SOME PROPERTIES OF CRYSTALLOGRAPHIC RELIABILITY INDEX - R_{factor}: EFFECTO OF TWINNING

GARIB N. MURSHUDOV

Abstract. Macromolecular crystallography (MX) is one of the popular technique available to the structural biology community that allows elucidation of atomic and sometimes electronic details of biologically important molecules. One of the important problems of the field is the overall atomic model quality estimation. In this paper some of the probability distributions used in MX is reviewed and generalised for one particular crystal growth phenomenon - twinning. It is shown that in the limiting case of perfect twinning the distributions are related to non-central χ^2 distribution. Analysis of the most popular reliability index - R_{factor} convincingly illustrates that it is a poor quality factor and if care is not exercised then the existing techniques may give misleading results and create an illusion of high model quality when it is not the case. It is concluded that designing an objective and preferably computationally efficient reliability index for the comparisons of the quality of atomic models is urgently needed.

1. Introduction

Macromolecular crystallography (MX) is one of the most popular experimental techniques available to the structural biology community that is able to give atomic and electronic details of biologically important molecules with high accuracy. According to the Protein Data Bank (PDB) [2] around 85% of more than 68000 available three dimensional structures have been analysed using this technique (www.pdb.org/pdb/). Other widely used experimental techniques used for the derivation of three dimensional structures of macromolecules are Nuclear Magnetic Resonance and Electron Microscopy.

The PDB is a valuable resource of the structural biology that have wide range of applications including structure based drug design, protein folding problem, study of protein mechanisms and signal transduction in organisms. The PDB is also recycled and used as a starting model for new MX structure analysis [11]. For all these applications it is important to be able to select high quality models. To select best quality models among similar ones it is necessary to have objective reliability indicators that work sufficiently well for all models independent of the data they were derived from. There are already many software tools to analyse quality of

Current address: YSBL,University of York, York, UK YO10 5EF
New address: Structural Studies Division, LMB-MRC, Cambridge, CB2 0QH, UK
email: garib@ysbl.york.ac.uk or garib@mrc-lmb.cam.ac.uk

the derived models and validate them [22, 7, 6]. These software tools either rely only on chemical and structural properties of the atomic models ignoring the fit to the experimental data or use available reliability index that is, as it will be shown in this paper, a poor quality indicator, especially when two different models from two different crystals are compared.

One of the most widely used reliability index is so-called R_{factor}:

$$R_{factor} = \frac{\sum_H ||F_{H,o}| - |F_{H,c}||}{\sum_H |F_{H,o}|} \tag{1}$$

Where $|F_{H,o}|$-s are observed data - amplitudes of the structure factors or Fourier coefficients of crystals and $|F_{H,c}|$-s are those calculated from the atomic model. The summation is carried over all data points included in the model derivation.

Behaviour of this index depending on the model quality has been the subject of studies for some time [12, 19]. However these studies included only very simple cases - single crystal without any complications. It is well known that in many cases data collected as a result of X-ray experiment are from more than one crystals [16, 23, 10] resulting in completely different statistical properties of these datasets. Since statistical properties of datasets from single and multiple crystals are different it can be expected that behaviour of any property will be different, R_{factor} is not an exception. Therefore model comparison based on these indicators will give misleading results.

It should be noted that all available MX structure optimisation [13, 1, 18, 5, 3] - refinement software produce R_{factor} as a reliability index and it is used to compare of quality of different models derived using different experimental conditions. Therefore it is important to anlalyse the behaviour of R_{factor} under different conditions. This paper describes its behaviour under two conditions, data from single and twinned crystals, and demonstrates that in general R_{factor} is a poor quality indicator.

Macromolecular crystallographers usually use so-called R_{free} to validate derived models. This value is calculated using a portion of the observed data that are left from remaining part of calculations, thus reducing influence of overfitting. The results described in this contribution are equally valid for R_{factor} and R_{free}.

Organisation of the paper: In the first section a brief review of relationship between crystallographic calculations and Fourier series is given. In the second section structure factor probability distributions under two simplified conditions are derived. In the third section behaviour of R_{factor} under different conditions are derived and analysed. The final section describes the shortcoming of the current analysis and gives the list of the problems to be solved to design new quality indicators that can be used for objective comparison of various Macromolecular structures.

2. Crystallographic calculations and Fourier series

MX experiment deals with crystals and observations from crystals are related to Fourier coefficients of these crystals. By definition crystals are three-dimensional periodic entities and therefore their Fourier transformation is expressed as Fourier series. Maximum non-repeating part of the crystals is called asymmetric unit. The minimal paralelepipedic part of the crystal that when repeated by translations in three directions fills the whole space is called unit cell of the crystal.

