Journal of
Applied Crystallography
ISSN 0021-8898

Received 9 August 2007
Accepted 16 November 2007

Determination of absolute structure using Bayesian statistics on Bijvoet differences

Rob W. W. Hooft,[a]* Leo H. Straver[a] and Anthony L. Spek[b]

[a]Bruker AXS, PO Box 811, 2600 AV Delft, The Netherlands, and [b]Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands. Correspondence e-mail: rob.hooft@bruker-axs.nl

A new probabilistic approach is introduced for the determination of the absolute structure of a compound which is known to be enantiopure based on Bijvoet-pair intensity differences. The new method provides relative probabilities for different models of the chiral composition of the structure. The outcome of this type of analysis can also be cast in the form of a new value, along with associated standard uncertainty, that resembles the value of the well known Flack x parameter. The standard uncertainty we obtain is often about half of the standard uncertainty in the value of the Flack x parameter. The proposed formalism is suited in particular to absolute configuration determination from diffraction data of biologically active (pharmaceutical) compounds where the strongest resonant scattering signal often comes from oxygen. It is shown that a reliable absolute configuration assignment in such cases can be made on the basis of Cu $K\alpha$ data, and in some cases even with carefully measured Mo $K\alpha$ data.

1. Introduction

Bijvoet, Peerdeman and van Bommel were the first to demonstrate that the absolute configuration of a chiral molecule could be determined by X-ray crystallography (Bijvoet *et al.*, 1951). Their method was based on the complex resonant scattering contributions to the atomic scattering factors that make the intensities of Friedel-related reflections (or their symmetry equivalents) different. This difference in intensity (the 'Bijvoet difference') depends both on the atom types present in the molecule and the wavelength of the radiation used (Flack & Shmueli, 2007). The concept of 'absolute configuration' has since been generalized to 'absolute structure' to include cases where the polarity of the structure rather than the absolute configuration is determined (Jones, 1984; Glazer & Stadnicka, 1989).

Traditionally absolute structure determination was based on analysis of Cu $K\alpha$ data collected on a diffractometer with a point detector for compounds containing atom types heavier than phosphorus. Currently, most small-molecule structure determinations are based on data collected on diffractometers equipped with CCD detectors using Mo $K\alpha$ radiation. The impact of this change is that often a more accurate, highly redundant and complete data set is obtained, which, however, often contains a weaker resonant scattering signal.

There exists a significant interest in the determination of the absolute configuration of biologically active molecules (van der Helm & Hossain, 1987). Unfortunately, many molecules of interest do not contain atoms heavier than sulfur. In the past, this problem was solved with the introduction of a heavier atom in the structure, *e.g.* with the addition of HBr (Spek, 1976). The current trend is to attempt absolute structure determination on the native compound, even when no atoms heavier than oxygen are present.

Over time a number of methods for the determination of the absolute structure have been proposed.

The most straightforward way of establishing the absolute structure of a small enantiopure molecule is to refine both enantiomers separately, subsequently select the absolute structure with the lowest crystallographic R factor and test for the statistical significance of the R-factor difference. The latter is commonly done with the Hamilton test (Hamilton, 1965).

A much more sensitive method (Zachariasen, 1965; Engel, 1972) is to select a subset of reflections from the measured data that are most sensitive to the absolute structure (relatively weak reflections with a large Bijvoet difference), and compare the calculated Bijvoet differences with the observed differences. Just by comparing the signs of these differences, the absolute structure can often be established even if the difference in R factor is inconclusive. Although the absolute structure can be determined using this method, it is not easy to quantify the degree of certainty of the assignment. Le Page *et al.* (1990) present a method to accompany an absolute structure determined in this way by a calculation of the probability that the absolute structure should be inverted. For this calculation use a binomial distribution. This method has not found widespread use, and therefore its performance is difficult to assess.

Another variation on this method is used by the Bijvoet program in the *DIRDIF* program suite (Beurskens *et al.*, 1980;

Beurskens *et al.*, 1999). This program uses a weighted average of the signs of the Bijvoet difference (B). This method can be very successful, but it needs a carefully selected subset of Bijvoet pairs to be effective. The absolute structure assignment using this calculation is accompanied by a standard uncertainty, but it is very hard to establish the statistical correctness of this value as it relies on distributions being Gaussian, and disregards the careful selection of the reflection subset. Also, no difference can be seen in this calculation between a racemic twin and a weak resonant scattering signal: both will result in smaller absolute values of B and larger standard uncertainties.

Rogers (1981) was the first to introduce a parameter that can be refined as part of the least-squares refinement. This parameter encodes the 'strength' and sign of the measured resonant scattering signal measured in units of f'', the imaginary component of the complex atomic scattering factor.

The Rogers η parameter was soon superseded by the Flack x parameter (Flack, 1983). The Flack x parameter encodes the relative abundance of the two components in an inversion twin. The value of the Flack x parameter can be determined using a full-matrix least-squares procedure [*e.g.* with the TWIN/BASF instructions in *SHELXL97* (Sheldrick, 1997)]. A reasonable estimate of the Flack x parameter can be obtained by determining the parameter separately; this is automatically performed for all non-centrosymmetric structures in the *SHELX97* package. Since the Flack x parameter can correlate with the atomic coordinates, especially for structures in space groups that do not have a fixed origin, the estimate can be inaccurate if its value deviates significantly from zero (Flack & Bernardinelli, 2006; Flack *et al.*, 2006).

