.58 (1.70-1.58)Å						
A-1	0.044 (0.359)	0.039 (0.311)	89.2	18.74 (2.46)	313006	59916 (9446)	5.22
S-1	0.044 (0.357)	0.038 (0.319)	85.6	18.47 (2.32)	283251	57530 (8602)	4.92
O-1	0.037 (0.285)	0.033 (0.273)	25.8	9.91 (1.93)	29755	17300 (2901)	1.72
A-2	0.067 (0.463)	0.062 (0.440)	83.0	13.92 (2.39)	311739	55898 (8667)	5.58
A-3	0.044 (0.336)	0.034 (0.317)	81.0	20.33 (3.01)	311689	54318 (8103)	5.74
(II)	resolution 0.77 (0.90-0.77)Å						
A	0.023 (0.056)	0.015 (0.042)	99.6	63.29 (24.69)	27025	3842 (1366)	7.06
S	0.020 (0.053)	0.014 (0.037)	72.8	64.84 (19.16)	15883	2804 (563)	5.66
O	0.032 (0.054)	0.023 (0.040)	88.2	34.98 (21.75)	11216	3402 (1365)	3.29
EVAL14	0.028 (0.066)	0.021 (0.050)	98.9	50.80 (20.26)	25475	3813 (1349)	6.68
(III)	resolution 2.30 (2.40-2.30)Å						
A	0.081 (0.364)	-	88.6	14.73 (4.80)	464971	83717 (8675)	5.55
S	0.051 (0.165)	-	29.8	16.94 (4.49)	93233	28153 (1135)	3.31
O	0.095 (0.375)	-	80.2	11.97 (4.61)	371738	75754 (8614)	4.91
EVAL14	0.103 (0.616)	-	87.4	11.45 (2.82)	395306	82557 (8801)	4.79
(IV)	resolution 0.77 (0.90-0.77)Å						
A	0.025 (0.041)	-	82.8	35.67 (17.66)	7513	1749 (650)	4.29
S	0.023 (0.042)	-	67.0	35.94 (17.45)	5833	1416 (530)	4.12
O	0.028 (0.038)	-	27.0	27.03 (15.34)	1685	569(166)	2.96
EVAL14	0.025 (0.052)	-	69.7	34.85 (13.95)	5870	1472 (555)	3.99

*number in parentheses refer to the highest resolution shell.

Table 4.3 Refinement results

	R(strong/all)	#(strong/all)	R_{free}(strong/all)*	#(strong/all)	$\Delta\rho_{max/min}$(Å^{-3})	S	wR_2	WGHT(a b)	-
(I)									
A-1	0.181/0.204	42223/51438	0.207/0.238	2165/2676	0.46/-0.31	2.604	0.4957	0.20 0.00	-
S-1	0.179/0.203	40755/49710	0.202/0.229	2113/2567	0.42/-0.30	2.595	0.4930	0.20 0.00	-
(II)									
A	0.023/0.026	5941/6218	-	-	0.36/-0.25	1.082	0.0584	0.03 2.00	-
S	0.022/0.023	4172/4318	-	-	0.26/-0.19	1.085	0.0522	0.02 2.34	-
EVAL14	0.027/0.031	5806/6155	-	-	0.35/-0.31	1.075	0.0649	0.03 2.16	-
(III)	R	data#	R_{free}	#R_{free}	fom	rms(Bond Å)	rms(angle °)	rms(chiral)	B(Å^2)
A	0.225	79448	0.256	4205	0.813	0.0113	1.424	0.091	36.17
S	0.216	27617	0.250	1467	0.836	0.0077	1.286	0.079	26.94
O	0.224	72411	0.256	3838	0.809	0.0108	1.416	0.091	39.965
EVAL14	0.233	78352	0.267	4152	0.803	0.012	1.469	0.094	36.709
(IV)									
A	0.021/0.025	1609/1749	-	-	0.38/-0.34	1.149	0.0557	0.02 0.33	-
S	0.018/0.022	1285/1418	-	-	0.27/-0.23	1.062	0.0307	0.01 0.32	-
EVAL14	0.020/0.028	1294/1472	-	-	0.26/-0.31	1.062	0.0413	0.01 0.30	-

*R_{free}(5% data)

$R=\Sigma||F_{obs}|-|F_{calc}||/\Sigma|F_{obs}|$

$\Delta\rho_{max/min}$=maximum and minimum difference density

$S=\{\Sigma[w(F_o^2-F_c^2)^2/(n-p)]\}^{1/2}$, where n = number of reflections, p = number of parameters refined

$wR_2=\{\Sigma[w(F_o^2-F_c^2)]^2/\{\Sigma[w(F_o^2)^2]\}^{1/2}$

WGHT = weighting scheme: $w=1/[\sigma^2(F_o^2)+(aP)^2+bP]$, where $P=(F_o^2+2F_c^2)/3$

fom= Figure of merit

rms=root mean square deviation

B=temperature factor

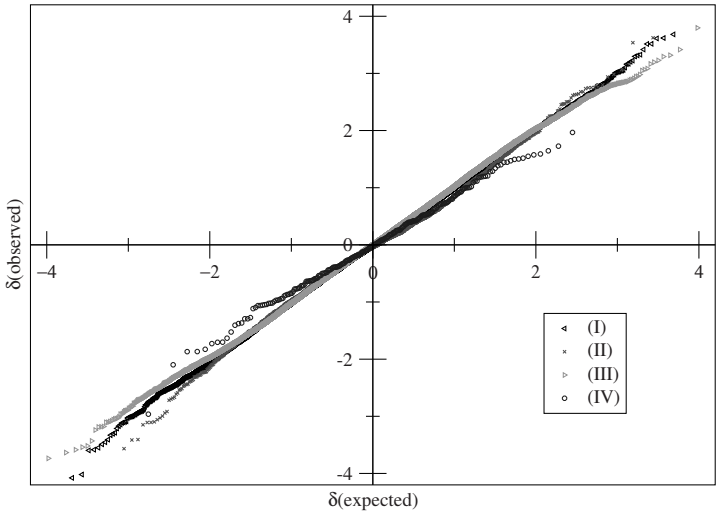


Fig. 4.5 Normal probability plot comparing the merged S- and O-data. $\delta(\text{observed})=(I(S)-I(O))/(\sigma(S)^2+\sigma(O)^2)^{1/2}$ is shown for (I) with 14913 data points, (II) with 2363 data points, (III) with 20184 data points and (IV) with 235 data points.

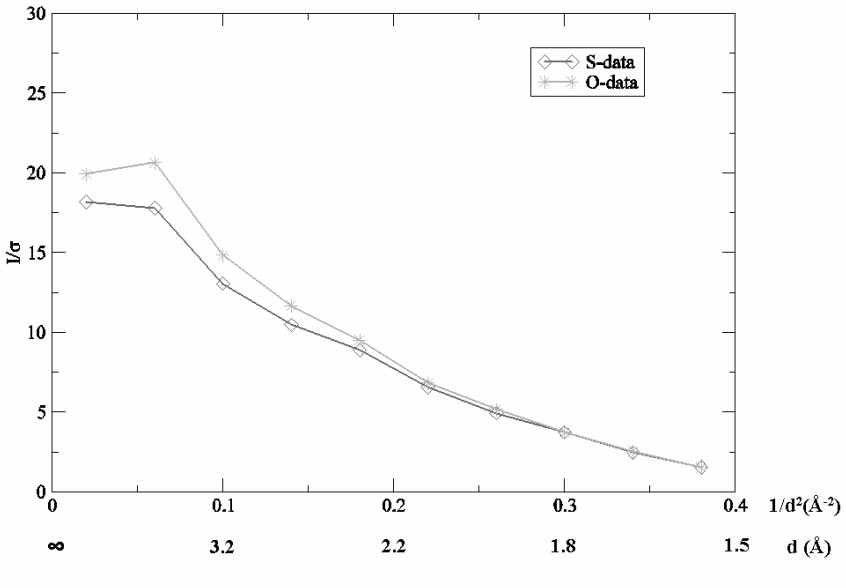


Fig. 4.6 Unmerged I/σ of (I) is plotted against the resolution for the S- and O-data.

It is evident that finding and refining the lattices of the three components of the glucose isomerase triplet is straightforward using DIRAX and PEAKREF, and that

EVAL15 can deconvolute overlapping reflections without any problem and with good accuracy. Since the current data only contain 9% overlapping reflections, the contribution of these reflections to the structure solution cannot be investigated and, as it appeared, they do not have much effect in the refinement as well. Therefore, data sets containing more overlap are investigated in the following sections.

4.3.2 Small unit cell with one long cell axis (II)

The data of $Mn(OTF)_2(Py(ProOH)_2$ were collected in-house on a Nonius KappaCCD, with rotating anode MoK_α radiation, using fine slicing of 0.3º per image to a resolution of 0.77 Å (images kindly provided by Martin Lutz (manuscript in preparation). Due to the systematic overlap in the c^*-direction, the overlap percentage is higher (41%) than for the Triplet twin data and the overlap fraction increases with resolution (Fig. 4.4). The maximum overlap fraction of 0.32 is rather low.
Two types of overlap occur: either in the ω-direction (Fig. 4.7a) or within one image (Fig. 4.7b). In general, overlap can be reduced by fine-slicing and by choosing a large crystal-detector distance. Both were applied here. The overlap in the ω-direction is largely resolved due to fine slicing (like in Fig. 4.7a), but within one image it can still be significant (like in Fig. 4.7b) and increases with *l*-index (Fig. 4.7c).

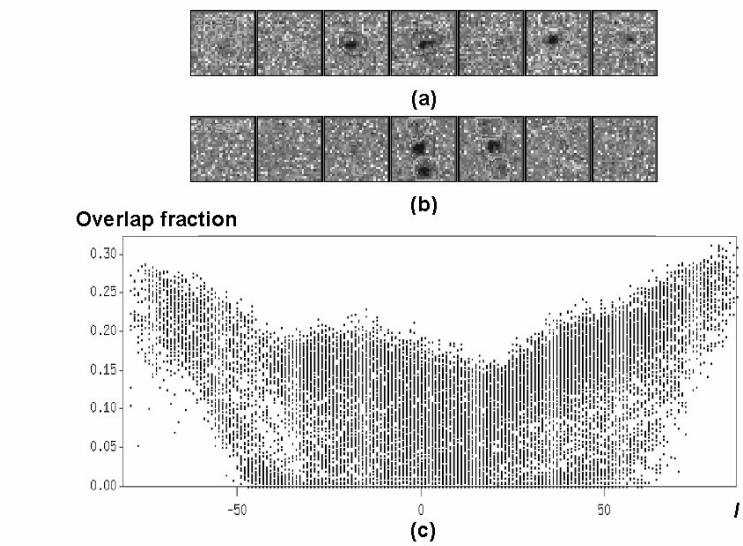


Fig. 4.7 (a) When reflections overlap in the ω-direction, they are separated due to fine slicing (reflection 3 9 1). (b) Even with low *l*-indices, the overlap fraction can be significant, which is due to overlap within one image (reflection 9 -3 -1). (c) The overlap fraction, plotted against the *l*-index for all reflections, increases with *l*/*l*.

R_{merge} for the O-data is larger than that for the S-data (Table 4.2), which might be due to the smaller amount of strong reflections (after scaling, only 414 have $I/\sigma \geq 50$ versus 3896). The normal probability plot of the difference intensity between the S-

and O-data reveals no systematic outliers and a slope of approximately one (Fig. 4.5) and thus they prove to be similar.
The data contain contributions of the anomalous scattering of manganese, where $\delta f''_A$ = 2.74 electrons at λ = 1.54178 Å. The anomalous differences of the Bijvoet-pairs are investigated. The normal probability plot of $(\Delta F_{calc}^2 - \Delta F_{obs}^2) / (\sigma(F^+_{obs}{}^2)+\sigma(F^-_{obs}{}^2))^{1/2}$ shows a normal behaviour for both data sets, with a correlation coefficient of 0.9996 and 0.9991, a slope of 0.9832 and 0.9291 and an intercept of -0.0384 and 0.0284 for the S and O-data respectively. Therefore, even small anomalous differences can be measured accurately by overlap deconvolution.
Because of the fact that the overlap split data add 41% to the number of reflections, and 25 % to the number of unique reflections, a significant effect is expected in the refinement. However, the quality indicators are similar for both data sets. Also the refined structures (coordinates, anisotropic U_{ij}'s) are not significantly different. This implies that the structure refinement problem is already over-determined with the S-data, but that on the other hand the O-data are of good quality and that the overlap deconvolution by EVAL15 is successful.
The integration with EVAL15 was straightforward and standard, whereas that with EVAL14 was not. For the latter an artificially small contour was used in order to split the overlapping reflections and the size of the shifts was limited to prevent weak main reflections shifting to strong neighbour reflections. The R values of merging and refinement of EVAL15 are slightly lower and the average I/σ is higher than that of EVAL14 (Tables 4.2 and 4.3). One of the advantages of EVAL15, using profile prediction, versus EVAL14, using summation integration, is the more accurate integration of weak intensities (Xian *et al.*, 2009). On account of the fact that very few (roughly 300 out of 6000) reflections are weak, this advantage cannot be exploited here.