Let us assume that we have a crystal and it contains N atoms with positions x_i $i = \overline{1, N}$. Then electron density in the crystal can be represented as a sum of the atomic electron densities:

$$\rho(x) = \sum_{i=1}^{N} \rho_i(x - x_i) \tag{2}$$

Where $\rho_i(x)$ is the electron density for the i-th atom. Usually atomic electron densities are approximated using the sum of Gaussian functions. It allows development of computationally efficient algorithms for crystallographic calculations, however the results in this paper do not depend on such approximations.

The Fourier coefficients of the electron density in the crystal can be expressed as:

$$F_H = \mathcal{F}(\rho(x)) = \sum_{i=1}^{N} \mathcal{F}(\rho_i(x - x_i)) = \sum_{i=1}^{n} f_i e^{2\pi\iota H^T x_i} \tag{3}$$

Where f_i is the Fourier transformation of $\rho_i(x)$ - formfactor of the i-th atom.

As it can be expected in general F_H is a complex number. However if a crystal is centrosymmetric, i.e. for each atom with the position x_i there is an atom with the position $-x_i$ then these coefficients will be real numbers, since in this case the total electron density will be centrosymmetric the Fourier coefficients of a centrosymmetric function will be real numbers. Macromolecules by their nature do not have centre of symmetry, therefore their crystals are in general non-centrosymmetric and therefore we consider only non-centrosymmetric cases. It should be noted that since in general crystals have some rotational symmetry some of the Fourier coefficients in general will behave like those from centrosymmetric crystals. There is no difficulty of extending the analyses carried out in this paper to the centrosymmetric cases, however it would not change the qualitative conclusions of the paper.

In general crystals obey one of the possible 230 space group symmetries and that put certain conditions on Fourier coefficients. For full review of crystallography and crystal symmetry see Vainshtein [21], International Table for crystallography [8]. For simplicity of the derivations, interpretations and to avoid unnecessary complications of notations we only consider crystals without any rotational symmetry

- a so-called P_1 space group. The results with minor modifications are applicable for all crystals.

We will need also expressions for real and imaginary parts of the structure factors:

$$F_H = A_H + \iota B_H = \sum_{i=1}^{n} f_i \cos 2\pi H^T x_i + \iota \sum_{i=1}^{n} f_i \sin 2\pi H^T x_i \tag{4}$$

It is clear that bi-scalar product - $H^T x_i$ must be invariant under coordinate transformation. Thus coordinate system in the Fourier - reciprocal space depends on the coordinate system in the space where atoms live - real space. When x_i are in the crystal coordinate system where $0 \leq x_i < 1$ then the corresponding coordinates in the reciprocal space are usually integers and they are called Miller indices. Miller indices are denoted by H. When atomic coordinates are in orthogonal system with atomic units then the corresponding coordinates in the reciprocal space are denoted by s . $1/|s_{max}| = 1/\max(|s|)$ for the data set is called the resolution of the diffraction data. Usually data set is collected in a spherical shell with between $s_{max} = \max(|s|)$ and $s_{min} = \min(|s|)$. $d_{max} = 1/s_{max}$ and $d_{min} = 1/s_{min}$ are called resolution limits of the data. For the indices of the structure factors or when working with Fourier series it is convenient to use Miller indices, however when atomic positions and errors in them are considered it is convenient to use orthogonal system. We will use H and s notations interchangeably and coordinate systems used will be clear from the context.

The result of crystal diffraction experiment from a single crystal is related to the amplitude of the structure factors:

$$I_{H,o} \propto |F_H|^2 = |A_H|^2 + |B_H|^2 \tag{5}$$

In general the observed data should be considered as random variables with some probability distributions. Usually the probability distribution of observed intensities are approximated using normal distribution:

$$P(I_{H,o}; I_{H,true}) = \frac{1}{\sigma_{H,o}\sqrt{2\pi}} e^{-\frac{(I_{H,o}-I_{H,true})^2}{2\sigma^2_{H,o}}} \tag{6}$$

Where $I_{H,o}$ is observed, $I_{H,true}$ is true intensities of the structure factors, σ_o is standard deviation. For simplicity we will assume that observed structure factors are equal to the true structure factors. Essentially it means that $\sigma_{H,o} \to 0$ and consequently $P(I_{H,o}; I_{H,true}) = \delta(I_{H,o} - I_{H,true})$, with $\delta(x)$ denoting Dirac's δ-function. Although it is a gross simplification it does not affect the qualitative conclusions drawn in this paper. In principle the effects of the experimental errors can be accounted for by adding a simple integration over experimental errors. In the following sections we will drop subscripts o and *true* assume that we are dealing with observed and/or "true" structure factors. We will also drop subscript H since we usually will be dealing with either single index or pair of indices.