Since the value of the Flack x parameter is the result of a least-squares refinement, its standard uncertainty can be derived from the covariance matrix. This standard uncertainty can be used to quantify the degree of confidence in the proposed absolute structure. Flack & Bernardinelli (2000) discuss criteria for the reliability of the absolute structure assignment based on the standard uncertainty in the Flack x parameter value. Their analysis, starting from only the standard uncertainty, has to assume that the distribution is normal also in its tails. The paper does not distinguish between the probability of obtaining the absolute structure given the observations and the probability of the observations given the absolute structure. The Bayesian prior relating the two probabilities is ignored. The Flack & Bernardinelli (2000) method does not result in a quantitative statement about the absolute structure assignment.

Parsons & Flack (2004) recently introduced a variation of the refinement of the Flack x parameter. Their method relies on the careful determination of a few selected Bijvoet differences [as the ratio $(I_+ - I_-)/(I_+ + I_-)$] which can either be obtained directly from a good redundant data set or by carefully adding some extra observations. Parsons & Flack (2004) show that this method increases the sensitivity of the absolute structure determination.

Dittrich *et al.* (2006) recently reported advances in absolute structure determinations made using 'invarioms'. Invarioms are aspherical scattering factors that take into account electron density deformations. Using invarioms instead of the normal spherical scattering factors can improve figures of merit as well as the standard uncertainties of all refined parameters. Experience has shown that by using invarioms, the standard uncertainty in the value of the Flack x parameter can be significantly reduced, and that the calculated value of the Flack x parameter is frequently closer to 0.0.

In the next section, we introduce a new way to determine the absolute structure. The method applies Bayesian statistics to the Bijvoet differences. The result of this approach is a series of probabilities for different hypotheses of the absolute structure. The solution with the highest probability can be determined, and this can be used to map the results to a value with standard uncertainty that can be directly compared with the value of the Flack x parameter.

2. Theory and methods

For each Bijvoet pair of reflections, we can define

$$\Delta F_c^2(h) = F_c^2(h) - F_c^2(-h) \tag{1}$$

and

$$\Delta F_o^2(h) = F_o^2(h) - F_o^2(-h). \tag{2}$$

Here, F_c^2 are calculated intensities and F_o^2 are observed intensities. If we assume completely independent observations of the two reflection intensities,

$$\sigma^2_{\Delta F_o^2(h)} = \sigma^2_{F_o^2(h)} + \sigma^2_{F_o^2(-h)}. \tag{3}$$

Now, we can define a variable z as follows:

$$z_h = \frac{\Delta F_c^2(h) - \Delta F_o^2(h)}{\sigma_{\Delta F_o^2(h)}}. \tag{4}$$

If the absolute structure of the model for calculation of the structure factors is correct, and we assume the calculated intensities to be correct (*i.e.* they do not carry a standard uncertainty) the probability distribution of z is a standard normal Gaussian[1]

$$p(z_h) = \frac{1}{(2\pi)^{1/2}} \exp(-z_h^2/2). \tag{5}$$

Based on all pairs of observations, we can now calculate the probability of the measured data, given the fact that the absolute structure is correctly specified in the model (the correct absolute structure is noted as the condition $y = 0$, it will become clear later in the paper why this notation is chosen):

$$p(\text{observations} \mid y = 0) = \prod_h p(z_h). \tag{6}$$

[1] One referee remarked that this may not be completely true in practice, especially far from the mean. However, large deviations from the mean occur when the differences are much larger than the standard uncertainty, in which case the absolute structure assignment should be obvious in any case. The practical Gaussian nature of the distribution and the usability of the calculated standard uncertainty can also be verified by the normal probability plot analysis as mentioned later in the paper.

In statistics, Bayes' theorem for conditional probabilities specifies that

$$p(x \mid y) = \frac{p(y \mid x)\, p(x)}{p(y)}. \tag{7}$$

In our case, we can use this theorem to invert our probability given above:

$$p(y = 0 \mid \text{observations}) = \frac{p(\text{observations} \mid y = 0)\, p(y = 0)}{p(\text{observations})}. \tag{8}$$

This value cannot be computed, as p(observations) (the probability of obtaining the current observations) is unknown. But to be able to make the absolute structure assignment, we would like to calculate the ratio of $p(y = 0 \mid \text{observations})$ and the similar term for the opposite absolute structure, designated as $p(y = 1 \mid \text{observations})$. The term of p(observations) disappears in the calculation of this ratio:

$$\frac{p(y = 0 \mid \text{observations})}{p(y = 1 \mid \text{observations})} = \frac{p(\text{observations} \mid y = 0)\, p(y = 0)}{p(\text{observations} \mid y = 1)\, p(y = 1)}. \tag{9}$$

And if no prior knowledge about the absolute structure exists [*i.e.* $p(y = 0) = p(y = 1)$],

$$\frac{p(y = 0 \mid \text{observations})}{p(y = 1 \mid \text{observations})} = \frac{p(\text{observations} \mid y = 0)}{p(\text{observations} \mid y = 1)}. \tag{10}$$

To be able to do this, we define a value q analogous to z:

$$q_h = \frac{-\Delta F_c^2(h) - \Delta F_o^2(h)}{\sigma_{\Delta F_o^2(h)}}. \tag{11}$$