4.3.3 Large unit cell with one long cell axis (III)

Hemagglutinin-esterase (Zeng *et al.*, 2008) was crystallized with 4 molecules per asymmetric unit. X-ray data were collected at the ESRF on Beamline ID14-3 up to 2.3 Å on a marmosaic225 detector (images kindly provided by Qinghong Zeng). Due to the very large c-axis (Table 4.1) and the unfortunate circumstance that it was parallel to the beam in parts of the rotation range, the overlap was large especially in that range (Fig. 4.8). 80% of the reflections are overlapping. The overlap fraction increases with higher resolution (Fig. 4.4) and the maximum overlap fraction of 0.93 is very high (Table 4.1). The R_{merge} for the O-data (Table 4.2) is significantly larger than that for the S-data. This is partly due to two reasons: firstly the amount of data is larger and secondly the overlap-split reflections are generally weaker, because they occur at higher resolution. The normal probability plot of the difference in intensity between the S- and O-data reveals no systematic deviation between the data (Fig. 4.5) and the I/σ values at the highest resolution shell are similar. In the structure determination from protein data sets, the phasing step is critical and here the quality of the data can be assessed the best (Xian *et al.*, 2009). Using the backbone of one molecule (the coordinates of which are kindly provided by Qinghong Zeng) as a search model, PHASER could successfully find 4 molecules in the asymmetric unit for the S- and A- data.
However, the fourth molecule was more difficult to find due to, as it turned out later, disorder. PHASER had no problems in finding the correct solution for the S-data, because these are rather complete at lower resolution (ca 70% at 4Å) and molecular

replacement can easily be done at this resolution. Thus, the phasing step by molecular replacement is not a good test for the data quality. Nevertheless the 2Fo-Fc maps do show significant differences between the S- and A-data. The map of the S-data has stretched features in the direction of the long cell axis (Fig. 4.9a). As already mentioned, the overlap is systematic in this direction, and in the S-data all reflections with *l*-indices larger than 40 (*l*-max = 120) are missing, which causes the artefacts in the electron density in this direction. After the phasing step, ARP/wARP was used to build a more complete structure. The number of residues built with the S-data is almost 5 times less than that with the A-data (Table 4.4) and the sequence coverage is only 27% compared to 83%. Figure 4.9 reveals why it is more easy to build the side-chains with the A-data than with only the S-data: the electron densities at e.g. the positions of the side chain or sugar moiety are much more well-defined. Thus, deconvolution of overlapping reflections adds valuable information to the data set. The structure used for the refinement was solved by molecular replacement using PHASER, but with a different data set that had less overlapping reflections and a higher resolution (1.8 Å) (Zeng *et al.*, 2008). In order to do a refinement that is adapted to the current resolution, the water molecules were deleted from the coordinate file. The R/R_{free} of the S-data are not significantly different from the A-data given the difference in number of reflections, which confirms that the data quality of the O-data is equally good. However, in the Fo-Fc difference map of the S-data some stretched features in the direction of the long cell axis are seen as before; this map has a lower resolution in the c-direction. Although the refinement with only 1/3 of data can be done without any problems, building a correct model would have been a cumbersome task.

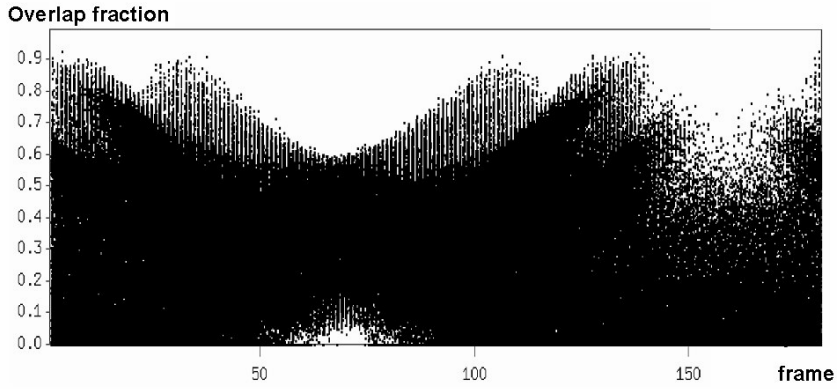
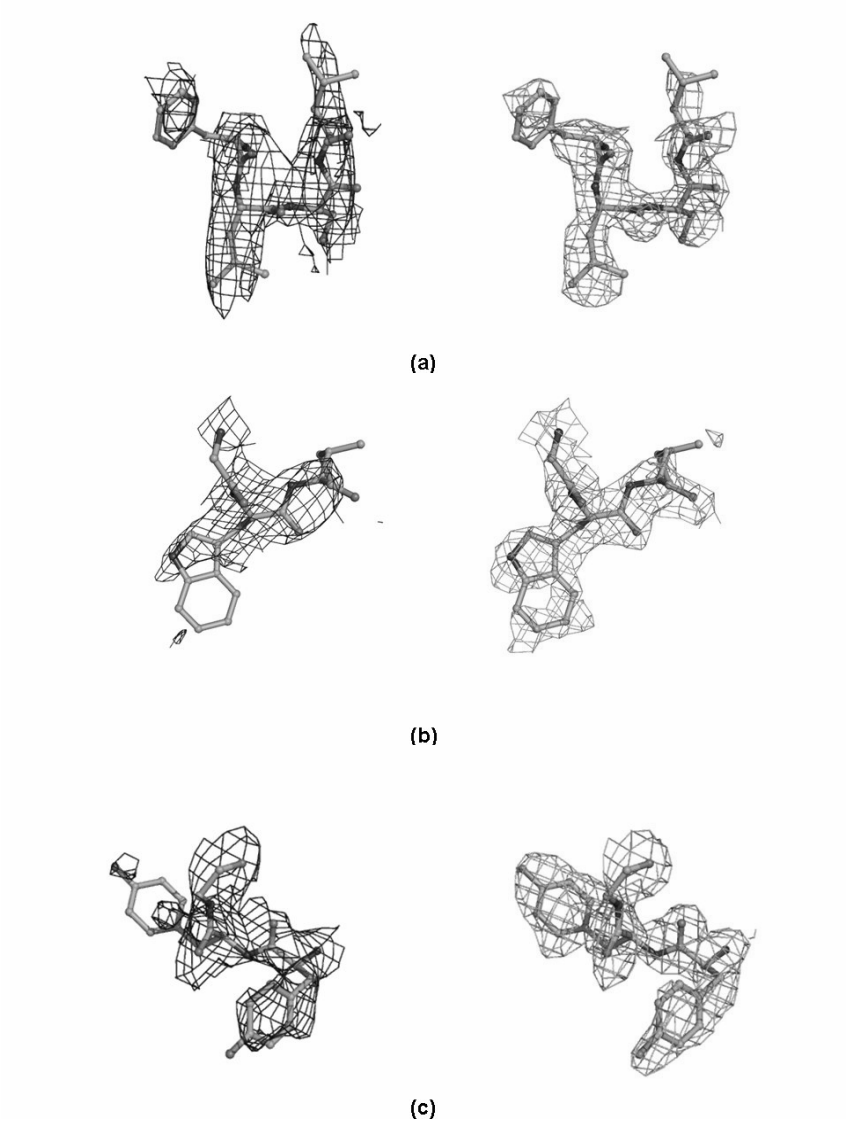
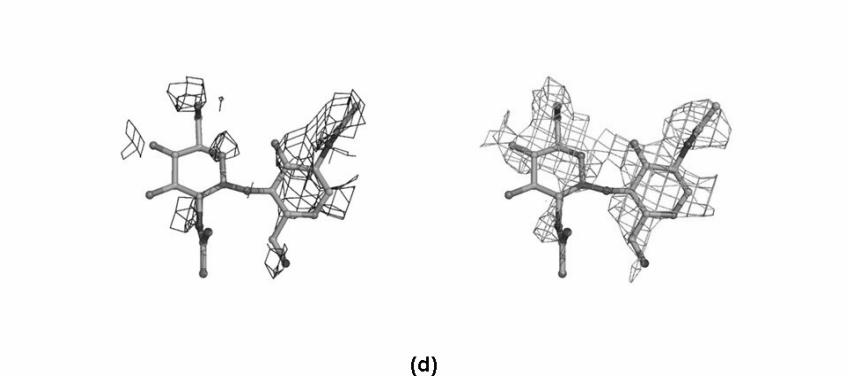


Fig. 4.8 The overlap fraction is plotted against the number of the frame for all reflections. Overlap is systematic and larger when the c-axis is parallel to the beam. In these ω-ranges it is sometimes even not possible to deconvolute the overlapping reflections (near frame 160). These overlap-sum reflections are omitted from the data and leave a gap in the plot.

The refinement of A-data shows that one molecule had much higher B-factors, than the others (~65 compared to ~30 $Å^2$ (including TLS)). The displacement of this molecule is especially large in the c-direction, due to limited crystal contacts. As the S-data have limited resolution at the c-direction, the fall-off of the scattering with resolution is not well determined, and the structure refined to a lower B-factor (~45 compared to ~25 $Å^2$). This means that some aspects of the structure are not well

described by the S-data, even though this is not seen in the R-values of the refinement.





(d)

Fig. 4.9 2Fo-Fc maps from the PHASER solution of the S-data (left) and A-data (right) are shown for four examples:
(a) the density of a sequence of amino acids (phenylalanine, leucine and serine) shows stretched features (running in the direction of the long cell-axis) with the S-data, but not with the A-data.
The density at (b) a tryptophan-, (c) a tyrosine- residue and (d) a sugar moiety show why side-chains are difficult to build with the S-data, whereas with the A-data that is no problem.

EVAL14 data were obtained by choosing an artificially small contour to avoid overlap. The advantage of more accurate integration of weak data by EVAL15 is seen in the significantly lower R_{merge} (Table 4.2). I/σ is smaller in EVAL14, especially for weak reflections, but also for strong reflections, which is normally not expected for summation integration (Xian *et al.*, 2009). However, the latter is caused by the artificially small contour. The better quality of the EVAL15 data is also reflected in an improved refinement (Table 4.3). The numbers of residues found, as well as the sequence coverage in ARP/wARP, are slightly higher (Table 4.4). We conclude that the overlap deconvolution produces good data.

Table 4.4 Automatic buiding-statistics of (III)

ARP/wARP	EVAL14	A	S
Number of residues	877	951	236
Number of chains	32	29	17
Correctness of the model (%)	94.1	95.4	80.7
sequence coverage(%)	80	83	27
Total data	78242	79335	27141
completeness data(%)	87.3	88.5	31.6

4.3.4 Small unit cell twin (IV)

A crystal of chloro(triethanolaminato)zinc(II), was grown as a non-merohedral twin with a 2-fold rotation around [010]. The data set was collected in-house on a Nonius KappaCCD, a rotating anode with MoK_α radiation, using coarse slicing of 1º per image to a resolution of 0.77 Å (Lutz & Bakker, 2003). As the overlap was coincidental, the overlap percentage was low (22%). The maximum overlap fraction is 0.69 (Table 4.1) and the overlap fraction did not increase with resolution (Fig. 4.4). The R_{merge} is slightly higher and the merged I/σ is lower for the O-data than for the S-

data (Table 4.2), because of lower redundancy. Therefore, the data quality of the O-data appears to be less satisfactory than that of the S-data. The normal probability plot (Fig. 4.5) shows that these data are not entirely similar and surprisingly the slope is even smaller than one, implying that the difference is smaller than expected (Abrahams *et al.*, 1971). However, as only 200 data points are included, the statistics is not reliable.

The structural refinement results with SHELXL (Table 4.3) are slightly better for the S-data than for the A-data. A couple of reflections are strong outliers in a normal probability plot of $(F_{obs}^2 - F_{calc}^2)/\sigma(F_{obs}^2)$ for the A-data (not shown here). These reflections turn out to be split overlapping reflections. The reason for these errors is actually a concurrence of circumstances: the overlap is coincidental, but with a peculiar relation: [*h k l*] and [-(*h*-0.5**k*) *k* –*l*] often overlap and their θ-difference of 0.1º is extremely small: thus their horizontal and vertical coordinates also lie very close (within one pixel). Apparently the deconvolution for this type of overlap combined with coarse slicing is less accurate. In this case shifts of reflection profiles in the ω-direction of ~0.2º may lead to an erroneous deconvolution, while equivalent reflections are integrated as overlap-sum. This means that situations occur where, of a set of equivalent reflections only one is left and SADABS has no opportunity to reject this reflection. In the refinement, these kind of reflections are on the top list of the most disagreeable reflections. An improvement to the above problem would have been fine slicing, so that overlapping reflections can be resolved in the rotational direction.