3. Some structure factor probability distributions

Let us assume that we have two crystals with exactly same unit cell dimensions and both crystals contain exactly same number of atoms and there is one to one correspondence between atoms in these crystals. Let us further assume that positions of the atoms in the second crystal can be expressed $y_i = x_i + \Delta x_i$. In addition let us also assume that atoms in each crystal are independent of each other and uniformly distributed over the asymmetric unit of the crystal, Δx_i are independent of each other and of x_i. Let us also assume that all Δx_i have exactly same distributions. Since the number of atoms is large and under these assumptions conditions of the central limit theorem are fulfilled (*e.g.* the first, second and third moments are finite) we can use four dimensional normal distribution as an approximation to the joint probability distribution of the structure factors of the first and second crystals. For this we need to derive the first and the second moments of the imaginary and real parts of the structure factors of these crystals. Let us use the following notations: subscripts 1 and 2 are crystal numbers, $F_1 = (A_1, B_1)$ and $F_2 = (A_2, B_2)$. Then for mean values we can write.

$$< A_1 >=< B_1 >=< A_2 >=< B_2 >= 0 \tag{7}$$

Here we used the fact that under above assumptions

$$< \cos 2\pi H^T x_i >=< \sin 2\pi H^T x_i >= \frac{1}{|V|} \int_V \cos 2\pi H^T x_i dx_i = 0$$

for all i-s. Where V is the unit cell of the crystal, $|V|$ is the volume of the unit cell, $< . >$ denotes expectation or mean value.

$$\begin{aligned} < A_1^2 >=< B_1^2 >=< A_1^2 >=< B_1^2 >= \frac{1}{2} \sum f_j^2 \\ < A_1 A_2 >=< B_1 B_2 >= \frac{1}{2} \sum f_i^2 < \cos 2\pi H^T \Delta x_i > \\ < A_1 B_1 >=< A_2 B_2 >=< A_1 B_2 >= 0 \end{aligned} \tag{8}$$

Here we additionally used the fact that under above assumptions

$$\begin{aligned} < \cos^2 2\pi H^T x_i >=< \sin^2 2\pi H^T x_i >= \frac{1}{2} \\ < \cos 2\pi H^T x_i \cos 2\pi H^T (x_i + \Delta x_i) >=< \sin 2\pi H^T x_i \sin 2\pi H^T (x_i + \Delta x_i) >= \\ \frac{1}{2} < \cos 2\pi H^T \Delta x_i > \end{aligned}$$

Since all Δx_i-s have the same distribution we can write $D(s) =< \cos 2\pi H^T \Delta x_i >$. If we use notations $\Sigma = \sum f_i^2$ then for the full covariance matrix we can write:

$$\mathbf{\Sigma_4} = \frac{\Sigma}{2} \begin{Bmatrix} 1 & 0 & D & 0 \\ 0 & 1 & 0 & D \\ D & 0 & 1 & 0 \\ 0 & D & 0 & 1 \end{Bmatrix} \tag{9}$$

Thus the joint probability distribution of structure factors of two crystals can be approximated by a four dimensional normal distribution:

$$P(A_1, B_1, A_2, B_2) = N_4(\mathbf{0}_4, \mathbf{\Sigma_4}) \tag{10}$$

where $\mathbf{0}_4 = (0, 0, 0, 0)$ is a four dimensional vector of zeros.

In the following discussions we will assume that $\Sigma = 1$. It can be achieved by dividing all A, B-s by the corresponding $\sqrt{\Sigma}$ and work with so-called normalised structure factors. This simple trick does not change the results of this paper. When we need to derive distributions of the structure factors then we simply can use denormalisation which is just a simple change of variables.

This the multivariate normal distribution forms the basis of the probability distributions derived and used in crystallography. For example from this distribution one can deduce that the marginal distribution of the structure factors of a single crystal is two dimensional normal distribution with zero mean and identity variance matrix:

$$P(A_1, B_1) = \int_{R^2} P(A_1, B_1, A_2, B_2) dA_2 dB_2 = N(\mathbf{0}_2, \mathbf{I}_2/2) \tag{11}$$

where $\mathbf{0}_2 = (0, 0)$ is a two dimensional zero vector and $\mathbf{I}_2$ is two dimensional identity matrix.

The conditional distribution of the structure factors of one crystal given another one is a two dimensional normal distribution with mean equal to (DA_2, DB_2) and variance matrix equal to $\mathbf{I_2}/(2(1 - D^2))$:

$$P(A_1, B_1; A_2, B_2) = \frac{P(A_1, B_1, A_2, B_2)}{P(A_2, B_2)} = N\left((DA_2, DB_2), \frac{1}{2(1 - D^2)}\mathbf{I_2}\right) \tag{12}$$

Two extreme cases of D are of special interest: $D = 0$ and $D = 1$. In the first case since the conditional distributionn of the first crystal given the second crystal is independent of the structure factors of the second crystal therefore two crystals are independent. This case is usually called unrelated crystals. In other words there is no information in the second crystal about the first crystal and *vice versa*. In the second case the conditional distribution can be expressed as a product of Dirac's δ function.

$$P(A_1, B_1 | A_2, B_2) = \delta(A_1 - A_2)\delta(B_1 - B_2)$$

That is the case when two crystals are identical, *i.e.* knowing one of the crystals is sufficient to describe everything in the second crystal.

