Error Estimation and Pattern Recognition Techniques in Protein Crystallography
Zwart, P.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 3

The influence of positional errors on the Debye effects

Abstract

The relation between a Gaussian perturbation of the atomic positional parameters and the average squared structure factor amplitude is presented. Using an error dependent radial distance distribution of an atomic protein model, it can be shown that the Debye effects diminish exponentially as a function of increasing positional errors. These relations can be used to estimate the quality of an atomic model and the corresponding phases. The limiting case of equal atoms with an infinitely large coordinate error results in the classical Wilson model.

3.1 Introduction

The deviations from a straight line in a Wilson plot (Wilson, 1942, 1949) are known as Debye effects (Giacovazzo, 1998). These deviations are mainly caused by the stereo-chemistry of the molecular structure and can be modelled by the Debye equation (Debye, 1915):

\[
E[|F_h|^2]_o = \sum_j \sum_k f_j(h)f_k(h) \frac{\sin(2\pi d_{jk}h)}{2\pi d_{jk}h}
\]  

(3.1)

The subscript \(o\) in expression (3.1) denotes that the averaging is carried out over all orientations of \(F_h\) for a given reciprocal lattice spacing \(h\). Note that in expression (3.1) we do not account for lattice periodicity or other packing effects and the expression is thus only strictly valid for a single unit cell or molecule. The effect of the shape of the
molecular envelope on the Wilson plot is not directly accounted for. This point is discussed by Morris et al. (2003a). The need for these assumptions will however be eliminated as will be discussed below.

Expression (3.1) can be rewritten in terms of the radial distance distribution \( f_{\text{rad}}(d) \). Assuming equal atoms, one arrives at:

\[
\mathbb{E}[|F_h|^2]_o = N f^2(h) \left( 1 + (N - 1) \int_0^\infty f_{\text{rad}}(d) \frac{\sin(2\pi dh)}{2\pi dh} dd \right)
\]  

(3.2)

For the trigonometric part of the structure factor amplitudes, expression (3.2) can be transformed into

\[
\mathbb{E}[|E_h|^2]_o = 1 + \gamma(h)
\]

(3.3)

with

\[
\gamma(h) = (N - 1) \int_0^\infty f_{\text{rad}}(d) \frac{\sin(2\pi dh)}{2\pi dh} dd
\]

(3.4)

In the Wilson approximation the atoms are independent and uniformly distributed throughout the unit cell, resulting in \( \gamma(h) = 0 \) (Giacovazzo, 1998). The excess or lack of specific interatomic distances results in a non-zero interference and affects the mean squared structure factor amplitude. This is demonstrated in Figure 3.1, where the radial distance distribution of a typical protein is multiplied by a \( \text{sinc}(2\pi hd) = \sin(2\pi hd)/(2\pi hd) \) term (see expression (3.4)), for \( 1/h \) equal to 1.1, 2.2 and 4.5 Å. In the same plots the curves for a structure with a uniform, independent distribution of atoms (hereafter denoted as a random structure) are shown. The interatomic distances arising from chemically bonded atoms (1-2 distances) at about 1.4 Å and atoms involved in bond angle distances (1-3 distances) at about 2.4 Å are the two major contributors into the differences between the radial distance distribution of a protein structure and a random structure. An excess of interatomic distances compared to the random case is also found around 3.8 Å, a typical C\textsubscript{\textalpha}(i)-C\textsubscript{\textalpha}(i+1) distance. At larger distances, differences between the radial distance distribution of a protein and a random structure are also found due to secondary structure. The qualitative effects of these interatomic distances on the average squared structure factor amplitude are summarised in Table 3.1 in terms of positive or negative contributions to the integral in expression (3.4). As shown by Morris & Bricogne (2003) the 1-2 and 1-3 distances are (in part) responsible for a large peak in the mean \( |E_h|^2 \) value at around \( 1/h = 1.1 \) Å. Both the 1-2 and 1-3 distances have a positive contribution to \( \gamma(h) \) for \( 1/h = 1.1 \) Å, Figure 3.2. Also, the lack of interatomic distances between the latter two coordination shells where the sinc function is negative, results in an effectively larger value of \( \gamma(h) \) compared to a random structure. Other pronounced peaks in the average of \( |E_h|^2 \) as a function of resolution lie around \( 1/h = 2.2 \) and \( 1/h = 4.5 \) Å. At \( 1/h = 2.2 \) Å the 1-2 distances have a negative contribution, while the 1-3 distances have a positive
### 3.2. The influence of the coordinate error on the Debye effects