The data of EVAL14 (again integrated with an artificially small contour) has a lower R_{merge} than the EVAL15 A-data, but not better than the S-data. In the refinement EVAL14 has lower R-value, S, wR_2 and a lower Δρ, but again not compared to the EVAL15 S-data (Table 4.3). An important advantage of EVAL15 is that it delivers much more data than EVAL14, despite the integration with artificially small contours.

4.4 Discussion

EVAL15 can deal with multiple lattices in a straightforward way and successfully deconvolute overlapping reflections. The quality of the deconvoluted reflections is similar to that of single reflections. In general, the number of overlapping reflections and the amount of overlap can be reduced by both fine slicing and a larger detector distance. However, the use of the fine slicing is limited by the mosaicity. The choice of crystal-to-detector distance is dependent on the required resolution of the data. Reflections, that lie too close, like in the case of the small molecule twin (IV), cannot be separated by a larger detector distance due to the beam divergence. However, fine slicing would help to improve the situation.

Bourgeois also examined the quality of deconvoluted data, although only of overlap due to a long cell axis (Bourgeois *et al.*, 1998). Similar to Bourgeois, we find that the higher completeness achieved by deconvolution of overlap only has a small penalty on R_{merge}. Therefore, the extra reflections have satisfactory quality. In case of the hemagglutinin-esterase data the extra information gained by deconvolution contributes to an improved electron density map, such that automatic model building is much more successful.

Acknowledgement

We thank Madhumati Sevvana, Martin Lutz and Qinghong Zeng for kindly providing the data and structures. We thank the Netherlands Technology Foundation STW for financial support of project UPC 6148.

References

Abrahams, S. C., Bernstein, J. L. & Keve, E. T. (1971). *J. Appl. Cryst.* **4**, 284-290.
Bourgeois, D. (1999). *Acta Cryst.* D**55**, 1733-1741.
Bourgeois, D., Nurizzo, D., Kahn, R. & Cambillau, C. (1998). *J. Appl. Cryst.* **31**, 22-35.
Buts, L., Dao-Thi, M.-H., Wyns, L. & Loris, R. (2004). *Acta Cryst.* D**60**, 983-984.
Duisenberg, A. J. M. (1992). *J. Appl. Cryst.* **25**, 92-96.
Duisenberg, A. J. M., Kroon-Batenburg, L. M. J. & Schreurs, A. M. M. (2003). *J. Appl. Cryst.* **36**, 220-229.
Kabsch, W. (1988). *J. Appl. Cryst.* **21**, 916-924.
Leslie, A. G. W. (1999). *Acta Cryst.* D5**5**, 1696-1702.
Lutz, M. & Bakker, R. (2003). *Acta Cryst.* E**59**, m74-m76.
McCoy, A. J., Grosse-Kunstleve, R. W., Storoni, L. C. & Read, R. J. (2005). *Acta Cryst.* D**61**, 458-464.
Morris, R. J., Perrakis, A. & Lamzin, V. S. (2002). *Acta Cryst.* D**58**, 968-975.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240-255.
Otwinowski, Z. & Minor, W. (1997). *Macromolecular Crystallography, Pt A* **276**, 307-326.
Perrakis, A., Harkiolaki, M., Wilson, K. S. & Lamzin, V. S. (2001). *Acta Cryst.* D**57**, 1445-1450.
Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458-463.
Pflugrath, J. W. (1999). *Acta Cryst.* D**55**, 1718-1725.
Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1986). *Numerical Recipes*. Cambridge: Cambridge University Press.
Schreurs, A. M. M. (1998). *VIEW.* University of Utrecht, The Netherlands.
Schreurs, A. M. M. (1999). *PEAKREF.* University of Utrecht, The Netherlands.
Schreurs, A. M. M., Xian, X. & Kroon-Batenburg, L. M. J. (2009). *J. Appl.Cryst.*
Sevvana, M. (2006). PhD thesis, Georg-August-Universität Göttingen.
Sheldrick, G. (2008). *Acta Cryst.* A**64**, 112-122.
Sheldrick, G. M. (1996). *SADABS.* University of Goettingen, Germany.
Sheldrick, G. M. & Schneider, T. R. (1997). *Methods Enzymol.* **277**, 319-343.
Xian, X., Schreurs, A. M. M. & Kroon-Batenburg, L. M. J. (2009). *to be submitted.*
Zeng, Q., Langereis, M. A., van Vliet, A. L., Huizinga, E. G. & de Groot, R. J. (2008). *Proc. Natl. Acad. Sci. U S A* **105**, 9065-9069.

Chapter 5

What lies beneath the profile surface

5.1 Introduction

The reflection profile is a convolution of broadening effects of the mosaicity, crystal shape/size, beam divergence, detector point spread and wavelength dispersion. EVAL15 predicts profiles by simulation of 10,000 randomly generated X-ray impacts (Schreurs *et al.*, 2009). With a few physical crystal and instrument parameters, a three-dimensional standard profile is predicted for each reflection that is used to integrate the data by least squares minimization.
The purpose of this chapter is to study profiles in detail. These details can reveal peculiar features of the crystal structure or unexpected instrumental characteristics.

5.2 Lattice distortion in the data of NspA

Data of an Au-derivative (I) of Neisserial Surface Protein A (NspA) (Vandeputte-Rutten *et al.*, 2003) were collected at 100 K on the EMBL beamline X11 at DESY with a rotation of 0.5º per image and a maximum resolution of 4 Å (images kindly provided by Lucy Rutten). The wavelength used was 0.811 Å, which is near the absorption edge of the Au-atom and the crystal has the space group R32 with cell dimensions a = 97.83 Å, b = 97.83 Å, c = 171.87 Å and $\gamma = 120º$.
We examined the data in detail, by comparing predicted profiles with those observed and with learned profiles of DENZO (Otwinowski & Minor, 1997). Reflection profiles vary across the detector, due to geometrical distortions. Neighbour reflections usually have similar profiles, and can be used to generate a learned profile. By default DENZO divides the detector in 9 sectors (Fig. 5.1) and an averaged profile is produced for each of these.
At a rotation of ω=0º such an averaged profile was compared with an observed profile (Fig. 5.2). It originates from the sector 1-3. The observed reflection -8 -17 -25 lies in that sector. Both show an asymmetric tail that points downwards to the direction of the equatorial line. However, the predicted profile of EVAL15 does not have such a tail.
We investigated the averaged profiles at $\omega = 0º$ for the other sectors of the detector. It turned out that all learned profiles above the equatorial line have an asymmetric tail that points downwards, whereas the profiles below that line have an asymmetric tail that points upwards (Fig. 5.3). The profiles of the sectors around that equatorial line are symmetric (data not shown). At further rotation e.g. $\omega = 45º$ and $\omega = 90º$ no tails are seen in the observed profiles, however at $\omega = 180º$ they reoccur (Fig. 5.4).
The predicted profile in EVAL15 was derived according to the procedure described in chapter 3. High I/σ reflections were used to optimize the parameter describing the point spread (γ = 0.879 pixels) and high duration reflections to determine the mosaicity (μ = 0.353º). The crystal dimensions are 0.1x 0.1 x 0.05 mm. By default the beam divergence of the synchrotron source is described with *focus dist* = 1000 mm and *focus width* = 0.1 mm and the wavelength distribution has σ_λ = 0.0001 Å. This initial model lacks the asymmetrical tail, as mentioned above.
Figure 5.5a shows the observed profile in 2-dimensions, projected as a front view. The tail points downwards and the predicted profile shows little resemblance with the observed profile (Fig. 5.5b). The reflection profile is a convolution of broadening effects of the above-mentioned parameters. In search of the origin of this tail, we have increased the individual broadening effects to see if any of them are responsible.

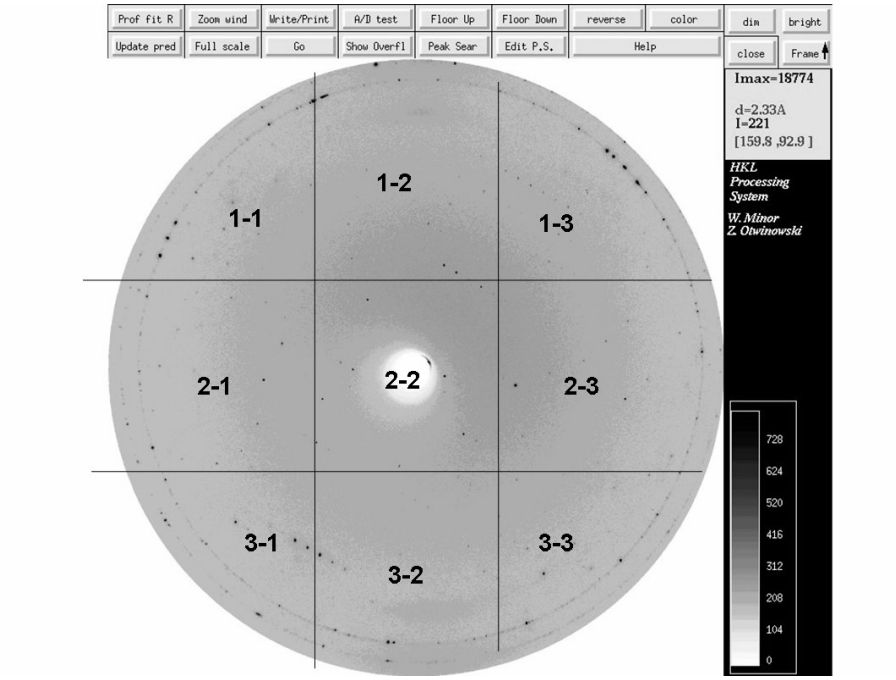


Fig. 5.1 DENZO divides the detector in nine parts.

A larger mosaicity stretches the profile along the powder-arc and a broader wavelength distribution stretches it along the radial direction (Fig. 5.5 c,d). A larger beam divergence or a larger point spread γ have a similar isotropic broadening effect (Fig. 5.5 e,f). If the crystal would consist of two fragments, the profile could be asymmetric. However, as the profiles vary across the detector in a single image, this cannot be the cause of the tail. Thus, none of the known parameters can be used to predict an asymmetric profile.

We found that the c-axis is perpendicular to the beam and parallel to the detector both at ω = -30º and ω = 150º, whereas it is oriented parallel to the beam at ω = 60º. Therefore, the asymmetrical tail is only visible when the c-axis is more or less perpendicular to the beam (e.g. at ω = 0º and ω = 180º). The tail has a slightly different θ-position than the centroid of the profile. As can be inferred from Bragg's law ($2d\sin\theta = n\lambda$), a change in θ can be caused by a different λ-value or a different d-spacing. We have already shown that the tail cannot be predicted with different values of the wavelength, thus the change in θ can only be caused by a different d-spacing in a specific lattice direction. As the tail points in the direction of a decreasing $|l|$-index when the c-axis is perpendicular to the beam, this lattice direction is **[0, 0, 1]**

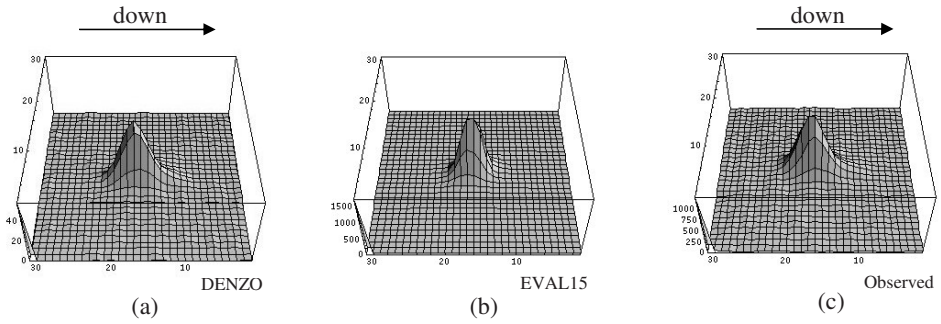


Fig. 5.2 (a) The learned DENZO profile of sector 1-3 at ω = 0º (b) the predicted profile of EVAL15 and (c) the observed profile of reflection -8 17 -25 is displayed in three-dimensions. The asymmetric tail points downwards to the equatorial line of the detector. The spike in the plot is pixel (1,1) at the upper left corner of a reflection box and it is given a high value for reasons of orientation.