This value represents z_h for the inverted structure. Now

$$p(q_h) = \frac{1}{(2\pi)^{1/2}} \exp(-q_h^2/2) \tag{12}$$

and

$$p(\text{observations} \mid y = 1) = \prod_h p(q_h). \tag{13}$$

Hence

$$\frac{p(y = 0 \mid \text{observations})}{p(y = 1 \mid \text{observations})} = \prod_h \frac{p(z_h)}{p(q_h)} = \exp\left[1/2\left(-\sum_h z_h^2 + \sum_h q_h^2\right)\right]. \tag{14}$$

If the correct absolute structure and wrong absolute structure hypotheses are the only two possibilities for a certain structure, this would be sufficient. However, in practice a structure may be a twin consisting of two inverses (so-called inversion twins), and a more general probability model is desired to express this. The twinning can be described as a linear combination of the two structure factors of the pure enantiomeric structures. For each Bijvoet pair,

$$\Delta F_c^2(\text{twin}) = x\Delta F_c^2(y = 1) + (1 - x)\Delta F_c^2(y = 0). \tag{15}$$

This linear combination is analogous to the definition of the Flack x parameter (Flack, 1983). Since $\Delta F_c^2(y = 1) = -\Delta F_c^2(y = 0)$, this equation can be simplified to

$$\Delta F_c^2(\text{twin}) = (1 - 2x)\Delta F_c^2(y = 0) \equiv \gamma \Delta F_c^2(y = 0). \tag{16}$$

We refer to the variable γ as the 'generalized absolute structure'. For the correct absolute structure $\gamma = 1.0$ (and $x = 0$), and for the wrong absolute structure $\gamma = -1.0$ (and $x = 1$). With the help of this parameter, we can now introduce for each reflection h the function $x(\gamma)$:

$$x(\gamma) = \frac{\gamma \Delta F_c^2 - \Delta F_o^2}{\sigma_{\Delta F_o^2}} \tag{17}$$

It can be seen easily that z_h is equal to $x_h(\gamma = 1)$ and q_h is equal to $x_h(\gamma = -1)$. Note that this computation is also allowed with physically impossible values of $|\gamma| > 1.0$. With this generalization, the probability distribution becomes

$$\log p(\text{observations} \mid \gamma) \simeq \sum_h \frac{-x_h^2(\gamma)}{2} = -\tfrac{1}{2}\sum_h x_h^2(\gamma), \tag{18}$$

and from Bayes' theorem,

$$p(\gamma \mid \text{observations}) = \frac{p(\text{observations} \mid \gamma)\, p(\gamma)}{p(\text{observations})}. \tag{19}$$

We can now avoid the need to calculate p(observations) in two ways: we can either have a discrete set of possible hypotheses for the value of γ, or we can study the continuum of all possible γ values.

In the case of the discrete set $\gamma_1, \gamma_2, \ldots, \gamma_n$, we can normalize easily:

$$p(\gamma_i \mid \text{observations}) = \frac{p(\text{observations} \mid \gamma_i)\, p(\gamma_i)}{\sum_j p(\text{observations} \mid \gamma_j)\, p(\gamma_j)}. \tag{20}$$

In most cases, all priors $p(\gamma_i)$ will be set to $1/n$. Two useful discrete sets of hypotheses that can be treated this way are the two-membered set *the absolute structure is either correct or wrong* and the three-membered set *the absolute structure can be correct or wrong, but the sample may also be a 50/50 inversion twin*.

If we want to consider the whole continuum of possible values of γ, the normalization of the probability function is less meaningful. In the case of a γ continuum, only ratios of different probabilities should be used, and these ratios do not depend on the normalization. However, all of the probability numbers are exceedingly small. For numerical stability reasons, it is advisable to bring all relative probabilities that we want to use in calculations to a reasonable size. To achieve this we can simply divide by a high value of the probability function. For this goal, we chose to use $p(\gamma)$ with $\gamma = \gamma_0$. We call the 'incompletely normalized' result p_u:

$$p_u(\gamma) = \frac{p(\text{observations} \mid \gamma)\, p(\gamma)}{p(\text{observations} \mid \gamma_0)\, p(\gamma_0)}. \tag{21}$$

Table 1
Samples studied.

Conditions are as follows. (1) Measured on a Bruker AXS SMART APEX system with an Mo $K\alpha$ sealed-tube X-ray generator at room temperature. (2) Measured on a Nonius KappaCCD system with a Nonius Mo $K\alpha$ rotating-anode X-ray generator at 100 K. (3) Measured on a Bruker AXS SMART 6000 system with a Cu $K\alpha$ sealed-tube X-ray generator with graphite monochromator at 100 K. (4) Measured on a Bruker AXS SMART 6000 system with a Siemens Cu $K\alpha$ rotating-anode tube and focusing multilayer optics at 100 K. (*a*) Data integrated using *EvalCCD* (Duisenberg *et al.*, 2003) and scaled using *SADABS* (Sheldrick, 1996). (*b*) Data integrated with *DENZO* (Otwinowski, 1993) and scaled using *SCALEPACK* (Otwinowski, 1993). (*c*) Data integrated using *SAINT* (Bruker, 2004) and scaled using *SADABS*. $R1 = \sum(|F_o| - |F_c|)/\sum|F_o|$. AMBI is ammonium bitartrate, M048A is threonine, M049A is glutamic acid, M050A is ammonium bitartrate, M051A is alanine. The rest of the data were supplied to us by a pharmaceutical company.