The error dependence of the mean squared structure factor amplitude was examined by introducing an error-dependent radial distance distribution, $f_{rad}(d|\sigma^2)$, into equation (3.2):

$$
E[|F_h|^2]_{\sigma} = N f^2(h) \left( 1 + (N - 1) \int_0^\infty f_{rad}(d|\sigma^2) \frac{\sin(2\pi h d)}{2\pi h d} dd \right)
$$

$$
= N f^2(h) \left( 1 + \gamma(h|\sigma^2) \right)
$$

(3.5)

**Table 3.1:** Qualitative contribution of specific interatomic distances to $\gamma(h)$ at pronounced extrema in a Wilson plot.

<table>
<thead>
<tr>
<th>$1/h$</th>
<th>1-2 distances</th>
<th>1-3 distances</th>
<th>C$\alpha$-C$\alpha$</th>
<th>Secondary structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>max</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+/-</td>
</tr>
<tr>
<td>min</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+/-</td>
</tr>
<tr>
<td>$1/h=1.1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$1/h=1.65$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$1/h=2.2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$1/h=2.65$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$1/h=4.5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$1/h=6.5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

max and min denote if $\gamma(h)$ is at a local maximum of minimum at the given value of $h$. + denotes a positive contribution to $\gamma(h)$, - denotes a negative contribution. +/- denotes both a positive and negative contribution to $\gamma(h)$. See main text and Figure 3.2 for further details.

The influence of a coordinate error on the Debye effects can therefore be used to assess the quality of an atomic model and the corresponding phases. The function $\gamma(h)$ is expected to be essentially the same for a wide range of protein structures as judged from the well known features of the Wilson plot (Popov & Bourenkov, 2003; Cowtan & Main, 1998; Blessing et al., 1996), although a dependence on the secondary structure is present (Morris et al., 2003a). An empirical $\gamma(h)$ curve obtained from a PDB analysis (Bernstein et al., 1977; Berman et al., 2000) can be used to estimate $\gamma(h)_{obs}$ from the observed structure factor amplitudes of the protein under consideration, which can be subsequently used to assess the model and phase quality.
3.2. The influence of the coordinate error on the Debye effects

Figure 3.1: Sine functions multiplied with the radial distance distribution of a random structure (thin line) and lysozyme (thick line, PDB id: 102L), for (top to bottom) \( h=1/1.1 \), \( 1/2.2 \) and \( 1/4.5 \) Å. See Table 3.1, Figure 3.2 and the main text for further details.

Figure 3.2: Integrals of root delimited line sections of the curves depicted in Figure 3.1 for \( 1/h = 1.1 \) Å. The values on the horizontal axis correspond to the midpoints of the root delimited line sections. The height of the bars are proportional to the numerical value of the integral. The difference between the integrals from the protein and random structure defines the qualitative contributions listed in Table 3.1.
where $\gamma(h|\sigma^2)$ is the error-dependent variant of expression (3.4). The expectation value of the trigonometric part of the structure factor amplitude becomes:

$$
E[|E_h|^2_{\omega,q}] = 1 + \gamma(h|\sigma^2)
$$

(3.6)

The subscript $q$ in expressions (3.5) and (3.6) denotes that the expectation value is obtained by integrating over the errors of the positional parameters. The errors of the positional parameters are assumed to be distributed independently for each atom, according to a spherically symmetric Gaussian with variance terms equal to $\sigma_m^2$ along the $x$, $y$ and $z$ directions. The change in the radial distance distribution can then be written in accordance with Zwart & Lamzin (2003a):

$$
\int_0^\infty f_{rad}(d|\sigma^2)NCM(d|d^\text{tar},\sigma^2)dd^\text{tar}
$$

where

$$
NCM(d|d^\text{tar},\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(d-d^\text{tar})^2}{2\sigma^2}}
$$

(3.8)

and $\sigma^2$ is the sum of the variances of the error terms of the positional parameters of a pair of atoms:

$$
\sigma^2 = \sigma_j^2 + \sigma_k^2
$$

(3.9)

For errors with a variance equal to $\sigma_m^2$ for all atoms, $\sigma^2$ becomes equal to $2\sigma_m^2$. When atom $j$ is not perturbed and an atom $k$ has a positional error with variance $\sigma_m^2$, $\sigma^2$ becomes equal to $\sigma_m^2$.