Using the distributions real and imaginary parts of the Fourier coefficients we can derive the distribution of amplitudes of structure factors. Since $|F_1|^2 = |A_1|^2 + |B_1|^2$ we can use a well known result from the Statistics [20] that if random variables z_l, $l = \overline{1,k}$ have normal distributions with mean and variance equal to μ_l and σ_l^2 then the distribution of $\sum_{l=1}^{k}(z_l/\sigma_l)^2$ is non-central χ^2 distribution [1]with k degrees of freedom and non-centrality parameter equal to $\sum_{l=1}^{k}(\mu_l/\sigma_l)^2$. Thus $\frac{2|F_1|^2}{1-D^2}$ has the non-central χ^2 distribution with the degrees of freedom equal to two and non-centrality parameter equal to $\frac{2D^2|F_2|^2}{1-D^2}$. And finally we can write:

$$P(|F_1|^2;|F_2|^2) = \frac{2}{(1-D^2)}\chi_2^2(\frac{2|F_1|^2}{1-D^2}, \frac{2D^2|F_2|^2}{1-D^2}) \tag{13}$$

This equation with some modification is used by many MX software that is based on MX Maximum likelihood modelling of the data from X-ray diffraction experiment [13, 1, 5, 3]. In crystallography it is known as Rice distribution [17, 4, 15]. More precisely the conditional distribution of distribution of $|F_1|$ given $|F_2|$ is Rice distribution and that is what is used in MX structure analysis. However it easily can be derived from the general theory of χ^2 distributions.

3.1. **Expression for D.** In general differences - Δx_i between two crystals can have any distribution. However for simple illustrative purposes we can assume that their distribution is three dimensional normal distribution with zero mean and diagonal covariance matrix:

$$P(\Delta x) = N(\mathbf{0}_3, \sigma_x^2\mathbf{I}_3) \tag{14}$$

where σ_x is the standard deviation of the differences.

Using this assumption we can write for D:

$$D = \int_{R^3} \cos(2\pi\Delta x s)P(\Delta x)d\Delta x = e^{-2\pi^2\sigma_x^2|s|^2} \tag{15}$$

This expression was used by Luzzati [12] to study the behaviour of R_{factor}-s in single crystal cases. Again in the limiting cases when 1) $\sigma_x \to \infty$ then two crystals are unreleated and when 2) $\sigma_x \to 0$ then two crystals are identical.

3.2. **Twinning in Crystallography.** It is often the case that two or more orientations of a a crystal are indistinguishable. These cases usually happen when the unit cell (more precisely crystal lattice) have higher symmetry than the crystal. As a result under certain physical conditions crystals grow in several orientations simultaneously. By analysing the whole PDB Lebedev *et al* [10] showed that the occurrence of this phenomenon in macromolecular crystals is non-negligible. This phenomenon crystal growth in multiple indistinguishable directions is called merohedral twinning. When there are only two orientations of the crystal related by two

[1]Density of the probability distribution for non-central χ^2 distribution with k degrees of freedom and non-centrality parameter equal to λ: $\chi_k^2(x,\lambda) = \frac{e^{-(x+\lambda)}x^{\frac{k}{2}-1}}{2^{\frac{k}{2}}}\sum_{i=0}^{\infty}\frac{(\lambda x)^k}{2^{2k}k!\Gamma(\frac{2k+i}{2})}$

fold rotation then the phenomenon is called hemihedral twinning. For simplicity we will consider only hemihedral twinning case.

Because of the nature of the diffraction experiment the total observations from two crystals orientations are the sum of the intensities of individual crystals. Let us denote fractional occupancy of one of the crystals by $1-\alpha$, where $0 \leq \alpha \leq 0.5$. Then the occupancy of the second crystal is α. Observed intensities can be written as a weighted sum of intensities from two crystals with weights equal to $1-\alpha$ and α:

$$\begin{aligned} I_{T,1} &= (1-\alpha)I_1 + \alpha I_2 \\ I_{T,1} &= \alpha I_1 + (1-\alpha)I_2 \end{aligned} \tag{16}$$

where I_1 and I_2 are two contributing intensities from two crystal, $I_{T,1}$ and $I_{T,2}$ are two corresponding observations.