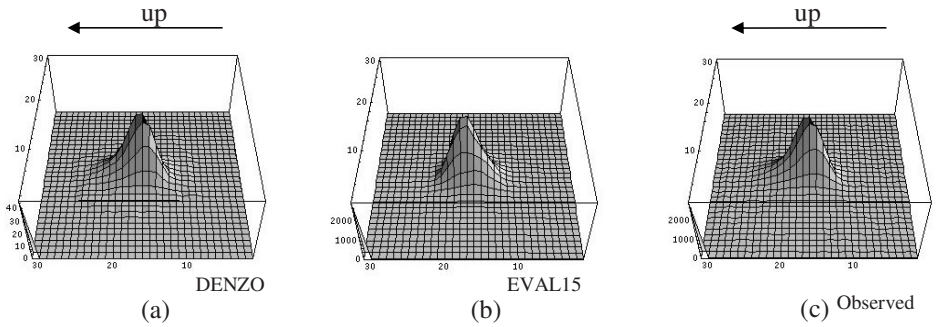


Fig. 5.3 (a) The learned DENZO profile of sector 3-1 at ω = 0º (b) the predicted profile of EVAL15 and (c) the observed profile of reflection 0 -11 26 is displayed in three-dimensions. The asymmetric tail points upwards to the equatorial line of the detector.

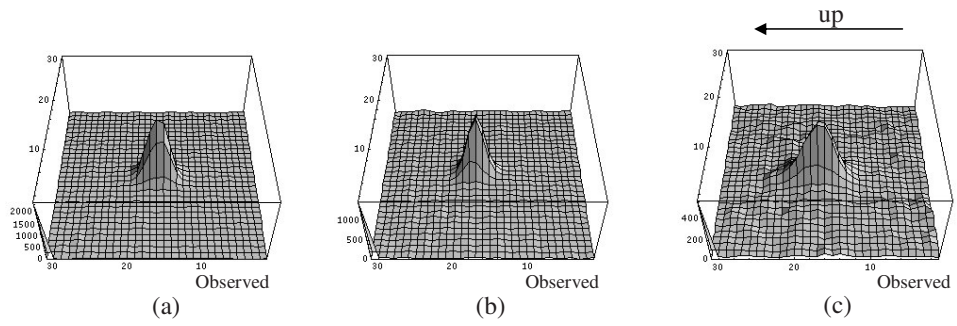


Fig. 5.4 (a) observed profile of reflection -1 11 -6 from sector 1-3 at ω = 45º, (b) observed profile of reflection 5 6 8 from sector 3-2 at ω = 90º and (c) observed profile of reflection 1 12 -29 from sector 3-1 at ω = 180º.

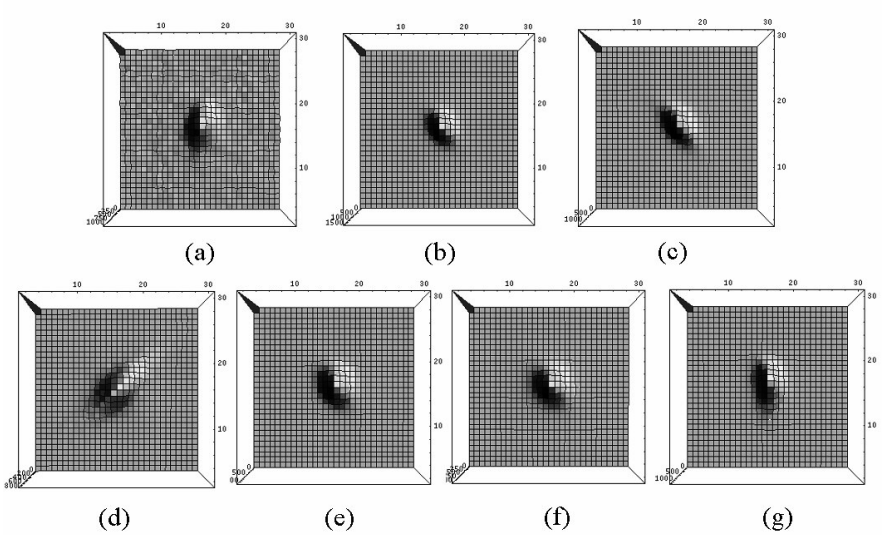


Fig. 5.5 (a) observed profile of reflection -8 -17 -25 at sector 1-3, (b) predicted profile with the initial model, (c) predicted profile with mosaic μ = 0.6° ,(d) σ_λ = 0.003, (e), focus dist = 500 mm, (f) point spread γ = 2.5 pixels and (g) *latt* = 0.035.

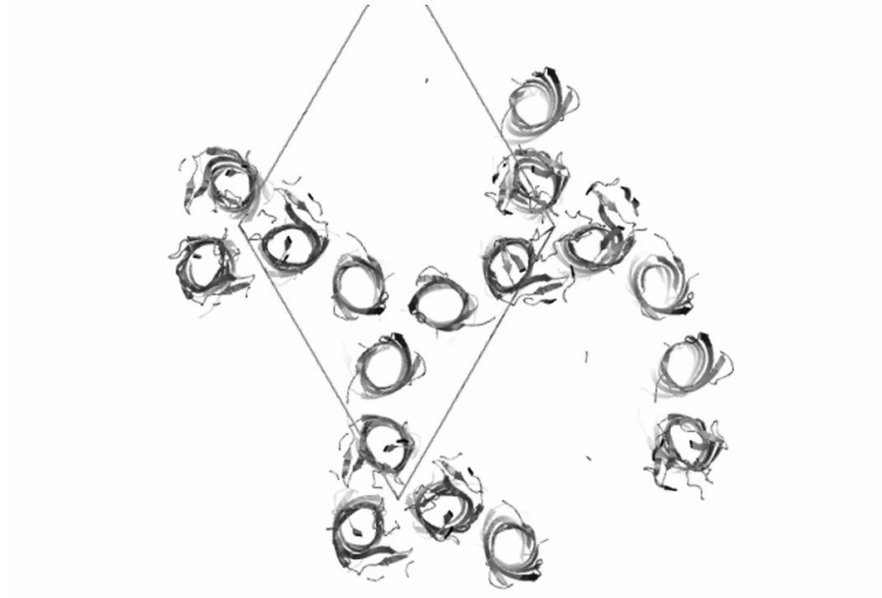


Fig. 5.6 Packing of NspA proteins molecules in the a,b- plane.

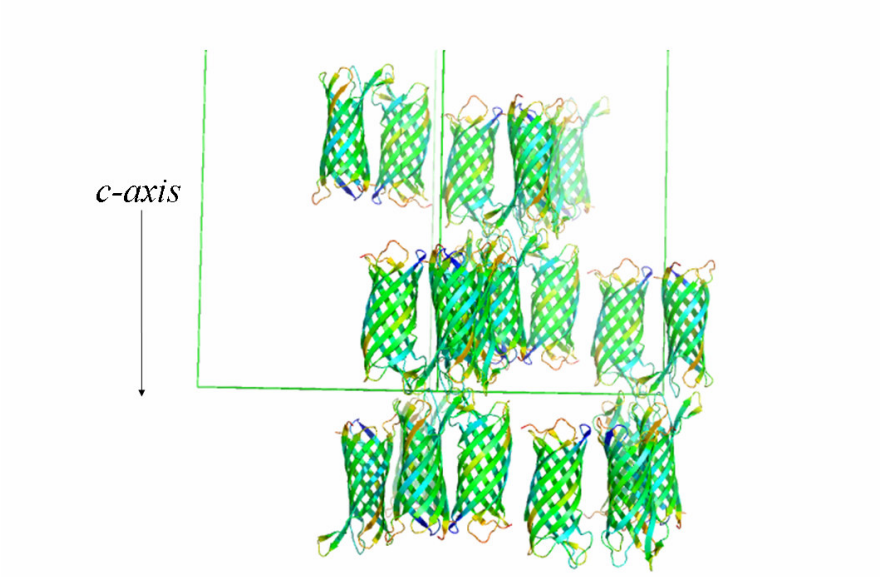


Fig. 5.7 Packing of NspA protein molecules in the direction of the c-axis.

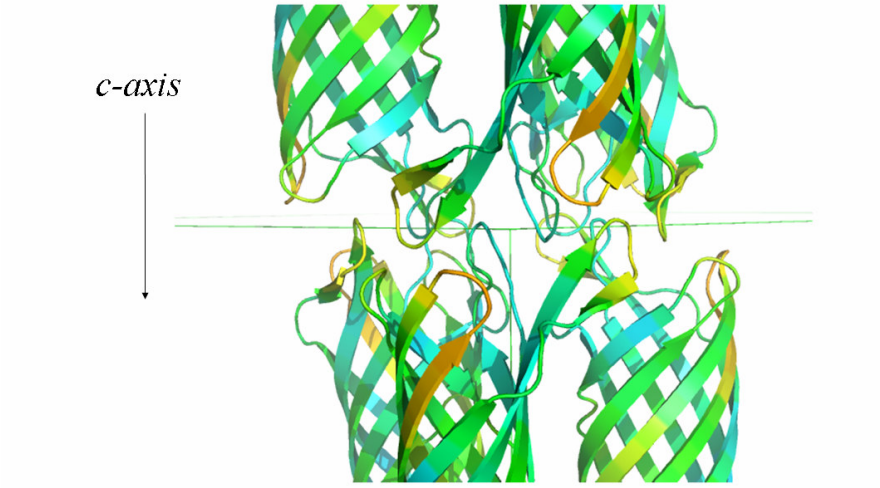


Fig. 5.8 Close up of the contact between two NspA molecules along the direction of the c-axis. The loops are mobile and a difference in conformation is likely to change the size of the c-axis. Colouring is according to B-factor

To understand the structural basis of the lattice distortion, we investigated the crystal packing (Fig. 5.6, 5.7). Layers of β-barrels are tightly packed in the ab-plane due to multiple contacts between the molecules (Fig. 5.6). The layers are only occasionally connected in the c-direction through loops, some of which are rather long (Fig. 5.7, 5.8). Conformational flexibility of the loops may lead to varying contacts and thus to

variation in the length of the c-axis. Since the reflection spots are not diffuse and neither clearly split, we interpret this phenomenon as the existence of multi-domain crystallites with varying length of c-axis. It must be concluded, that a distortion in the packing of molecules along the c-axis is the cause of the anomalous reflection shapes. We have developed the following procedure to incorporate lattice-distortion effects in EVAL15. The reciprocal axis matrix R_{mat} describes the orientation of the reciprocal axes $\mathbf{a}^*$, $\mathbf{b}^*$ and $\mathbf{c}^*$ in the goniostat-zero position (all goniostat angles are set to 0°).

$$R_{mat} = \begin{pmatrix} a_x^* & b_x^* & c_x^* \\ a_y^* & b_y^* & c_y^* \\ a_z^* & b_z^* & c_z^* \end{pmatrix}. \qquad \textbf{(5.1)}$$

To incorporate the lattice distortion effect, this matrix is deformed in a specific direction, namely that of the direct lattice vector ***lattvec*** = (l_x,l_y,l_z)**.** This vector is normalized and multiplied by *latt*/2, where *latt* is the magnitude of the distortion, giving

$$\begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} = \frac{latt}{2\,|\,lattvec\,|} \begin{pmatrix} l_x \\ l_y \\ l_z \end{pmatrix}. \qquad \textbf{(5.2)}$$

The components of vector $\mathbf{[v_x,v_y,v_z]}$ are added to a unity matrix giving the distortion matrix

$$U = \begin{pmatrix} 1+v_x & 0 & 0 \\ 0 & 1+v_y & 0 \\ 0 & 0 & 1+v_z \end{pmatrix}. \qquad \textbf{(5.3)}$$

The direct axes matrix D_{mat} is obtained by inverting the reciprocal lattice matrix R_{mat}. By multiplying U with D, the direct lattice is deformed as

$$U \cdot D_{mat} = \begin{pmatrix} 1+v_x & 0 & 0 \\ 0 & 1+v_y & 0 \\ 0 & 0 & 1+v_z \end{pmatrix} \begin{pmatrix} a_x & a_y & a_z \\ b_x & b_y & b_z \\ c_x & c_y & c_z \end{pmatrix} = \begin{pmatrix} a_x(1+v_x) & a_y(1+v_x) & a_z(1+v_x) \\ b_x(1+v_y) & b_y(1+v_y) & b_z(1+v_y) \\ c_x(1+v_z) & c_y(1+v_z) & c_z(1+v_z) \end{pmatrix}. \qquad \textbf{(5.4)}$$

Inversion of this matrix gives the distorted reciprocal lattice matrix R_{mat_latt}

$$R_{mat\_latt} = \begin{pmatrix} \frac{a_x^*}{1+v_x} & \frac{b_x^*}{1+v_y} & \frac{c_x^*}{1+v_z} \\ \frac{a_y^*}{1+v_x} & \frac{b_y^*}{1+v_y} & \frac{c_y^*}{1+v_z} \\ \frac{a_z^*}{1+v_x} & \frac{b_z^*}{1+v_y} & \frac{c_z^*}{1+v_z} \end{pmatrix}. \qquad \textbf{(5.5)}$$

Our goal is to generate reciprocal lattice vectors that are affected by the lattice distortion. The undistorted **S**-vector is generated with

$$R_{mat}\begin{pmatrix} h \\ k \\ l \end{pmatrix} = \begin{pmatrix} S_{0x} \\ S_{0y} \\ S_{0z} \end{pmatrix}, \qquad \textbf{(5.6)}$$

which is the normal of the reflecting plane h k l. When the Bragg-conditions are met at a rotation ω, the **S**-vector produces an impact on the detector. The distorted $\mathbf{S}_l$-vector is calculated in the same way using the R_{mat_latt} matrix. Vectors

$$\begin{pmatrix} S_x \\ S_y \\ S_z \end{pmatrix} - x\left[\begin{pmatrix} S_x \\ S_y \\ S_z \end{pmatrix} - \begin{pmatrix} S_{lx} \\ S_{ly} \\ S_{lz} \end{pmatrix}\right] \qquad \textbf{(5.7)}$$

are generated lying between the **S**-and $\mathbf{S}_l$-vector, by choosing x between -*latt*/2 and *latt*/2 according to a Lorentzian distribution with 3σ = *latt*/2 (Fig. 5.9). An asymmetric reflection profile can be obtained by limiting the distribution to a half Lorenztian function.