It can be shown (Appendix II-1) that:

$$
\gamma(h|\sigma^2) = \exp(-2\pi^2h^2\sigma^2)\gamma(h)
$$

(3.10)

The exponential multiplier has the same form as the $D$ term in the work of Luzzati (1952) and Read (1986). In the limiting case of an infinitely large error and $h \neq 0$, $\gamma(h|\sigma^2)$ in expression (3.10) becomes zero, effectively resulting in the Wilson approximation of independent, uniformly distributed atoms and thus $E[|F_h|^2] = 1$. The dependence of $E[|F_h|^2]$ as a function of the coordinate error and the resolution can be used to estimate $\sigma_m^2$ and corresponding figures of merit. Let us model the expected, average calculated intensity as a function of resolution and model error as follows:

$$
E[|F_h|^2]_{\omega,q} = k_p \exp[-B_{Wh}h^2/2] \sum_{j}^{N} f_j^2(h)(1 + \exp[-4\pi^2\sigma_m^2h^2]\gamma(h))
$$

(3.11)

When $\gamma(h)$ is known, or a good estimate of it is available, denoted as $\gamma(h)_{\text{obs}}$, a least-squares minimisation of the difference between the average calculated, squared structure
factor amplitude as a function of resolution ($<|F_h|^2>$ obtained from a model) and its expected value ($E[|F_h|^2]$) can be carried out. The latter expectation value is calculated using expression (3.11) thus allowing to estimate the scale factor $k_p$, the Wilson plot B value $B_{wi}$ and the variance of the error model $\sigma_m^2$. This variance can in turn be used to estimate phase probabilities (Sim, 1958, 1959) and corresponding figures of merit, which are defined as the expected value of the cosine of the phase difference between the available and error-free phases. $\gamma(h)_{\text{obs}}$ (an estimate of $\gamma(h)$ of the protein under consideration) can be obtained from the observed X-ray data and from an empirically obtained, standard $\gamma(h)$ curve, denoted as $\gamma(h)_{PDB}$, using a procedure outlined in Appendix II-2.

The standard $\gamma(h)_{PDB}$ curve was obtained from the analysis of 100 protein structures from the PDB. The structure factor amplitudes were calculated and have subsequently been normalised in resolution bins:

$$<|E_h|^2>_o = \frac{\sum_h |F_h|^2}{N_h \sum_j f_j^2(h)}$$

(3.12)

The subscript $h$ in expression (3.12) denotes the summation over the $N_h$ reflections that fall within the resolution bin $h$. Although this is more computationally intensive than obtaining the $\gamma(h)$ profiles via the radial distance distribution and the Debye equation, expression (3.2), it has the advantage that lattice periodicity and non-equal atom effects are incorporated. The resulting mean $\gamma(h)$ profile is shown in Figure 3.3 together with a $\gamma(h)$ profile obtained using expression (3.4) and a $\gamma(h)$ profile obtained using (3.12) for equal atom structures for comparison. As shown in Appendix II-2, the use of the empirically obtained $\gamma(h)_{PDB}$ curve for the estimation of $\gamma(h)_{\text{obs}}$ avoids the need for the single, equal atom molecule approximation.