Sample	Conditions	Redundancy	Space group	$R1$ (%)	Asymmetric unit	Resonant scattering signal ($\times10^4$)
AMBI	1*a*	3.5	$P2_12_12_1$	2.40	$C_4H_9NO_6$	9.0
M006C	2*b*	2.2	$P1$	3.61	$C_5H_5LiN_2O_5$	8.4
S3130A	2*b*	8.1	$P2_12_12_1$	3.09	$C_9H_{10}N_2O_3$	7.0
S3350A	2*b*	6.0	$P2_1$	3.47	$C_{13}H_{14}O_5$	7.3
S3351A	2*b*	5.8	$P2_1$	3.89	$C_{13}H_{14}O_5$	7.3
S3456A	2*b*	11.2	$P2_12_12_1$	3.05	$2C_{21}H_{22}N_4O_8 + CH_3OH$	7.3
S3385A	2*b*	6.8	$P2_12_12_1$	2.67	$3C_6H_8O_4$	8.1
M048A	2*a*	11.5	$P2_12_12_1$	2.58	$C_4H_9NO_3$	8.1
M049A	2*a*	12.8	$P2_12_12_1$	2.66	$C_5H_9NO_4$	8.2
M050A	2*a*	13.2	$P2_12_12_1$	2.25	$C_4H_9NO_6$	9.0
M051A	2*a*	8.7	$P2_12_12_1$	2.53	$C_3H_7NO_2$	7.9
T0001	3*c*			2.23	$C_3H_7NO_2$	43
N0951	4*c*			2.32	$C_{35}H_{48}O_{10}$	37
N1045	3*c*			6.64	$C_{25}H_{31}NO_5$	34
N1021	3*c*			2.51	$C_{25}H_{31}NO_5$	34
T0002	3*c*			2.31	$C_5H_{10}N_2O_3$	42
T0003	3*c*			2.72	$2C_{13}H_{21}NO_2$	32
N1099	3*c*			2.71	$C_{23}H_{30}N_2O_2$	29
N0965	3*c*			2.32	$C_{15}H_{14}N_2O_2$	32
N1040	3*c*			2.31	$4C_{15}H_{14}N_2O_3$	34
N0942	3*c*			2.44	$C_{19}H_{23}NO_3$	33
N1069	3*c*			2.62	$C_{26}H_{28}N_4O_2$	29
T0004	3*c*			2.44	$0.5C_3H_{12}N_6O_3$	42
N1000	3*c*			4.13	$4C_{15}H_{14}N_2O_3$	34
N0990	3*c*			6.87	$C_{35}H_{30}N_4O_4$	31
N0973	3*c*			6.15	$2C_{16}H_{26}N_2O_5$	37

Since this approach is most useful if no prior knowledge is assumed at all (note that we always have prior information, namely $-1 \leq \gamma \leq 1$, but here we explicitly choose to ignore this), we simplify it to

$$p_u(\gamma) = \frac{p(\text{observations} \mid \gamma)}{p(\text{observations} \mid \gamma_0)}. \tag{22}$$

It is observed in practice that $p_u(\gamma)$ (in the second definition) is a reasonably well behaved Gaussian-like function. We can therefore calculate[2] a quantity G:

$$G = \frac{\int \gamma p_u(\gamma)\,\mathrm{d}\gamma}{\int p_u(\gamma)\,\mathrm{d}\gamma}. \tag{23}$$

Using this definition, G is the best approximation of γ for the structure based purely on the observations and not using any prior knowledge (not even the physical restriction that γ must be in the interval $[-1, 1]$). Since in our practical experience $p_u(\gamma)$ looks very much like a symmetric Gaussian distribution, G will also be very close to the most probable value of γ. Like the value of the Rogers η parameter, the value of G will be close to 1.0 for structures for which the absolute structure of the model is correct, and close to -1.0 for structures for which the absolute structure of the model is incorrect. Continuing along this path, we can calculate the variance of the distribution using

$$\sigma^2(G) = \frac{\int (\gamma - G)^2 p_u(\gamma)\,\mathrm{d}\gamma}{\int p_u(\gamma)\,\mathrm{d}\gamma}. \tag{24}$$

This can be used to estimate an uncertainty in the obtained value of G.

The concept of the unrestricted absolute structure parameter G follows naturally from the comparison of the definitions of z and q. This is, however, a new concept. With a simple change of parameter expression we can cast our result in a form comparable with the Flack x parameter:

$$y = (1 - G)/2 \tag{25}$$

and

$$\sigma_y = \sigma_G/2. \tag{26}$$

With this definition, y behaves like the Flack x parameter in that it will have a value of 0.0 for the correct absolute structure model, and 1.0 for the inverted model.

3. Test calculations

Table 1 lists several data sets that were collected on different instruments. Some of these data sets happened to be of

[2] The integrals can be computed using a summation with a suitably small step size, where the bounds of γ are chosen such that $p_u(\gamma)$ at the bounds is insignificantly small.

interest at that time; others were specially collected to test the statistical methods introduced in this paper.