When $\alpha = 0$ then there is no twinning and it is a single crystal case. When $\alpha = 0.5$ then the twinning is called a perfect twinning. In principle when $\alpha < 0.5$ the equation 16 can be solved to find I_1 and I_2, however errors in the resulting data increases with proportionality coefficient of $1/(1-2\alpha)$. Moreover for $\alpha = 0.5$ it is imposible to solve the equation 16. Therefore it is usually advised to use the experimental data directly. For perfect twin case the probability distribution of structure factors can be derived using properties of the non-central χ^2 distributions [20]. To derive the necessary distribution we use the assumption that different structure factors (*i.e.* I_1 and I_2) from one crystal are independent of each other. Since the distributions of $\frac{2I_{11}}{1-D^2}$ and $\frac{2I_{12}}{1-D^2}$ are non-central χ^2 then $2\frac{2(I_{11}+I_{12})}{1-D^2}$ will have non-central χ^2 distribution with degrees of freedom four and non-centrality parameter $\frac{D^2(I_{21}+I_{22}}{1-D^2}$. Then for the distribution of I_{T11} we can write:

$$P(I_{T11} = \frac{4}{(1-D^2)}\chi^2_4(\frac{4I_{T11}}{1-D^2}, \frac{4D^2I_{T21}}{1-D^2}) \tag{17}$$

This distribution can be used for MX crystallographic modelling in the presence of perfect twinning. There is no difficulty of generalising it to more than two crystal cases , *i.e.* more general merohedrally twinned cases. Note that when $D = 0$ then the distribution reduces to the one that is used in crystallography for diagnosis of twinning [16].

4. Behaviour of R_{factor}-s under different conditions

R_{factor}-s given by the equation 1 are calculated between structure factors of two crystals. Data from related crystals either can be from two different crystalline entities produced during crystal structure analysis or during fitting of the atomic model into the experimental data. In the first case both data sets are observations and in the second case one of them corresponds to the observations and another one corresponds to the model under consideration. Although our purpose is to

analyse the reliability index during model building and optimisation, the results are equally applicable when data are from two different actual crystals.

In the first case actual twin fractions of the crystals may be different and in the second case α is estimated as a part of model building procedure and it is not necessary that the estimated value is equal to the actual value. Therefore we analyse R_{factor}-s between two crystals with different twin fractions.

Let us assume that we have two data sets from two related crystals. Both these crystals may have been twinned. Let us denote as the first index of the subscripts crystal number and the second index the contributor number. For example I_{11} is the first contributor from the first crystal. We can now write an equation for R_{factor} between data sets from two twinned crystals:

$$R_{factor} = \frac{\sum |\sqrt{(1-\alpha)I_{11} + \alpha I_{12}} - \sqrt{(1-\beta)I_{11} + \beta I_{12}}|}{\sum \sqrt{(1-\alpha)I_{11} + \alpha I_{12}}} \tag{18}$$

where α and β are corresponding twin fractions for two crystals. During the model building and optimisation α is the "true" and β is the estimated twin fraction.

To estimate R_{factor} using the probability distributions of the structure factors let us assume that in the reciprocal (Fourier) space in narrow spherical shells frequency of the structure factors represent their true frequency. This assumption works sufficiently well in practice. Let us additionally assume that points in the reciprocal space are sufficiently dense and the summation over H can be replaced by integration. Then for an approximation of R_{factor} we can write:

$$R_{factor} = \frac{\int_{s_{min}}^{s_{max}} \int_{R^8} |F_{T,11} - F_{T,21}| P(\mathbf{F}) |s|^2 d\mathbf{F} d|s|}{\int_{s_{min}}^{s_{max}} \int_{R^8} |F_{T,11}| P(\mathbf{F})) |s|^2 d\mathbf{F} d|s|} \tag{19}$$

where

$$F_{T,11} = \sqrt{(1-\alpha)I_{11} + \alpha I_{12}}$$
$$F_{T,21} = \sqrt{(1-\beta)I_{21} + \beta I_{22}}$$
$$\mathbf{F} = (F_{11}, F_{12}, F_{21}, F_{22}) = (A_{11}, B_{11}, A_{12}, B_{12}, B_{21}, A_{21}, B_{21}, A_{22}, B_{22})$$
$$d\mathbf{F} = dA_{11} dB_{11} dA_{12} dB_{12} dA_{21} dB_{21} dB_{21} dA_{22} dB_{22}$$
$$P(\mathbf{F}) = P(A_{11}, B_{11}, A_{21}, B_{21}) P(A_{12}, B_{12}, A_{22}, B_{22})$$

s_{min} and s_{max} are the reciprocal space vector lengths correponding to the minimum and maximum resolutions of the data used for analysis.

To estimate R_{factor} given by the equation 19 we use a numerical integration. The integration over the structure factors is performed using Monte Carlo method where sampling is performed using two four dimensional normal distributions and integration over $|s|$ is performed using a simple trapesoid rule. Calculation are done using a script written in statistical package language - R [14]. To produce figures after the integration smooth spline function [9] as implemented in the package R was used.

All analysis have been done for typical resolution range between $20\mathring{A}$ and $2\mathring{A}$. Note that the results are for illustration purpose and changing the resolution range alters results only quantitatively.