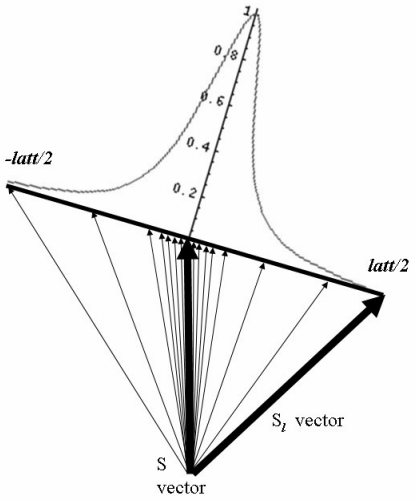


Fig. 5.9 Generated impacts around the central vector **S** follow a Lorentzian distribution.

We simulated new profiles with ***lattvec*** = (0,0,1) and *latt* = 0.035, while limiting x to positive numbers. The profile using this *latt*-distribution does now resemble the observed profile (Fig. 5.5g). With an optimized *latt* value of 0.035 the c-axis is therefore increased by 1.75 % at maximum. The other parameters are unchanged with respect to the initial model.
A couple of examples of observed and predicted profiles with and without the implemented lattice distortion are shown in three-dimensions (Fig. 5.10-5.12). fom_{peak}

is an indicator of the profile fit (explained in chapter 2). Indeed, the fom_{peak} of the improved profiles is lower than that of the initial model (Table 5.1). This is especially the case with |*l*-index| >15, where the asymmetric shape occurs most clearly.
Integration was carried out with the initial and the improved model. The merging statistics calculated with Xprep (Bruker AXS, Madison, USA) shows that the data with the *latt*-model are slightly better (Table 5.2). A correlation plot of the data of the initial model against that of the improved model reveals that the intensities are not similar, especially for reflections with higher *l*-index (plot not shown here). However, the majority of the anomalous ΔF's of both data are alike, although some differ and cause a slightly different anomalous Patterson. The anomalous signal of Au is very strong, which makes the location of the heavy atom sites rather straightforward. At the same time this makes it more difficult to compare the phasing quality. Au-sites were successfully located through anomalous scattering with SHELXD (Sheldrick, 2008) using data up to 4.5 Å and are quite similar for both data. The correlation coefficient CC_E between E_o and E_c, where E_c is calculated from the located sites and E_o is derived from the observed F_A's, is slightly higher for the data with the initial model (20.80% compared to 17.99%). However, its Pseudo-free CC after SHELXE is lower (51.98% compared to 57.78%). The electron density maps of both data are very similar and the β-barrels are clearly visible. The conclusion is, that both data are of good quality, and that they are not very sensitive to the tail of the profile. However, structural refinement could show different results due to the differently integrated high resolution reflections.

Table 5.1 Quality of two profile models

(I)		Initial model	Improved model
		without *latt*	with *latt*
h k l	ω(°)	fom_{box}/fom_{peak}	fom_{box}/fom_{peak}
(-8 17 -25)	0	2.14/6.75	1.75/3.38
(-1 11 -6)	45	2.08/4.31	1.99/2.27
(5 6 8)	90	2.01/2.14	1.98/2.28
(1 12 -29)	180	2.24/6.54	1.83/2.27
$<fom_{peak}>$	|*l*|>15	2.42	1.98
$<fom_{peak}>$	|*l*|<15	2.91	2.76
$<fom_{peak}>$	*l=0*	3.91	3.91
$<fom_{peak}>$	Complete data	2.64	2.43
(II)		Initial model	Improved model
		$w_{\lambda1}$=2.0, $w_{\lambda2}$=1.0	$w_{\lambda1}$=2.0,$w_{\lambda2}$=0.6
h k l	θ(°)	fom_{box}/fom_{peak}	fom_{box}/fom_{peak}
(4 -10 4)	25	0.95/1.4	0.93/1.25
(-1 10 39)	26	0.99/1.65	0.97/1.48
(-7 -7 39)	26	1.07/2.31	1.03/1.99
(-3 0 16)	8	2.89/4.18	3.27/4.62
(2 3 -17)	9	2.07/8.22	2.19/8.00
$<fom_{peak}>$	θ < 23°	2.85	2.83
$<fom_{peak}>$	θ > 23°	1.11	1.08
$<fom_{peak}>$	Complete data	2.46	2.43

(I)= Neisserial Surface Protein A (NspA), (II)= $Mn(OTF)_2(Py(ProOH)_2)$, where $Py(ProOH)_2$ is 2,6-Bis[[(s)-2-(hydroxymethyl)-1-pyrrolidinyl]methyl]pyridine and OTF is the triflate F_3CSO_3]

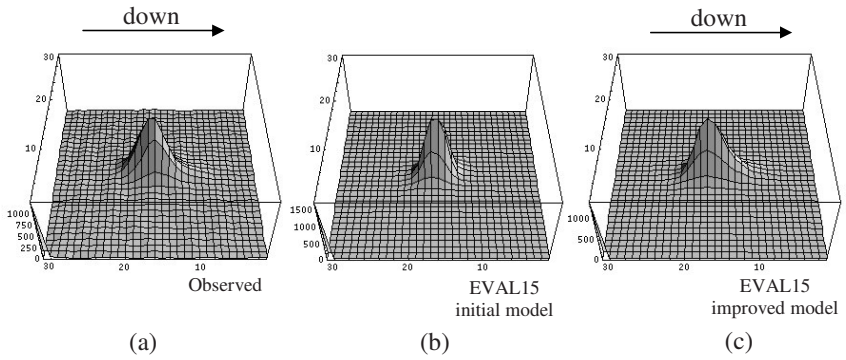


Fig. 5.10 (a) The observed profile of reflection -8 17 -25, (b) the predicted profile of EVAL15 with and (c) without *latt* are displayed in three-dimensions. The asymmetric tail points downwards to the equatorial line of the detector. This reflection is found at ω = 0º and at the sector 1-3.

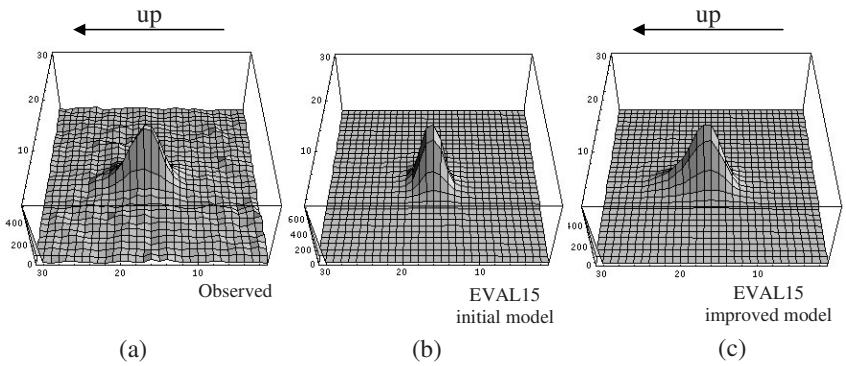


Fig. 5.11 (a) The observed profile of reflection 1 12 -29, (b) the predicted profile of EVAL15 with and (c) without *latt* are displayed in three-dimensions. The asymmetric tail points upwards to the equatorial line of the detector. This reflection is found at ω = 180º and at the sector 3-1.

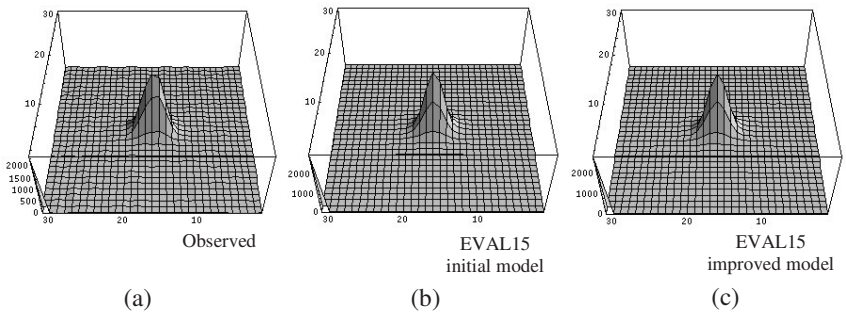


Fig. 5.12 (a) The observed profile of reflection -1 11 -6, (b) the predicted profile of EVAL15 with and (c) without *latt* are displayed in three-dimensions. At ω = 45º the tail is not visible and the predicted profiles with and without latt are both similar to the observed profile.

Table 5.2 Data statistics after scaling and merging

	R_{merge} (Friedels merged)*	R_{merge} (Friedels not merged)*	completeness	I/σ*	# data	# unique data*	redundancy
(I)	resolution 4.0 (4.10-4.0)Å						
EVAL15 (initial model)	0.058 (0.098)	0.043 (0.080)	99.6	30.77 (20.12)	25072	2833 (196)	8.82
EVAL15 (improved model)	0.057 (0.094)	0.042 (0.029)	99.6	32.22 (20.76)	24699	2834 (194)	8.68
(II)	resolution 0.77 (0.90-0.77)Å						
EVAL15 (initial model)	0.023 (0.056)	0.015 (0.042)	99.6	63.29 (24.69)	27025	3842 (1366)	7.06
EVAL15 (improved model)	0.022 (0.056)	0.015(0.041)	99.3	63.49 (24.45)	25904	3828 (1367)	6.72

*number in parentheses refer to the highest resolution shell.

Table 5.3 Refinement results

	R(strong/all)	#(strong/all)	$\Delta\rho_{max/min}$(Å^{-3})	S	wR_2	WGHT(a b)
(II)						
EVAL15 (initial model)	0.023/0.026	5941/6218	0.36/-0.25	1.082	0.0584	0.03 2.00
EVAL15 (improved model)	0.023/0.025	5940/6217	0.36/-0.27	1.07	0.0577	0.03 1.92

$R=\Sigma||F_{obs}|-|F_{calc}||/\Sigma|F_{obs}|$

$\Delta\rho_{max/min}$=maximum and minimum difference density

$S=\{\Sigma[w(F_o^2-F_c^2)^2/(n-p)]\}^{1/2}$, where n = number of reflections, p = number of parameters refined

$wR_2=\{\Sigma[w(F_o^2-F_c^2)]^2/\{\Sigma[w(F_o^2)^2]\}^{1/2}$

WGHT = weighting scheme: $w=1/[\sigma^2(F_o^2)+(aP)^2+bP]$, where $P=(F_o^2+2F_c^2)/3$

5.3 Unexpected λ-distribution in the data of $Mn(OTF)_2(Py(ProOH)_2$

The data of $Mn(OTF)_2(Py(ProOH)_2$ (II) (by courtesy of Martin Lutz) were collected in-house on a Nonius KappaCCD, with a graphite monochromator and rotating anode MoK_α radiation, using fine slicing of 0.3º per image to a resolution of 0.77 Å. The molecule crystallized in space group $P4_12_12$ with cell dimensions: a,b = 9.02 Å and c = 66.48Å.

The initial model was simulated with the following parameters: point spread $\gamma = 0.5$ pixels, *focus dist* = 150 mm, *focus width* = 0.3 mm; the crystal is face indexed with the shortest dimension 0.2 mm and longest dimension 0.4 mm, mosaicity $\mu = 0.3$º and a double Lorentzian λ-distribution is obtained with $\lambda_{\alpha1}/\sigma_{\alpha1}/w_{\alpha1}$ = 0.71073/0.0001/2 and $\lambda_{\alpha2}/\sigma_{\alpha2}/w_{\alpha2}$ = 0.71359/0.0001/1, where $w_{\alpha1}$ and $w_{\alpha2}$ are the relative weights.