3.3 Results

3.3.1 Coordinate error dependent $\gamma(h)$ profiles

Monte Carlo simulation has been carried out to compute the average value of $<\frac{\sin[2\pi hd]}{2\pi hd}>$, with $d$ distributed according to the non-central Maxwell distribution in order to test the validity of expression (3.10). This has been carried out for a number errors and various values of $h$. The numerical results have been subsequently compared with the results from expression (3.10). A plot of the average values $<\frac{\sin[2\pi hd]}{2\pi hd}>$ as a function of the expected values is shown in Figure 3.4. To visualise the effect of the reduction of the Debye effects with increasing positional error, the atomic model of lysozyme (PDB id: 102L) has been used to compute a number of error dependent radial distance distributions and corresponding $\gamma(h)\sigma^2$ profiles via (3.5) and (3.7). The resulting profiles are shown in Figure 3.5. The contribution of hydrogen atoms has been omitted.
3.3. Results

Figure 3.3: Empirical $\gamma(h)$ curves determined from a selection of good quality atomic protein models using expression (3.12) for the deposited protein models (crystal structure), the same protein models but with all atoms as oxygen and B values set to 20 Å$^2$ (crystal structure, equal atoms) and using the radial distance distribution, expression (3.5), (single molecule, equal atoms). The differences between the curves are ascribed to packing effects and the assumption of equal atoms in expression (3.5).

Figure 3.4: The expected value of $\gamma(h|\sigma^2)$, given by relation (3.10), is plotted against the average value of $<\frac{\sin(2\pi h d)}{2\pi h d}>$ for a distance distributed according to the non-central Maxwell distribution with a target distance equal to 2.5 Å at various values of $\sigma^2$ and $h$, black diamonds. The least-squares line fitted through the points has a slope of 1 and correlation coefficient of 1.
3.3.2 Model and phase quality estimates

Estimates of $\sigma^2_m$ and corresponding figures of merit have been obtained using the described least-squares procedure, with a number of different errors on the model.

The first model used was the ARP/wARP (Perrakis et al., 1999) distributed example (with the X-ray data) of Leishmanolysin (PSP; courtesy of P. Metcalf). The final model has been randomised by adding a Gaussian error to the positional parameters with an rmsd of 1.5 Å. Structure factor amplitudes have been calculated from this model using REFMAC5 (Murshudov et al., 1997) and have been used to estimate $\sigma^2_m$. The overall scale factor, Wilson plot B value, the bulk solvent parameters and $\gamma(h)_{obs}$ have been estimated from the measured, experimental data (a zero coordinate error was assumed) as outlined in Appendix II-2. $\gamma(h)_{obs}$ has been subsequently used to estimate the scale factor, the Wilson plot B value and the variance of the Gaussian error model, $\sigma^2_m$, given the structure factor amplitudes calculated from the randomised model. In Figure 3.6 a bulk solvent corrected Wilson plot from the experimental data is shown (Observed I) as well as a fitted curve on the basis of $\gamma(h)_{obs}$ (Theoretical I, 0 Å rmsd). The least squares residual estimated the coordinate error to be 1.7 Å. The corresponding Wilson plots using the calculated structure factor amplitudes of the randomised model are shown in Figure 3.6 as well.
3.3. Results

Figure 3.6: Wilson plots for experimental, bulk solvent corrected PSP structure factor amplitudes (Observed I) and the fit using the estimated $\gamma(h)_{obs}$ (Theoretical I, 0 Å rmsd). Similar curves are shown for the structure factor amplitudes calculated from the model with an rmsd of 1.5 Å on the positional parameters (Calculated I). A fit of the Wilson plot of the calculated structure factor amplitudes using $\gamma(h)_{obs}$ and assuming a coordinate error of 1.7 Å is also shown (Calculated I, 1.7 Å rmsd). In the upper right corner, the exponentiated negative least squares criteria is shown versus the rmsd.

The same final model of PSP was scrambled by introduction of a Gaussian error to the atomic parameters and underwent five unrestrained refinement cycles with REFMAC5 using the full resolution range of the observed structure factor amplitudes, with the use of cross validation. The resulting coordinates have been used to calculate model structure factor amplitudes and figures of merit were estimated. This has been carried out for a number of different errors as well as for a phase set calculated from a model with an rmsd of 3 Å which was used for free atom modelling with ARP/wARP, without the use of cross validation. A similar set of tests has been carried out on the 1HF8 model and X-ray data set from the PDB. For 1HF8, the data was truncated at the low resolution side because of poor data quality for $1/h > 7$ Å. The results of these test are summarised in Figures 3.7 and 3.8. Another test has been carried out on a number of intermediate free atom modelling structures. Solvent-flattened experimental phases of PSP were used to carry out a free atom modelling experiment without the use of cross validation. The estimated figures of merit and rmsd values are shown in Figure 3.9.
Figure 3.7: Estimated figures of merit for the PSP data set (resolution: 20–2.0 Å, B_{WH} is 18 Å²) for various scrambled models refined without restraints, as well as from a free atom model obtained from phases generated from the final model randomised by 3 Å rmsd. In the latter case, no cross validation has been used. <estimated fom> denotes the average figure of merit estimated via the described method. For comparison, the REFMAC5 estimate is given as well. <true fom> is defined as the average value of the cosine of the phase differences between the final and scrambled model.