4.1. **Limiting case 1: unrelated crystals.** Recall that when crystals are unrelated then $D = 0$. For this case we consider four combination of α and β equal to 0.5 and 0. During model building and omptimisation these cases have the following interpretation: 1) $\alpha = 0.5$, $\beta = 0.5$, data are from the perfectly twinned crystal and twin is modelled; 2) $\alpha = 0.5$ $\beta = 0$, data are from the perfectly twinned crystal and twin is not modelled; 3) $\alpha = 0$, $\beta = 0.5$, data are from single crystal and perfect twin is modelled; 4) $\alpha = 0$, $\beta = 0$, data are from single crystal and twin is not modelled. Note that the case number four was analysed by Luzzati [12] and Srinavasan and Parthasaraty [19]. Their results are in perfect agreement with the results shown here.

R_{factor}-s for these four limiting cases are shown on Table 1. It is seen that if the data are not from twinned crystals and twin is modelled then R factor is lower than if twin is not modelled. It shows that in the beginning of the crystal structure analysis if twin is modelled R_{factor} may be lower thus creating an illusion that model is related to the crystal when it is not.

TABLE 1. R_{factor}-s between unrelated crystals for four different combination of limiting twin fractions. These results are resolution independent

Twin \Modelled	Yes	No
Yes	0.41	0.49
No	0.52	0.58

Figure 1 shows the behaviour of R_{factor}-s when there is no twin and twin is modelled with different twin fractions. As it can be seen the R_{factor} has minimum when $\beta = 0.5$.

This simple limiting case demonstrates shortcomings of this reliability index at the early stages of crystal structure analysis.

4.2. **Limiting case 2: identical crystals.** Another interesting case is when two crystals are identical and the only difference between them is the presence or absence of twinning. As it was noted above this case corresponds to $D = 1$ which is the same as $\sigma_x \to 0$.

Two most common cases at the end stages of structure analaysis are when 1) twin is not present and it is not modelled and 2) twin is present it is not modelled. In the first case in very unlikely scenario of $D = 1$, R_{factor} would be zero as expected. In the second case R_{factor} would be very high thus wrongly indicating that model has large errors. It is very unlikely that when atomic model error is zero and modelled twin fraction is 0.5. Optimisation programs would give much

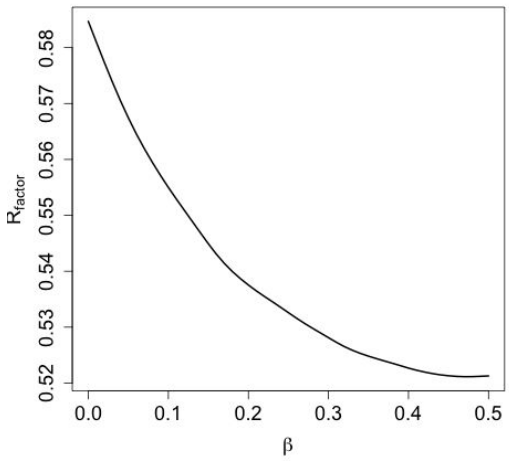

FIGURE 1. R_{factor} vs modelled twin fraction. No twinning is present, i.e. α=0 and β is estimated. As it can be seen the minimum of R_{factor} is when β is near to 0.5

TABLE 2. $R_{factors}$ between identical crystals with four different combinations of limiting twin fractions

Twin \Modelled	Yes	No
Yes	0	0.28
No	0.29	0

smaller value of β However it is still interesting to look at the behaviour of $R_{factors}$ with different β values. It can be seen from 2 that in this case R_{factor} has minimum when β is zero. We can conclude that towards the end of structure analysis twin fraction is optimised close to its true value.

4.3. **Related crystals: atomic model with error.** More interesting and practical cases are when crystals are not identical but related, *i.e.* $0 < D < 1$ and twin is present in both crystals. This case corresponds to the modelling of MX data when the model is essentially correct but has some errors.

Figure 3 shows dependence of R_{factor} on model errors σ_x for four different limiting values of α and β. From the figure it is seen that if twin is present then modelling twin could reduce R-factors by more than 10% without improving quality of the atomic model. Moreover when there are no twin and model error is large then modelling twin may give lower R_{factor} thus creating a false illusion of model improvement. For example if the standard deviation of the model error is around 0.4Å then modelling twin will improve reliability index immediately and give wrong twin fractions thus resulting in wrong model.