Characteristic radiation has a theoretical $K_{\alpha1}/K_{\alpha2}$-ratio of 2. At high θ, the observed profiles show a clear $K_{\alpha1}$ and $K_{\alpha2}$ separation (Fig. 5.13a-5.15a), but their ratio is not 2:1 as it should be (Fig. 5.13b-5.15 b). We only get similar profiles in our prediction, if we use $w_{\alpha1}:w_{\alpha2}$ = 2:0.6 (Fig. 5.13c-5.15c). The λ-distribution is shown in Fig. 5.16. The model is only slightly improved in terms of fom_{peak} (Table 5.1), although the profiles look significantly different. The merging and refinement statistics are surprisingly similar for the initial and improved model (Table 5.2, 5.3). However, close inspection learns that the thermal displacement parameter u is systematically smaller by 1~2% for the improved model and the averaged ratio between F_{obs}^2 and F_{calc}^2 is systematically closer to one for the high resolution range. Very high resolution data could be more sensitive to the ratio of $K_{\alpha1}$,$K_{\alpha2}$. An explanation for the unexpected K_α-distribution could be that the monochromator is not optimally aligned. Therefore, not all $K_{\alpha2}$-radiation passes the monochromator.

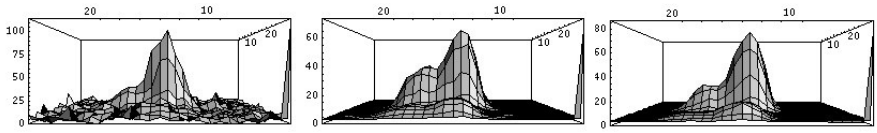


Fig. 5.13 (a) The observed profile of reflection 4 -10 4, (b) the initial model of EVAL15 and (c) the improved model in three-dimensions. The profiles are summed over a few successive ω-frames.

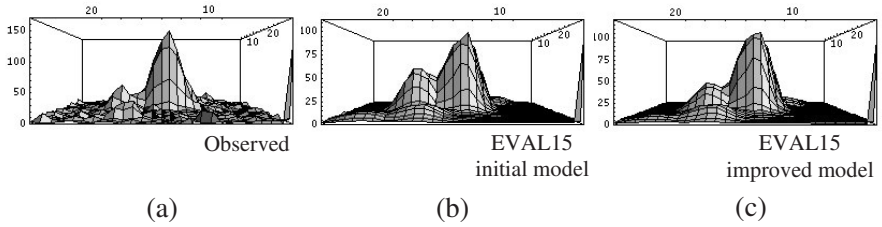


Fig. 5.14 (a) The observed profile of reflection 1 -10 39, (b) the initial model of EVAL15 and (c) the improved model in three-dimensions.

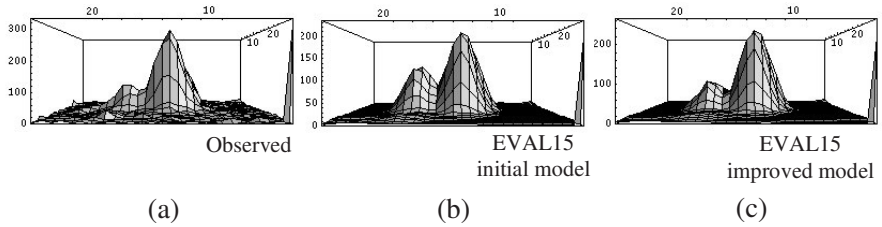


Fig. 5.15 (a) The observed profile of reflection -7 -7 39, (b) the initial model of EVAL15 and (c) the improved model displayed in three-dimensions.

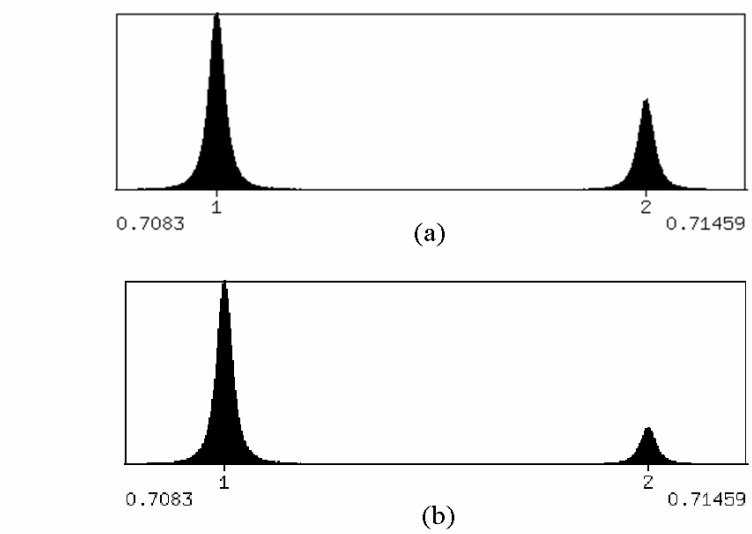


Fig. 5.16 (a) the initial spectrum with $w_{\alpha 1}$:$w_{\alpha 2}$ =2:1 and (b) the improved spectrum with $w_{\alpha 1}$:$w_{\alpha 2}$ =2:0.6.

5.4 Detector read-out and distortion correction errors

Diffraction data of lysozyme were collected on an ADSC Q4R detector at the ESRF Beam line ID14-3 in Grenoble, France. The molecules crystallized in space group $P4_32_12$ with the cell dimensions: a,b = 79.29 Å and c = 37.21 Å.
One diffraction image is shown in Figure 5.17. The close-up of the lower left side of the detector shows stretched profiles in the horizontal direction. This is not the case at the right side of the detector (Fig. 5.18). After a couple of degrees ω-rotation this phenomenon disappears. This implies that this effect cannot be due to anisotropic pointspread or to distortion correction errors (later we will show an example of this). At first we attempted to improve our profile by incorporating the previously introduced lattice distortion effects, but soon it became clear we had to attribute the effects to read-out errors. Another diffraction image shows clear read-out errors (Fig. 5.19). In the lower left panel reflections have large streaks in the horizontal direction and this suggests that when read-out errors occur in this area, they are in the horizontal direction. Such errors may have affected the image in Fig. 5.17.
This kind of occasionally occurring problems cannot be incorporated in the profile prediction. However, profile-learning methods where profiles are obtained by averaging reflection spots on nearby positions on the detector, may be in advantage, because similar artefacts occur both for the reflection and the learned profiles.

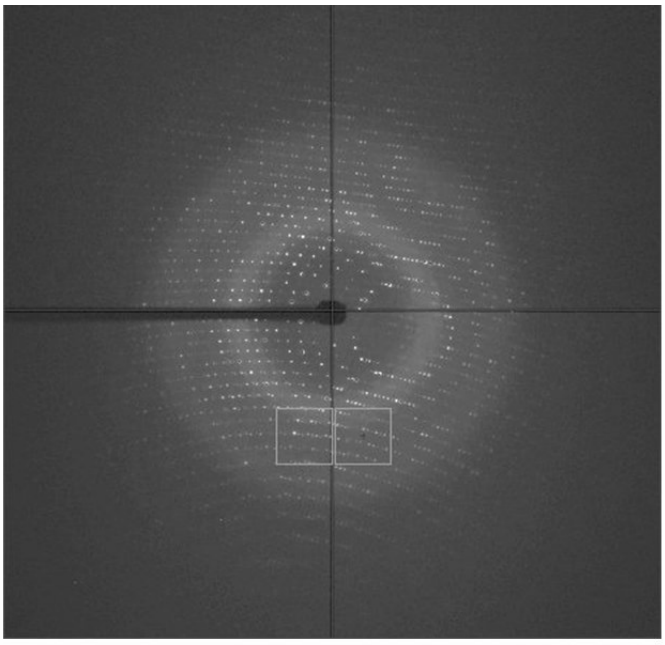


Fig. 5.17 An diffraction image of the lyzosyme data.

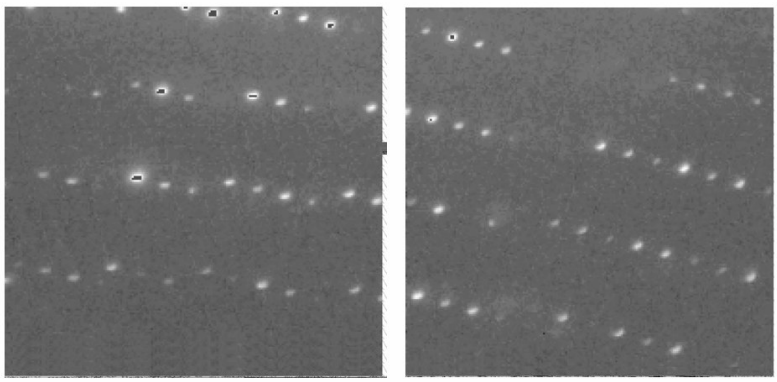


Fig. 5.18 (a) close-up of the left lower side of the diffraction image and (b) close-up of the right lower side of the diffraction image in Fig. 5.17.

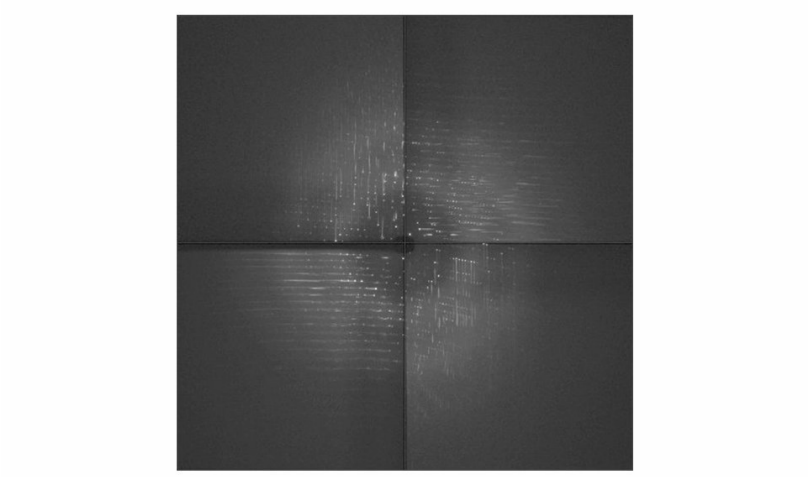


Fig. 5.19 A diffraction image with obvious readout-errors.

Hemagglutinin-esterase (Zeng *et al.*, 2008) was crystallized with 4 molecules per asymmetric unit. X-ray data (kindly provided by Qinghong Zeng) were collected at the ESRF on Beamline ID14-3 up to 2.3 Å on a marmosaic225 detector. The molecules crystallized in the space group $P2_12_12$ with the cell dimensions: a = 67.63 Å, b = 113.28 Å and c = 273.52 Å.

For the integration of the intensity the accuracy of the predicted reflection position is vital (Pflugrath, 1999). By minimizing the difference between the observed and predicted impact positions with the simplex method (Press *et al.*, 1986) in PEAKREF (Schreurs, 1999), the cell orientations, as well as the detector and crystal position (horizontal, vertical and rotational direction) are refined. This leads to more accurately predicted impact positions, but these are never exactly the same as those of the observed reflections. As even sub-pixel differences can produce a significant

difference in the calculated intensity, EVAL15 shifts predicted reflections slightly by minimizing fom_{box} (see chapter 2), such that they coincide as well as possible with the observed reflections.
After integration of the data, we observed a peculiar systematic behaviour of the horizontal shift in a certain area of the detector. The detector is made up of 9 CCD-chips with 2048 x 2048 pixels per chip. Just below the boarder between two detector panels, the reflections all shift to the positive horizontal direction and thus their positions are systematically wrongly predicted (Fig. 5.20). We think this is caused by a wrong distortion correction.
The marmosaic225 detector is a position sensitive detector. X-rays hit the detector through a beryllium window and excite the phosphor on the fluorescent screen. The resultant visible light is transmitted through a fibre optic taper, which reduces the image to the size of the CCD, where the light is converted into electrons. This signal is read out and fed into a computer (Fig. 5.21). The image is distorted because of the taper. The distortion is measured during production of the detector with a mask with holes: the grid image. This grid image is fitted by a 2D-distortion polymonal. Using the polynomal the pixel position in the image can be back-calculated to impact position on the detector. If the distortion correction is not correct, the impact positions would be wrongly predicted and the profile shifts will be large and systematic.