Figure 3.8: Estimated figures of merit for the 1HF8 data set (resolution: 7–2.0 Å, B_{WH} is 32 Å²) for various unrestrained refined scrambled models. The true figure of merit is taken to be equal to cosine of the phase difference of the final and current model. <estimated fom> denotes the average figure of merit estimated via the described method. For comparison, the REFMAC5 estimate is given as well.
3.4 Discussion and Conclusions

As seen from Figure 3.4, the exponential multiplier describes the changes in the Debye effects as a function of coordinate error rather well and offers an easy way of modelling these effects. The differences in the $\gamma(h)_{PDB}$ profiles computed using expression (3.4) and via binning of structure factor amplitudes are ascribed to the underlying assumptions. The effect of packing only affects the low resolution part of the $\gamma(h)$ curve, whereas the effect of an equal atom assumption is introduces substantial differences over the whole resolution range. The curves can however be scaled together using an exponential model similar to the Babinet terms used to model the effects of the bulk solvent.

The estimates of the figures of merit shown in Figures 3.7, 3.8 and 3.9 are close to the REFMAC5 estimates, indicating that the effects studied contain enough information to predict, within certain limits, the accuracy of the model and corresponding phases. In the case of phases originating from a model with an rmsd of 3 Å and extreme model bias due to the subsequent free atom modelling without cross validation (see Figure 3.7, entry rmsd=3 Å + f.a.m) the average estimated figure of merit is still larger than the true average cosine of the phase error but is a better estimate than the one obtained by REFMAC5.

A key point is that the presented error estimation method is rather sensitive to the quality of the low resolution part of the data set used. This is ascribed to the fact that the Debye effects at high resolution diminish faster than those at lower resolution. Most information on the value of $\sigma^2$ when the error is (moderately) large is thus obtained from the low
resolution part of the data. More appropriate weighting schemes in the least-squares procedure or a maximum likelihood approach can possibly account for this sensitivity and might be useful in reducing the observed bias in the estimates. Linked to this is the need for a model describing the behaviour of the bulk solvent and its effect on the average structure factor amplitude. The exponential model (Tronrud, 1997) used here is known for its limitations (Fokin & Urzhumtsev, 2002) but has been widely used because of its simplicity. The presented method is sensitive to non-randomly incomplete data, such as missing strong, low angle reflections. These effects can in principle be modelled by using the characteristics of the truncated Wilson distribution (Parthasarathy & Sekar, 1993b), rather than ignoring a subset of valuable structure factor amplitudes as done for the 1HF8 data set. Furthermore, the method is based on an assumption of independent Gaussian errors on each atom. Violation of this assumption undermines the basic principles of the method, which is largely designed for usage during free atom modelling experiments.

Ideally, the dependence of the average squared structure factor amplitude as a function of resolution and model error, should be used in conjunction with other error estimation methods, such as $\sigma_A$ (Read, 1986), in a hope to enhance the overall quality of the error estimates. Further incorporation of prior knowledge in the error estimation method outlined in this paper might enhance the results. If secondary structural information is available or if the classification procedure outlined by (Morris et al., 2003a) proves to be reliable and robust, then protein specific $\gamma(h)_{PDB}$ curves can be utilised to obtain a more accurate $\gamma(h)_{obs}$ estimates.

Using the expected averaged squared structure factor amplitude as a function of resolution and model error as a source of information during refinement, seems to be an interesting option. This is in effect a reciprocal space variant of the addition of restraints to atoms on the basis of known radial distance distributions in proteins (Sheldrick, 1998; Scheres & Gros, 2001, 2003). A more radical and possibly better approach would be to improve the description of structure factor probability distributions by taking into account stereochemistry a priori, as suggested by Bricogne (1997b,a).