All of the structures have weak resonant scattering signals. Roughly half of the data sets were collected using Mo $K\alpha$ radiation. The theoretical resonant scattering signal at $2\theta = 0^\circ$ was estimated for each of the data sets from

$$\Delta F/F = \left(\frac{2\sum_i N_i f_i''^2}{\sum_i N_i f_i^2}\right)^{1/2}. \qquad (27)$$

Both summations run over atom types i, N_i is the number of atoms of type i in the structure, f is the scattering factor of the atom type, and f'' the imaginary part of the resonant scattering factor (Weiss *et al.*, 2001). For Mo $K\alpha$ radiation, $f''(\mathrm{O}) = 0.0060$, $f''(\mathrm{N}) = 0.0033$ and $f''(\mathrm{C}) = 0.0016$. For Cu $K\alpha$ radiation, $f''(\mathrm{O}) = 0.0322$, $f''(\mathrm{N}) = 0.018$ and $f''(\mathrm{C}) = 0.0091$. $\Delta F/F$ is called the 'signal'. This does assume a random distribution of atoms in the cell; locations of resonant scatterers close to symmetry elements can cause weakening of the signal. On the other hand, this formula can be a pessimistic guess since f will decrease for increasing diffraction angles 2θ, whereas the resonant scattering factor f'' is nearly independent of the diffraction angle.

All structures were refined using *SHELXL97*. After refinement, the observed Bijvoet pairs listed in the FCF output file of *SHELXL* were used for an analysis of the value of y. Care was taken not to use FCF files produced by *SHELXL* run using the TWIN/BASF instructions, as in such a case the calculated structure factors already have the Flack x calculation embedded and this would invalidate the analysis. Where available, the absolute configuration assignment was cross-checked with prior information; in other cases the structure for which the value of y was closest to 0.0 was chosen. Results of the analyses are given in Table 2.

Table 2
Absolute structure analyses.

The absolute structure for all samples is determined using four different techniques. (1) The Flack x parameter is refined together with all other structural parameters. (2) The value of y is determined. (3) For a two-hypotheses model (the structure is either right or it is wrong), the probability $p2$(wrong) that the absolute structure assignment was wrong is given. (4) For a three-hypotheses model (the structure is either right or wrong, or it is a 50% inversion twin), the probabilities $p3$(ok), $p3$(twin) and $p3$(wrong) that each of the hypotheses is correct are given.

Data set	Flack x	y	$p2$(wrong)	$p3$(ok)	$p3$(twin)	$p3$(wrong)
AMBI	−0.10 (90)	−0.05 (16)	2×10^{-10}	0.997	0.002	2×10^{-10}
M006C	−0.15 (81)	−0.28 (50)	0.04	0.721	0.248	0.031
S3130A	0.24 (91)	0.31 (41)	0.2	0.398	0.473	0.129
S3350A	−1.01 (81)	−0.50 (44)	0.006	0.868	0.126	0.005
S3351A	0.39 (92)	−0.13 (47)	0.06	0.671	0.289	0.041
S3456A	−0.28 (51)	0.06 (17)	2×10^{-7}	0.969	0.031	2×10^{-7}
S3385A	0.16 (48)	0.17 (20)	3×10^{-4}	0.726	0.274	2×10^{-4}
M048A	0.70 (107)	0.24 (32)	0.07	0.491	0.470	0.039
M049A	−0.20 (97)	0.24 (35)	0.1	0.480	0.461	0.059
M050A	−0.34 (81)	0.14 (18)	1×10^{-5}	0.846	0.154	1×10^{-5}
M051A	−0.00 (60)	−0.06 (20)	2×10^{-6}	0.976	0.024	1×10^{-6}
T0001	−0.02 (20)	0.01 (3)	$<10^{-100}$	1.000	1×10^{-39}	6×10^{-164}
N0951	0.00 (9)	0.00 (1)	$<10^{-100}$	1.000	7×10^{-80}	$<10^{-300}$
N1045	−0.15 (26)	0.02 (8)	3×10^{-33}	1.000	1×10^{-8}	3×10^{-33}
N1021	0.01 (11)	0.00 (1)	$<10^{-100}$	1.000	4×10^{-63}	3×10^{-259}
T0002	0.07 (18)	0.05 (5)	$<10^{-100}$	1.000	2×10^{-65}	2×10^{-292}
T0003	−0.05 (12)	−0.01 (4)	$<10^{-100}$	1.000	3×10^{-34}	2×10^{-130}
N1099	0.04 (15)	0.07 (5)	7×10^{-84}	1.000	4×10^{-18}	7×10^{-84}
N0965	−0.10 (16)	−0.04 (5)	$<10^{-100}$	1.000	1×10^{-47}	2×10^{-173}
N1040	0.06 (9)	0.10 (2)†	$<10^{-100}$	1.000	1×10^{-43}	2×10^{-213}
N0942	0.01 (12)	0.01 (3)	$<10^{-100}$	1.000	2×10^{-45}	5×10^{-188}
N1069	−0.07 (14)	−0.02 (5)	4×10^{-90}	1.000	7×10^{-24}	4×10^{-90}
T0004	0.05 (28)	0.07 (6)	2×10^{-60}	1.000	2×10^{-13}	2×10^{-60}
N1000	0.00 (19)	−0.05 (6)	3×10^{-77}	1.000	1×10^{-21}	3×10^{-77}
N0990	−0.01 (17)	−0.03 (8)	3×10^{-41}	1.000	2×10^{-11}	3×10^{-41}
N0973	−0.04 (28)	−0.06 (13)	8×10^{-15}	1.000	1×10^{-4}	8×10^{-15}

† For N1040, y deviates significantly from an enantiopure value.