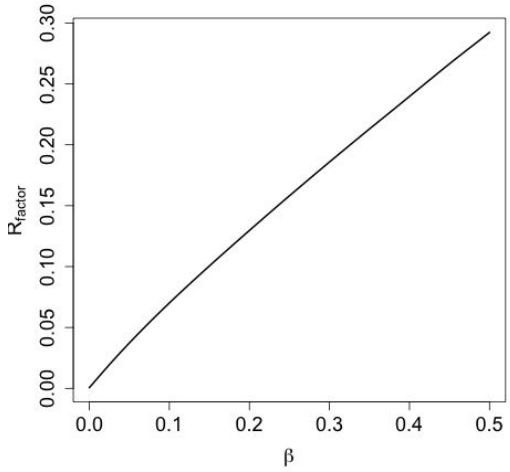

FIGURE 2. R_{factor} vs modelled twin fraction. No twinning is present, i.e. α=0 and β is estimated. In this case twin fraction would be estimated correctly to be equal to the true value - 0

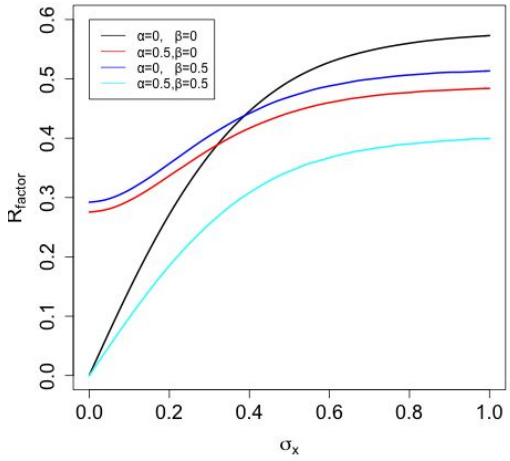

FIGURE 3. R_{factor} vs σ_x for four different limiting cases. 1) Perfect twin and twin modelled - cyan line; 2) Perfect twin with unmodelled twin - red line 3) No twin with perfect twin modelled - blue line 4) No twin and no modelled twin - black line. σ_x is in Å

As figure 4 shows if there are some errors in the atomic model then even if twin is not present then minimum of R_{factor} is when β is different form zero, *i.e.* the

actual value of α. This figure demonstrates that if the estimated twin fractions are very small they should be should be interpreted with care.

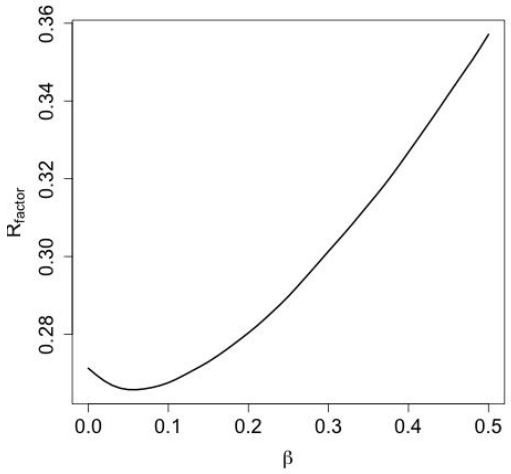

FIGURE 4. R_{factor} vs modelled twin fraction. No twinning is present, i.e. α=0, β is estimated and σ_x=0.3.

5. Conclusions and Future perspectives

In this paper we reviewed some of the basic structure factor distributions used in MX structure analysis and extended them to one particular peculiar crystal growth case - perfect twinning. It was shown that this distribution is related to the non-central χ^2 distribution. One of the future version of the MX optimisation software - refmac [13] will use a more general form of this distribution.

It was also demonstrated that the most popular reliability index - R_{factor} is very poor quality indicator and can give misleading results. It is important to search and find more robust reliability indices. The best available quality indicator is the estimated covariance matrix of the derived models. However for very large systems that is case for MX, its calculation requires prohibitively large computing resources. Moreover the results depend on the likelihood functions used which may be a challenging problem by itself for general case. Finding an objective reliability index for model comparison is an outstanding open question of MX structure analysis.

Here only very simple cases were considered. The main reason for this was that we wanted to demonstrate that even in very simple cases R_{factor} can give misleading results. Full analysis would account for combination of many different situations. For example some of the complicating circumstances that need to be accounted for in crystal structure analysis are:

1) Strictly speaking atoms are not uniformly distributed over all unit cell. Proteins are usually compact and their atoms occupy only a region (typically 50%) of the unit cell. As a result even for unrelated single crystals R_{factor}-s may be lower than 0.58 and behaviour of reliability indices may be different. Moreover by their nature atoms in molecules are bonded with each other and therefore assumption of independent atoms is a gross simplification.

2) Distribution of differences between two crystals may not be Gaussian. For example it is often the case that one part of the crystal is better defined than other parts of the crystal resulting in mixture of wildly different distributions for differences between atomic positions.

3) With or without twinning there may be other crystal growth peculiarities. A well known example is a so-called pseudo translation. It means that for every atom with position x in the crystal there is an atom with position $x + T + \Delta x$. This phenomenon has a dramatic effect on the distribution of the structure factors. For instance when T is a rational fraction of the crystal period (crystal translation) then the distribution of the structure factors will be multimodal and special treatment is needed to deal with them. Usually the effect of pseudo translation is opposite to that of twinning: $R_{factors}$ tends to become larger than that for standard single crystal case.