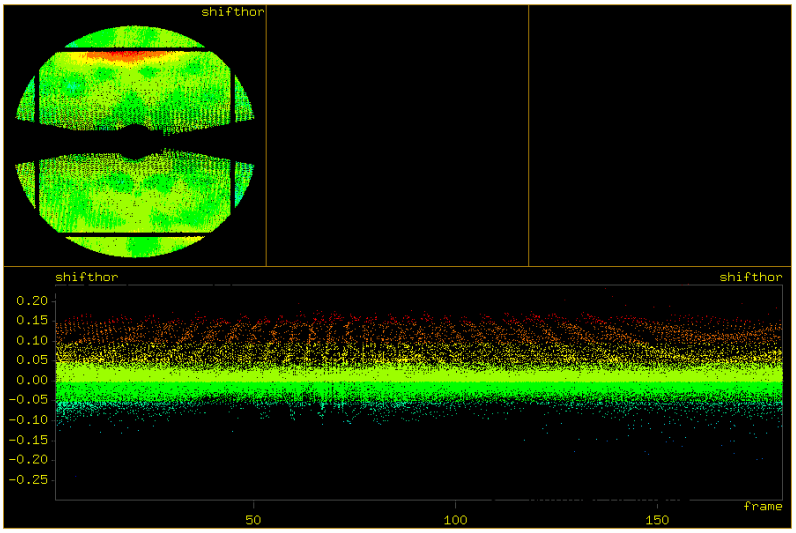


Fig. 5.20 Top: One area of the detector shows systematic horizontal shifts (red area). The rectangular gaps are the borders between detector panels, which are excluded from the integration.
Bottom: The shift of reflections in the horizontal direction versus the frame number. The positive horizontal shifts are coloured red, the negative blue and the shifts in between yellow and green.
There is clearly a correlation between the large horizontal shift and the position on the detector.

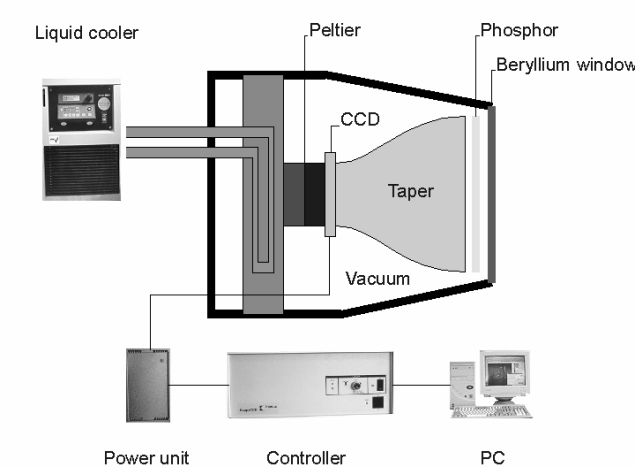


Fig. 5.21 An area detector coupled to a CCD-chip. X-ray photons enter on the right and are transformed into light photons by the phosphor screen. The CCD is cooled by a Peltier element stack. This picture is copied from www.nonius.nl.

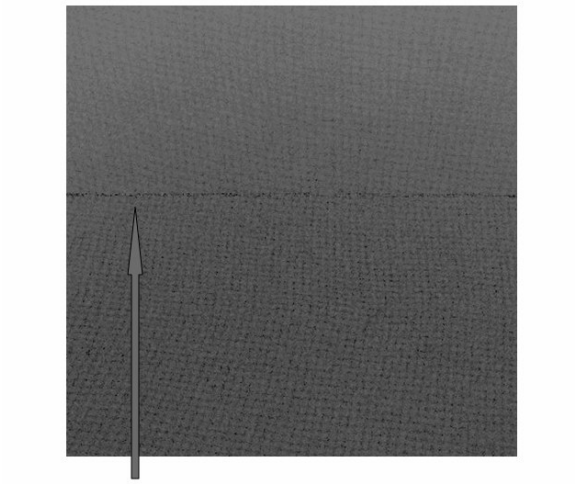


Fig. 5.22 Close-up of the grid after distortion correction in the problem area.

An image is made of the background scattering on the detector (VIEW low3-image (Schreurs, 1998)). Figure 5.22 is a close up of that image and shows the problem area of the detector. The grid lines seen in Fig. 5.22 show how the distortion changes over the detector. It is obvious that the two panels would need their own distortion polynomal. For us it is not clear how the manufacturer treats the distortion across the discontinuous boarder between the panels. In all cases, we suspect that distortion correction errors are introduced in this area.

5.5 Concluding remark

Studying profiles in detail and comparing with predicted profiles based on physically realistic parameters can give insight into specific crystal packing effects or instrumental characteristics.
The lattice distortion effect of NspA is caused by a variation of the c-axis of the unit cell. The loops connecting strands in the β-barrel are mobile and each conformation of the loops leads to a different length of the c-axis. This effect produces an asymmetric tail in the reflection profile. At this moment the lattice distortion parameter *latt* is incorporated in the profile prediction in one direction. Such an effect could potentially be found in two or even three dimensions, especially at higher space group symmetries.
We noticed that with sealed tube radiation, using graphite monochromator, sometimes the $K_{\alpha1}/K_{\alpha2}$ ratio is not as it should. We suspect that the monochromator is not optimally aligned.
Detector read-out and distortion correction errors can cause a deviation from the expected profiles or can lead to problems in the prediction of reflection positions. An alternative profile model cannot solve these problems.
Improving our profile methods to incorporate the first two effects lowers the fom_{peak}'s, but does not influence the data significantly. In the case of a different $K_{\alpha1}/K_{\alpha2}$ ratio, the improved profile mainly changes high-resolution reflections and the overall effect on the data is small. However, for charge density studies (0.3 Å resolution) detailed accuracy of the profile may be of importance. Finally, we have shown that when one thoroughly investigates what lies beneath the surface of a profile, unexpected effects can be discovered.

References

Otwinowski, Z. & Minor, W. (1997). *Macromolecular Crystallography, Pt A* **276**, 307-326.
Pflugrath, J. W. (1999). *Acta Cryst.* D**55**, 1718-1725.
Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1986). *Numerical Recipes*. Cambridge: Cambridge University Press.
Schreurs, A. M. M. (1998). *VIEW.* University of Utrecht, The Netherlands.
Schreurs, A. M. M. (1999). *PEAKREF.* University of Utrecht, The Netherlands.
Schreurs, A. M. M., Xian, X. & Kroon-Batenburg, L. M. J. (2009). *J. Appl.Cryst.*
Sheldrick, G. (2008). *Acta Cryst.* A**64**, 112-122.
Vandeputte-Rutten, L., Bos, M. P., Tommassen, J. & Gros, P. (2003). *J. Biol. Chem.* **278**, 24825-24830.
Zeng, Q., Langereis, M. A., van Vliet, A. L., Huizinga, E. G. & de Groot, R. J. (2008). *Proc. Natl. Acad. Sci. U S A* **105**, 9065-9069.

Summary and Conclusions

Summary and Conclusions

Molecular structures can be determined in X-ray crystallography up to an atomic level. The molecules are crystallized and then exposed to X-radiation by which Bragg reflections are produced. Their intensities can be measured and are related to the structure factor, but the phase information is lost. Both are needed in order to determine the molecular structure. The phases can be retrieved from additional experimental or structural information and are refined during density modification. With the structure factors and phases, the electron density distribution in the unit cell is calculated by an inverse Fourier transformation and the molecular structure is revealed.

This thesis deals with an essential part in structure determination, i.e. the accurate integration of reflection intensities. The two main goals to be achieved are to get the most out of weak reflections and to solve complicated diffraction problems. High-resolution reflections are often weak, especially in case of protein crystals. Several benefits are obtained if these weak reflections are integrated accurately: they contribute to a higher resolution in the electron density map; lead to lower FOM's and lower standard deviations of the atomic coordinates.

The application of commonly used data integration programs is mostly limited to "normally" scattering single crystals. However, problems associated with anisotropic reflection shapes (which can have different orientations on the detector and can depend on the ω-rotation), anisotropic mosaicity, lattice distortion, $K_{\alpha 1}$,$K_{\alpha 2}$-splitting, overlapping reflections (either due to a long cell axis or to multiple lattices), satellite reflections and fine slicing, need to be solved. We developed the data integration method EVAL15 that predicts *ab initio* 3-dimensional profiles using a ray-tracing algorithm. As this method does not use analytical expressions for the various broadening effects, it is relatively easy to incorporate additional effects. Unlike profile learning methods, EVAL15 can integrate reflections in area's of reciprocal space where no strong reflections occur like at high resolution. The prediction algorithm also takes into account all variations in the reflection profile due to geometrical deformation. An additional advantage of our method is that careful inspection of the profiles can reveal specific physical properties of the crystal and unexpected characteristics of the diffraction experiment.

Chapter 1 introduces crystallography in general and the history of data integration in particular.

In Chapter 2 we describe the new data integration method EVAL15. The predicted profile is generated based on the principle of general impacts, whereby impacts of X-rays are traced originating from random sample points of the focus, the crystal, the wavelength spectrum and the mosaic distribution. The simulated profile needs to be convoluted with a point-spread function of the detector and is then used in a profile fitting algorithm using Singular Value Decomposition. This chapter describes the non-uniform distributions of the parameters used in the profile prediction and the derivation of standard deviations.

In chapter 3 the quality of EVAL15-data integration is assessed for several standard diffraction experiments. We investigate the merging statistics and for small molecule data we verify the refinement residuals, difference electron densities and the weighting scheme, whereas for protein data the phasing quality and the resulting electron densities are important indicators. A procedure to optimize the prediction parameters is developed. In practice only a few high I/σ reflections are needed. We compared the EVAL15-data with that of EVAL14 (its predecessor that uses

summation integration) in particular to investigate if the EVAL15 profile method results in improved weaker data. Indeed, we found that these are significantly improved and that this is notable especially in the phasing quality.
The capacity of EVAL15 to deal with spatially overlapping reflections such as caused by multiple lattices or a long cell axis is established in chapter 4. It is shown that EVAL15 successfully deconvolutes overlapping reflections through a least-squares fit using Singular Value Decomposition. The quality of the deconvoluted reflections is similar to that of single reflections. The gained reflections contribute to a higher completeness, redundancy and improvement in the electron density map. In general, we find that the higher completeness achieved by deconvolution of overlap only has a small penalty on R_{merge}.
In the last chapter some profiles are studied in detail, which provides insight into specific crystal packing effects or instrumental characteristics. In a protein data set, we found asymmetric reflection profiles caused by variation of one unit cell dimension. This lattice distortion effect is incorporated into the predicted profile with a new physical parameter, the lattice distortion *latt*. In one case, where the data were recorded with sealed tube radiation, we learned that reflection profiles can show an abnormal distribution of $K_{\alpha 1}$ and $K_{\alpha 2}$ radiation. This is probably due to a non-optimally aligned graphite monochromator. This chapter also shows that the reflection profile prediction can indicate possible errors caused by distortion correction or by the read-out process of the detector. Although we can try and predict profiles in as much detail as we want, we found this is not absolutely needed to derive data of good quality. Apparently the predicted profile is accurate enough even without trying to incorporate the minor details.

This thesis shows that it is possible to make ab initio predictions of reflection profiles to high accuracy using only a small number of physical parameters. The method of general impacts, on which the predictions are based, allows full flexibility to incorporate all sorts of crystal and instrument characteristics and is implemented in EVAL15. The profiles are used in accurate integration of reflections through least-squares minimization using Singular Value Decomposition. The data obtained are of good quality both for small molecules and proteins, and overlap deconvolution, which is straightforward in EVAL15, is successful. Deviation between the prediction and observation in EVAL15 can also be used as a diagnostic tool for specific physical crystal properties. This method has great potential in solving complicated problems where other programs fail, like those related to aperiodic crystals, accurate electron density studies, anisotropic crystal shapes/mosaicity and lattice distortion.