4. Results and discussion

The power of the introduced method comes from the fact that it weights each observed Bijvoet difference based on its expected accuracy directly, rather than relying on the weight of the reflection intensities. Calculating these proper weights for a least-squares procedure is very difficult, but proper weighting can be rather easily accomplished with the derived maximum-likelihood procedure instead of using least-squares. Bijvoet differences can be much smaller than the residual differences between the observed and calculated intensities, and the calculated differences are accurate as long as the resonant scatterers have been accurately positioned.

It is essential to measure Bijvoet pairs for the calculation of y, where the Flack x parameter can be determined even if the data set covers at least the asymmetric unit corresponding to the space group with an added inversion centre.

We only tested our methods on data sets with close to 100% coverage of Bijvoet pairs.

When prior information is given, *e.g.* that the sample must be either the structure or its inverse, the method presented can be used to calculate probabilities of the two possible hypotheses [$p2$(ok) and $p2$(wrong)]. These probabilities can be surprisingly decisive, even when the resonant scattering signal is very weak. For the test data sets measured using Cu $K\alpha$ radiation, the chirality of all structures can be proven beyond reasonable doubt if it is assumed (prior knowledge) that the original compound was enantiopure. For the three-hypotheses model where the additional possibility of a 50% inversion twin cannot be ruled out, the distinction given by the probabilities [we call these $p3$(ok), $p3$(twin) and $p3$(wrong)] is less pronounced, but even in that case many of the determined values for the Cu $K\alpha$ data sets would satisfy the most stringent pharmaceutical requirements. The least surprising results are obtained when the whole continuum of inversion twin structure compositions must be considered. In this case, the estimate that is obtained as the value of y has a smaller standard uncertainty than the value of the Flack x parameter. In most cases, the value of y is also closer to zero than the value of the

Flack x parameter. Structures for which the continuum approach is *required* have not been studied in this paper. Using the continuum approach to solve the binary absolute structure question (as is commonly done with the Flack x parameter in existing studies of bioactive compounds) is suboptimal. In contrast, the use of $p2$(ok) and $p2$(wrong) directly gives quantitative reliability information.

For some data sets, calculations were performed both on the correct and on the inverted model, refined in *SHELXL*. For the AMBI data set, the $p3$(wrong) value of 1.64×10^{-10} increases to $p3(\text{ok}) = 2.3 \times 10^{-10}$ when the inverted structure is refined. The small difference between these values shows that the inverted refinement cannot absorb more than a small fraction of the resonant scattering signal into the other refined structural parameters. Comparing the equivalent numbers for the M006C data set, which has a weaker resonant scattering signal and which does not have a fixed origin, shows a similarly sized relative increase from $p3$(wrong) = 0.031 to $p3$(ok) = 0.043 for the inverted structure. The magnitude of this structural bias is largely insignificant for the absolute structure determination of pharmaceutically active compounds. It may, however, be significant for accurate determination of the twin ratio of inversion twins; this has not been the subject of our study.

There are two assumptions in the derivation of the probabilistic model: firstly, that the standard uncertainty of the two reflections that form each Bijvoet pairs are independent; secondly, that the standard uncertainties of the individual reflections are accurate.

Both of these conditions are necessary conditions for $x(\gamma)$ to follow a standard normal distribution. These assumptions can be verified by making a normal probability plot (Abrahams & Keve, 1971) from all values $x(\gamma = 1.0)$. Such normal probability plots, made for the data sets above, show that the observed distribution of $x(\gamma = 1.0)$ for most data sets indeed follows a Gaussian distribution (the correlation coefficient of the normal probability plot is 0.999) but with $\sigma < 1.0$. Two possible reasons can be suggested. (i) The used scaling programs overestimate the errors in the reflection intensities. This is highly unlikely. (ii) The measurement error in the Bijvoet difference is smaller than could be expected if the two errors in the reflection intensities were independent. The errors are in fact positively correlated, and the error in the Bijvoet difference is really smaller.[3]

The second hypothesis is most likely. Even without knowing the source of the smaller standard uncertainty, it is possible to use the information obtained from the normal probability plot to scale the standard uncertainties in the Bijvoet differences, thereby obtaining a corrected $x(\gamma)$. This correction scales down the standard uncertainties in γ in all but two of the cases that were examined for this paper.

[3] The positive correlation could be caused by the fact that there are many Friedel pairs in our data sets. For a Friedel pair the diffraction geometry could be more similar than for general Bijvoet pairs. This could cause systematic errors to cancel. A four-circle goniostat could be employed to extend these advantages. This is an interesting subject for a future study.

Table 3
Correction of the calculated value of y for the error in the standard uncertainties as derived from a normal probability plot.

Z is the deviation of y from the enantiopure value expressed in units of σ_y.