4) In many cases macromolecules have their internal symmetry and it may happen that this symmetry is very close to the symmetry operator causing twinning. In these cases assumption of independence of intensities with different Miller indices may not be true. For these cases usually differences between R_{factor}-s shown above are not that dramatic.

4) Usually during the model building, part of the atoms are missing and that has an effect in the second central moment - covariance matrix. The effect of missing atoms for single crystal cases was analysed by Srinavasan and Partasarathy [19].

Even with all these caveats in mind it is clear that R_{factor} is very poor model quality indicator and there is an urgent need to improve the situation. In future this problem will be addressed.

This work would be impossible without support by Wellcome Trust University Research Fellowship.

References

[1] PD Adams, PV Afonine, G Bunkóczi, VB Chen, IW Davis, N Echols, JJ Headd, L-W Hung, GJ Kapral, RW Grosse-Kunstleve, AJ McCoy, NW Moriarty, R Oeffner, RJ Read, DC Richardson, JS Richardson, TC Terwilliger, and PH Zwart. *PHENIX*: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D*, 66(2):213–221, 2010.

[2] HM Berman, T Battistuz, TN Bhat, WF Bluhm, PE Bourne, K Burkhardt, Z Feng, GL Gilliland, L Iype, S Jain, P Fagan, J Marvin, D Padilla, V Ravichandran, B Schneider, N Thanki, H Weissig, JD Westbrook, and C Zardecki. The Protein Data Bank. *Acta Crystallographica Section D*, 58(6 Part 1):899–907, 2002.

[3] E Blanc, P Roversi, C Vonrhein, C Flensburg, SM Lea, and G Bricogne. Refinement of severely incomplete structures with maximum likelihood in buster-tnt. *Acta Crystallographica Section D*, 60:2210–2221, 2004.
[4] G Bricogne. Bayesian statistical viewpoint on structure determination: Basic concepts and examples. *Methods in Enzymology*, 276:361–423, 1997.
[5] A.T. Brunger. Simulated annealing in crystallography. *Annual Reviews in Physical Chememistry*, 42:197–223, 1991.
[6] VB Chen, W. B Arendall-III, JJ Headd, DA Keedy, RM Immormino, GJ Kapral, LW Murray, JS Richardson, and DC Richardson. *MolProbity*: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D*, 66:12–21, 2010.
[7] EJ Dodson, GJ Davies, VS Lamzin, GN Murshudov, and Wilson KS. Validation tools: can they indicate the information content of macromolecular crystal structures? *Structure*, 6(6):685–690, 1998.
[8] T Hahn, editor. *International Tables for Crystallography. Volume A: Space-group symmetry.* Wiley, 2006.
[9] TJ Hastie and R J Tibshirani. *Generalized Additive Models.* Chapman and Hall, 1990.
[10] AA Lebedev, AA Vagin, and GN Murshudov. Intensity statistics in twinned crystals with examples from the PDB. *Acta Crystallographica Section D*, 62:83–95, 2006.
[11] F Long, AA Vagin, P Young, and GN Murshudov. *BALBES*: a molecular-replacement pipeline. *Acta Crystallographica Section D*, 64:125–132, 2008.
[12] V. Luzzati. Traitement statistique des erreurs dans la determination des structures cristallines. *Acta Crystallographica*, 5(6):802–810, 1952.
[13] GN Murshudov, AA Vagin, and EJ Dodson. Refinement of Macromolecular Structures by the Maximum-Likelihood Method. *Acta Crystallographica Section D*, 53(3):240–255, 1997.
[14] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2010.
[15] RJ Read. Structure-factor probabilities for related structures. *Acta Crystallographica Section A*, 46(11):900–912, 1990.
[16] D. C. Rees. The influence of twinning by merohedry on intensity statistics. *Acta Crystallographica Section A*, 36:578–581, 1980.
[17] SO Rice. Mathematical analysis of random noise. *Bell System Technical Journal*, 24:46–156, 1945.
[18] GM Sheldrick. A short history of *SHELX*. *Acta Crystallographica Section A*, 64(1):112–122, 2008.
[19] R Srinavasan and Parthasarathy. *Some Statistical Applications in X-Ray Crystallography.* Pergamon Press, 1976.
[20] A Stuart, K Ord, and S Arnold. *Kendall's Advanced Theory of Statistics, Volume 2A: Classical Inference.* Wiley, 2009.
[21] BK Vainshtein. *Modern Crystallography, Vol. 1. Fundamentals of Crystals. Symmetry, and Methods of Structural Crystallography.* Springer-Verlag, Berlin, 1995.
[22] G. Vriend. What if: A molecular modeling and drug design program. *J. Mol. Graph.*, 8(1):52–56, 1990.
[23] T. O. Yeates. Simple statistics for intensity data from twinned specimens. *Acta Crystallographica Section A*, 44:142–144, 1988.