Samenvatting en Conclusies

Samenvatting en Conclusies

Moleculaire structuren kunnen door Röntgen kristallografie tot in atomaire details worden bepaald. De moleculen worden gekristalliseerd en door de blootstelling aan Röntgen stralen ontstaan zogenaamde Bragg reflecties. De gemeten intensiteiten zijn gerelateerd aan de Structuurfactor, maar de fases zijn niet te meten. Om de moleculaire structuur te kunnen bepalen, zijn echter beiden nodig. De fases kunnen verkregen worden door aanvullende experimentele of structurele informatie en worden verfijnd door dichtheidsmodificatie. Als eenmaal de structuurfactoren en fases bekend zijn, wordt de elektronendichtheid in de eenheidscel door een inverse Fourier transformatie berekend en de moleculaire structuur wordt daardoor onthuld.
Dit proefschrift behandelt een essentieel deel in het structuurbepalingsproces, namelijk de nauwkeurige integratie van reflectie intensiteiten. Hierbij zijn de twee hoofddoelen als volgt: we willen de informatie van zwakke intensiteiten maximaal kunnen benutten en gecompliceerde diffractie patronen kunnen oplossen. Reflecties, die in het hoge resolutie gebied worden gedetecteerd, zijn vaak zwak, vooral in het geval van eiwitkristallen. De nauwkeurige integratie van deze zwakke reflecties levert meerdere voordelen op: ze dragen bij tot een hogere resolutie in de elektronendichtheid en leiden tot lagere FOM's en ook lagere standaarddeviaties van de atomaire coördinaten.
Gebruikelijke data integratie programma's kunnen meestal alleen worden gebruikt voor eenkristallen die "normaal" verstrooien. Echter, problemen veroorzaakt door bijvoorbeeld anisotrope reflectievormen (welke verschillende oriëntaties op de detector kunnen hebben en ook van de ω-rotatie af kunnen hangen), anisotrope mosaiciteit, roosterdistortie, $K_{\alpha 1}$,$K_{\alpha 2}$-splitsing, overlappende reflecties (door een lange cel-as of door meerdere roosters), satelliet reflecties en "fine slicing" wachten nog op oplossingen. Wij hebben de data-integratiemethode EVAL15 ontwikkeld, die *ab initio* 3-dimensionale profielen voorspelt, door gebruik te maken van een "ray-tracing" algoritme. Aangezien deze methode geen analytische vormen gebruikt om verbredende effecten te incorporeren, kunnen bepaalde effecten in de reflectievorm gemakkelijk worden nagebootst. In tegenstelling tot programma's die afhankelijk zijn van profielleren, kan EVAL15 reflecties in gebieden van de reciproke ruimte integreren waar geen sterke reflecties aanwezig zijn, zoals bij hoge resoluties. Alle variaties in de profielvorm die door geometrische deformatie ontstaan, worden door het voorspellingsalgoritme in rekening gebracht. Onze methode levert nog een voordeel op: door zorgvuldige inspectie van het profiel kunnen bepaalde specifieke fysische eigenschappen van een kristal en onverwachte karakteristieken van het diffractie-experiment worden onthuld.
Hoofdstuk 1 behandelt kristallografie in het algemeen en de geschiedenis van data integratie in het bijzonder.
In hoofdstuk 2 beschrijven we de nieuwe data-integratiemethode EVAL15. Het voorspelde profiel wordt op basis van het principe van "general impacts" gegenereerd, waarbij verstrooide straling wordt gevolgd vanuit willekeurige "sample points" van het Focus, het kristal, het golflengtespectrum en de mosaicverdeling. Voordat het gesimuleerde profiel in een profielfittingsalgoritme met de "Singular Value Decomposition" gebruikt kan worden, moet het nog geconvolueerd worden met een "point-spread"-functie, die afhankelijk is van het type detector. Dit hoofdstuk beschrijft de niet-uniforme verdeling van de parameters, die gebruikt worden in de profielvoorspelling en ook de afleiding van de standaarddeviaties.

In hoofdstuk 3 wordt de kwaliteit van de EVAL15-data-integratie voor verschillende standaard diffractie-experimenten bepaalt. We onderzoeken de "merging"-statistiek en ook de verfijningsresiduen, elektronen-dichtheidsverschillen en het gewichts-schema voor kleine moleculen. Voor eiwit data zijn de faseringskwaliteit en de resulterende elektronendichtheid belangrijke indicatoren. Er is een procedure ontwikkeld, die de voorspelling van parameters optimaliseert, hierbij is alleen een aantal hoge I/σ reflecties nodig. We vergelijken de EVAL15-datakwaliteit met die van EVAL14 (zijn voorganger, waarbij sommatie-integratie wordt gebruikt), om te onderzoeken of de EVAL15 profielmethode de zwakke data verbetert. Dit is inderdaad het geval. De significante verbetering van zwakke data is in het bijzondere merkbaar in de faseringskwaliteit.
De capaciteit van EVAL15 om met deels overlappende reflecties, die door meerdere roosters of een lange cel-as worden veroorzaakt, wordt onderzocht in hoofdstuk 4. De succesvolle deconvolutie van overlappende reflecties door EVAL15 met behulp van een kleinste-kwardraten-fit met Singular Value Decompositie wordt getoond. De kwaliteit van de gedeconvolueerde reflecties is vergelijkbaar met die van enkelvoudige reflecties. De op deze wijze verkregen reflecties dragen bij tot een hogere compleetheid, "redundancy" en een verbetering in de elektronendichtheid. In het algemeen is de hogere compleetheid die bereikt wordt door de deconvolutie van overlap, gepaard met een iets hogere R_{merge}.
In het laatste hoofdstuk worden sommige profielen in detail bestudeerd, waardoor inzichten worden verkregen in specifieke kristalpakkingseffecten of instrumentele karakteristieken. We hebben asymmetrische reflectieprofielen in een eiwit data set verkregen, welke wordt veroorzaakt door de variatie van een eenheidscelldimensie. Dit roosterdistortie-effect wordt met behulp van een nieuwe fysische parameter, de roosterdistortie *latt*, geïncorporeerd in het voorspelde profiel. In een experiment, waar de data met behulp van "sealed tube radiation" wordt opgenomen, ontdekken we dat sommige reflectieprofielen een abnormale $K_{\alpha1}$,$K_{\alpha2}$-verhouding vertonen. Dit is waarschijnlijk terug te voeren tot een niet optimaal ingestelde grafietmonochromator. Dit hoofdstuk laat ook zien, dat de voorspelling van reflectieprofiel mogelijke fouten, veroorzaakt door distortiecorrectie of read-out van de detector, kan ontdekken. Het is gebleken dat, om data van goede kwaliteit te verkrijgen, niet per se profielen nodig zijn, die tot in het kleinste detail voorspeld zijn. Blijkbaar zijn de voorspelde profielen nauwkeurig genoeg, ook al doen we niet ons uiterste best om kleine details te beschrijven.

Dit proefschrift laat zien dat het mogelijk is om een nauwkeurige ab initio voorspelling met behulp van alleen een klein aantal fysische parameters van reflectieprofielen te maken. De "general impacts"-methode, waarop deze voorspellingen gebaseerd zijn, maakt het mogelijk om alle soorten van kristal- en instrumentkarakteristieken te incorporeren en deze methode is in EVAL15 geïmplementeerd. Deze profielen worden gebruikt in de nauwkeurige integratie door de kleinste-kwardraten methode met Singular Value Decomposition. Zowel voor de kleine moleculen als voor de eiwitten data zijn de verkregen data van goede kwaliteit en de overlapdeconvolutie, welke in EVAL15 heel gemakkelijk uitgevoerd kan worden, is succesvol. Verschillen tussen de voorspelling in EVAL15 en de observatie kunnen gebruikt worden als een diagnose voor specifieke fysische kristaleigenschappen. Deze methode heeft een enorm potentieel om gecompliceerde problemen waar andere programma's moeite mee hebben, zoals gevallen met

aperiodieke kristallen, nauwkeurige elektronendichtheidstudies, anisotrope kristalvormen/mosaiciteit en roosterdistortie, op te lossen.

Curriculum Vitae

De auteur van dit proefschrift werd op 8 december 1976 geboren in Qingdao te China. Zij behaalde haar Abitur (duitse VWO-diploma) aan het Gymnasium St.Ursula in Aachen in 1996. Van 1997 tot 1998 studeerde ze aan de Universiteit Utrecht wiskunde en vanaf 1998 studeerde ze aan dezelfde universiteit scheikunde. In september 2003 werd het doctoraal scheikunde behaald. Het bijvak werd in de afdeling theoretische chemie verricht onder begeleiding van Dr. Joop van Lenthe en het hoofdvak voltooide ze in de afdeling Van 't Hoff laboratory for Physical and Colloid Chemistry onder leiding van Dr. Andrei Petukhov. Tijdens het hoofdvak heeft ze ook als studentenassistente gewerkt bij experimenten in de synchrotron te Grenoble.
In januari 2004 trad ze in dienst als assistent in opleiding bij de sectie Kristal- en Structuurchemie aan de Universiteit Utrecht, gefinancierd door STW. In deze functie werd het in dit proefschrift beschreven onderzoek verricht onder begeleiding van Dr. Loes Kroon en Prof.dr. Piet Gros.

Dankwoord

Als je het dankwoord schrijft, dan is je proefschrift echt af en het werk van bijna 5 jaar loopt ten einde.
Als ik terug kijk op deze tijd, kan ik het het beste met een lange bergwandeling vergelijken. Stapje voor stapje kom je verder, soms is de wandeling moeilijk en steil, soms word je beloond door prachtige uitzichten. Dankzij menige mensen heb ik deze wandeling af kunnen leggen.

Graag zou ik allereerst Loes willen bedanken. Bedankt, dat ik met mijn vragen altijd bij jou terecht kon en voor je kritische houding ten opzichte van mijn werk. Je hebt me in zoveel opzichten geholpen, zowel op wetenschappelijk gebied, persoonlijk gebied als ook op het laatst met alle administratieve dingen, die komen kijken bij een promotie. Bedankt voor al je inzet en moeite! Toine: jij, Loes en ik vormden het team dat aan EVAL15 werkte. Bedankt dat ik jou altijd lastig kon vallen met vragen, en zonder jou zou EVAL15 niet mogelijk zijn. Piet, bedankt voor je begeleiding tijdens mijn onderzoek!
Mijn (oud-)collega's zou ik graag bedanken voor de gezelligheid tijdens het werken. Huub, bedankt voor je goede uitleg, toen ik voor het eerst studenten werkcolleges gaf. Hans, bedankt voor je hulp om me wegwijs te maken in de CCP4. Martin, je bent een bron van kennis, niet alleen op het gebied van kristallografie. Bedankt voor je small molecule data en je behulpzaamheid. Chiara, it was great to have you as a room-mate and shared cookies are indeed more delicious. Hope to see you soon in Finland! Lucy, Harma, Els, Qinghong en Jin: onze lab-girls avond-etentjes waren altijd leuk en lekker (de hoelahoep-wedstrijd zal ik niet vergeten). Lucy, bedankt voor je data van NspA, je hulp bij het (her)oplossen ervan en ook vooral voor de gezelligheid buiten het werk, zoals met klimmen, dim-sumen enzovoort! Harma, je Terry Pratchett boeken hebben zeker geholpen bij de ontspanning. Els, je bent een gezellige kamergenoot en bedankt voor de lekkere pannenkoeken. 曾庆红，邬金 ，我很高兴有你们这两位热心地，友好地中国同事.在此我也谢谢庆红给我的蛋白数据和帮助. Bert, Fin, Michael, Arjen en Lucio bedankt voor de gezelligheid tijdens het werken in de terminal room en ook voor het gebruiken van jullie (Fin, Bert en Michael) eiwit data! Verder wil ik nog graag Arie, Ton, Roland, Wieger, Dennis, Lars, Eric, Wietske, Aike, Rachel en Tom bedanken voor de gezellige gesprekken tijdens de koffie- en lunch-pauzes. Marjan en Kaoutar, erg bedankt voor het regelen van de administratie. Graag wil ik ook Andrea bedanken voor haar enthousiaste inzet tijdens de 3-maanden-stage aan EVAL15.
Madhumati Sevvana uit de groep van George Sheldrick in Duitsland wil ik graag danken voor de data van glucose isomerase en de mooie foto van het vertwinnde kristal.
Er waren menige vergaderingen over EVAL15 met Frank van Meurs, Mike Hursthouse, Raimond Ravelli en Anastassis Perrakis. Bedankt voor jullie tijd en input!
Albert Duisenberg en Rob Hooft: bedankt voor het kritische doorlezen van het proefschrift.
Verder zou ik ook graag de Bijvoet-mede-aio's bedanken voor de leuke aio-avonden, die met een borrel of etentje werden gehouden.

Vrienden en familie zorgen voor de nodige afwisseling en steun. Carmen, ook al zit je in Spanje, het contact is toch zeker niet verwaterd. Het is altijd leuk om met jou diepgaande gesprekken te hebben, ook over de telefoon! Tot de volgende keer in Nederland of Spanje! Nina, onze theekransjes zijn altijd gezellig. Hopelijk kunnen we deze traditie nog lang voortzetten. Louise, so great to see you soon in the Netherlands! Marjolijn en Suzanne, we hebben samen de scheikundestudie gedaan en ook bijna tegelijk het aio-schap. Onze sleep-overs zijn erg gezellig, ik heb ook al zin in jouw promotie, Marjolijn! Ciska, Carmi, Gaby en Maartje: tijdens onze gezellige etentjes wordt er veel gelachen en lachen is gezond ;) Maartje, leuk dat je als paranimf me ter zijde zult staan. Je vertrouwen in mij heeft me al die jaren veel geholpen. Meine lieben Ursulinnen: Hanh, Silke, Brigita und Angelina. Wir sind schon seit Ewigkeiten befreundet, auch wenn wir alle in anderen Länder wohnen…gute Freunde sind im Geiste immer vereint! Angel, schön dass du meine Paranimfe bist und mir den Rücken stärken wirst:) 爸爸, 妈妈, 谢谢你们给我的爱和鼓励, 还有我的小Minka, 你是我们家的小太阳.
Ruben, van alle cadeaus in mijn leven, ben ik het meest dankbaar voor jou! Dank voor alles.