Structure	y (before)	Z (before)	NPP slope	y (after)	Z (after)
AMBI	−0.05 (15)	−0.301	0.840	−0.05 (13)	−0.358
M006C	−0.28 (50)	−0.577	0.990	−0.28 (49)	−0.582
M048A	0.24 (32)	0.753	0.881	0.24 (28)	0.855
M049A	0.24 (35)	0.680	0.912	0.24 (32)	0.745
M050A	0.14 (18)	0.798	0.931	0.14 (17)	0.855
M051A	−0.06 (20)	−0.279	0.940	−0.06 (19)	−0.292
S3130A	0.31 (41)	0.748	0.822	0.31 (34)	0.909
S3350A	−0.50 (44)	−1.136	1.050	−0.50 (46)	−1.087
S3351A	−0.13 (47)	−0.277	1.035	−0.13 (49)	−0.265
S3456A	0.06 (17)	0.327	0.947	0.06 (16)	0.344
S3385A	0.17 (20)	0.832	0.840	0.17 (17)	0.988

The validity of such a downscaling of the errors can be confirmed by studying the result for a group of independent structure determinations and determining the value $Z = y/\sigma_y$ for each of them. If all standard uncertainties have been determined correctly, the values of Z from a random population of structure determinations should form a standard normal distribution. For the structure determinations using Mo $K\alpha$ radiation given in this paper, the average absolute value of Z is 0.61 (the expected value is 0.85) and the root mean square (r.m.s.) value of Z is 0.67 (expected 1.0) (Table 3). These results suggest that the error is indeed systematically overestimated. After applying the slope from the normal probability plot to correct the estimated standard uncertainties in the observed Bijvoet differences, the average absolute value of Z is 0.66 and the r.m.s. value of Z is 0.72. These values are still smaller than the expected values. The current benchmark set is too small for this to be considered proof of the merits of the downscaling procedure.

4.1. Centrosymmetric structures

Flack & Bernardinelli (2006) and Flack *et al.* (2006) investigated the value of the Flack x parameter for a set of centrosymmetric structures that were refined in a non-centrosymmetric space group. Looking at the definition of the Flack x parameter,

$$F_h^{\text{ref}} = (1 - x)F_h^{\text{calc}} + xF_{-h}^{\text{calc}}, \tag{28}$$

it can be clearly seen that for the correct model, x is indeterminate since the two terms F_h^{calc} and F_{-h}^{calc} are equal. The determination of x in these cases is therefore based purely on the random incorrect differences between the two 'half-structures' in the refinement. In this light, it is at first sight surprising and discomforting that the values observed have such small standard uncertainties. It is clear that for the Flack x parameter the assumption that the off-diagonal elements of the covariance matrix may be ignored is wrong. The assumption that all other parameters have been determined correctly by the least-squares refinement has been violated.

We have attempted a non-centrosymmetric solution and refinement of a centrosymmetric ruthenium-containing compound (Hotze *et al.*, 2005) ourselves to investigate this

effect further. For this structure the Flack x parameter is 0.56 (4) and the value of y is 0.45 (3). Both values are close to 0.5 with a relatively small standard uncertainty. A detailed analysis of the data set indicated that the small standard uncertainty is due to a few reflections for which the differences between the two half-structures create a significant Bijvoet difference ΔF_c while, as expected for a centrosymmetric structure, the ΔF_o value is statistically insignificant. Such pairs are normally indicative of twinning by inversion.

The only statistical difference in reciprocal space between a real inversion twin and a wrongly refined centrosymmetrical structure is that the calculated Bijvoet differences are much smaller than for a normal non-centrosymmetric structure with the same elemental composition. This is due to the fact that the configuration of the atoms is almost centrosymmetric (with respect to a suitably chosen origin, the phases of many reflections are close to 0 and π and the phases of the resonant scattering contributions are close to $\pi/2$ and $3\pi/2$) and hence the resonant scattering contribution to the scattering factors only results in relatively small scattering amplitude differences. It is difficult to determine a reliable criterion for this effect.

It appears then that the distinction between a true inversion twin and a non-centrosymmetrically refined centrosymmetric structure is best made in real space by a symmetry-detection procedure like *ADDSYM* (Spek, 2003), followed by a detailed inspection of the weak reflections after refinement in the suggested centrosymmetric space group.

5. Recommended procedure

Current versions of refinement programs cannot use the value of y to take absolute structure into account. We therefore recommend to refine the structure including the Flack x parameter (*e.g.* use the TWIN/BASF instructions in *SHELXL*). The value of y can then be determined separately using a utility that explicitly calculates structure factors for the Bijvoet pairs (*e.g.* the Bijvoet Pairs option in *PLATON*).[4]

This procedure will account for any correlation between the structural parameters and the absolute structure.

6. Conclusions

A new probabilistic procedure was introduced that can be used to establish the absolute structure. The procedure is especially suitable for biologically active compounds, which often do not contain atoms with a larger resonant scattering signal than that of oxygen.

The only special requirement for the data collection procedure imposed by the new probabilistic calculation is that Bijvoet pairs should be present in the data set. In contrast, the determination of the Flack x parameter also works for data sets that have a Bijvoet pair coverage of 0%, although this is not recommended practice.

One of the results of the procedure is a value y, which can be directly compared with the value of the Flack x parameter. We observe for our test data sets that the standard uncertainty in the value of y is roughly half of the standard uncertainty in the value of x. The observed deviation from 0.0 is consistent with the standard uncertainty. These observations are comparable with the results obtained using invarioms but without the significant efforts associated with the calculation of invarioms.

The calculations also give explicit probabilities for the absolute structure assignment, without referring to the value of y and without the assumption that the distribution of y is Gaussian. The explicit probability of an absolute structure assignment error makes our procedure suitable to regulate the probability of erroneous assignments in pharmaceuticals. The probability calculations can be based either on a model with two hypotheses for the two absolute structures or optionally take the chance of a racemate into account as a third hypothesis.

The procedure was tested on a number of light-atom structures (no atoms with a stronger resonant scattering signal than that of oxygen). For those data sets collected using Mo $K\alpha$ radiation, a mixed result was obtained: some structures could receive quite a good absolute structure assignment; most structures show at least a clear direction. For all data sets measured using Cu $K\alpha$ radiation (resulting in a roughly five times larger resonant scattering signal) an excellent absolute structure discrimination was obtained with the chance of error for most structures below 10^{-100}. Of course, for most of the Cu $K\alpha$ structures the standard uncertainty in the Flack x parameter is small enough for an unambiguous assignment.

The current method offers an alternative method to look at the same experimental data as addressed by the Flack x approach.

7. Availability

An implementation of the described algorithm by one of the authors is available in his *PLATON* (Spek, 2003) program (http://www.cryst.chem.uu.nl/platon/pl000000.html).

The authors wish to thank Anita Coetzee, Martin Lutz, George Sheldrick and Bill David for stimulating discussions before and during the 2005 Crystallographic Computing School in Siena, Italy, and Martin Lutz, Huub Kooijman and Trixie Wagner for testing the method on their data sets. The authors also express their appreciation to all referees of this paper for very valuable comments on the subject.

References

Abrahams, S. C. & Keve, E. T. (1971). *Acta Cryst.* A**27**, 157–165.
Beurskens, G., Noordik, J. H. & Beurskens, P. T. (1980). *Cryst. Struct. Commun.* **9**, 23–28.
Beurskens, P. T., Beurskens, G., de Gelder, R., García-Granda, S., Gould, R. O., Israel, R. & Smits, J. M. M. (1999). *The DIRDIF99 Program System*, Technical Report of the Crystallography Laboratory, University of Nijmegen, The Netherlands.

[4] For a TWIN/BASF refinement, *SHELXL* will write calculated structure factors into the FCF file that take the inversion twinning ratio into account. Such an FCF file can therefore not be used to calculate the value of y.

Bijvoet, J. M., Peerdeman, A. F. & van Bommel, A. J. (1951). *Nature (London)*, **168**, 271–272.
Bruker (2004). *SAINT*. Bruker AXS Inc., Madison, Wisconsin, USA.
Dittrich, B., Strumpel, M., Schäfer, M., Spackman, M. A. & Koritsánszky, T. (2006). *Acta Cryst.* A**62**, 217–223.
Duisenberg, A. J. M., Kroon-Batenburg, L. M. J. & Schreurs, A. M. M. (2003). *J. Appl. Cryst.* **36**, 220–229.
Engel, D. W. (1972). *Acta Cryst.* B**28**, 1496–1509.
Flack, H. D. (1983). *Acta Cryst.* A**39**, 876–881.
Flack, H. D. & Bernardinelli, G. (2000). *J. Appl. Cryst.* **33**, 1143–1148.
Flack, H. D. & Bernardinelli, G. (2006). *Inorg. Chim. Acta*, **359**, 383–387.
Flack, H. D., Bernardinelli, G., Clemente, D. A., Linden, A. & Spek, A. L. (2006). *Acta Cryst* B**62**, 695–701.
Flack, H. D. & Shmueli, U. (2007). *Acta Cryst.* A**63**, 257–265.
Glazer, A. M. & Stadnicka, K. (1989). *Acta Cryst.* A**45**, 234–238.
Hamilton, W. C. (1965). *Acta Cryst.* **18**, 502–510.
Helm, D. van der & Hossain, M. B. (1987). *Patterson and Pattersons*, edited by J. P. Glusker, B. K. Patterson & M. Rossi, pp. 482–495. Oxford University Press.
Hotze, A. C. G., van der Geer, E. P. L., Kooijman, H., Spek, A. L., Haasnoot, J. G. & Reedijk, J. (2005). *Eur. J. Inorg. Chem.* **2005**, 2648–2657.
Jones, P. G. (1984). *Acta Cryst.* A**40**, 663–668.
Le Page, Y., Gabe, E. J. & Gainsford, G. J. (1990). *J. Appl. Cryst.* **23**, 406–411.
Otwinowski, Z. (1993). *Data Collection and Processing*, pp. 56–62, *Proceedings of the Daresbury CCP4 Study Weekend*, compiled by L. Sawyer, N. Isaacs & S. Bailey. Warrington: Daresbury.
Parsons, S. & Flack, H. (2004). *Acta Cryst.* A**60**, s61.
Rogers, D. (1981). *Acta Cryst.* A**37**, 734–741.
Sheldrick, G. M. (1996). *SADABS*. University of Göttingen, Germany.
Sheldrick, G. M. (1997). *SHELX97*. University of Göttingen, Germany.
Spek, A. L. (1976). *Acta Cryst.* B**32**, 870–877.
Spek, A. L. (2003). *J. Appl. Cryst.* **36**, 7–13.
Weiss, M. S., Sicker, T. & Hilgenfeld, R. (2001). *Structure*, **9**, 771–777.
Zachariasen, W. H. (1965). *Acta Cryst.* **18**, 714–716.