Scaling and Phase-Difference Determination in Solvent Contrast Variation Experiments

BY M. ROTH
Institut Laue-Langevin, 156X, 38042 Grenoble CEDEX, France
AND A. LEWIT-BENTLEY AND G. A. BENTLEY
European Molecular Biology Laboratory, 156X, 38042 Grenoble CEDEX, France

(Received 24 June 1983; accepted 1 November 1983)

Abstract

In the solvent contrast variation technique, the structure factor is a linear function of contrast, and the intensity is therefore a parabolic function of contrast. A method has been developed which scales together data measured at different contrasts on different samples. Once the data are scaled, structure-factor amplitudes or intensities for any other contrast can be obtained by interpolation or extrapolation. The magnitude of the relative phase change between contrast, an important piece of phasing information in single-crystal studies, can be determined.

1. Introduction

Solvent contrast variation is a technique used in low-resolution structural studies of complex macromolecules to distinguish between their chemically distinct components, such as protein and nucleic acid in viruses or chromatin, by virtue of their different mean scattering densities. The technique depends on varying the scattering density of the solvent around the macromolecules so as to match out the scattering at low resolution from each of the species in turn. In this way the structure of each component may be analysed separately. Contrast variation is applicable to crystallographic studies, as well as to solution scattering, since the volumic fraction of mother liquor present in crystals of macromolecules is significant. Furthermore, these studies may be performed with X-rays (Stuhrmann & Kirste, 1965; Luzzati, Tardieu, Mateu & Stuhrmann, 1976), neutrons (Stuhrmann, 1974, 1976; Jacrot, 1976) or the electron microscope (Kühlbrandt, 1982). The scattering density of the solvent or medium is adjusted with electron-dense salts or sucrose in the case of X-rays, with D2O in the case of neutrons and with a stain such as gold glucose in the case of electron microscopy.

In a series of neutron diffraction experiments performed at different concentrations of deuterium, c, the complex structure factor for a given reflection hkl will be a linear function of c (Worcester & Franks, 1976):

\[ F = F_0 + cF_1, \]  

where \( F_0 \) and \( F_1 \) are two constants which depend on \( hkl \). The value \( c \) is defined as the mole fraction of atoms of D (i.e. \( ^2\text{H} \)) with respect to \( ^1\text{H} \) in the solvent and in all the hydrogen sites on the solute where the \( ^1\text{H} \) atoms can be exchanged for \( ^2\text{H} \) atoms by equilibration with the solvent. We assume that the D/H ratio is the same in the solvent and on these sites in the solute. One may therefore equate this mole fraction to the % of \( \text{D}_2\text{O} \) in the solvent, provided all solutes that contain exchangeable hydrogens have been taken into account for the calculation of the D/H ratio. Since many authors use the terms solvent contrast or % D2O when referring to \( c \), we use these terms interchangeably.

The experimentally measured quantities are the intensities \( I = FF^* \), which therefore vary as

\[ I(hkl) = F_0F_0^* + c(F_0F_1^* + F_0^*F_1) + c^2F_1F_1^* \]  

or, in a different notation,

\[ I = I_0 + cI_1 + c^2I_2. \]  

In the case of solution scattering the measured intensity for a given value of the scattering vector \( \mathbf{q} \) is
also a parabolic function of \( c \). One can write
\[
I(q) = \langle F_0 F_0^* \rangle + c \langle F_0 F_1^* + F_0^* F_1 \rangle + c^2 \langle F_1 F_1^* \rangle. \tag{4}
\]
The mean values indicated by the symbol \( \langle \rangle \) are a consequence of the orientational disorder of the molecules in solution, of their polydispersity, of the non-ideality of the incident beam and the spatial resolution of the detector (smearing).

In the following section we describe a method of estimating values of \( I_0, I_1 \) and \( I_2 \), from a series of experimentally determined intensities from crystal diffraction or solution scattering in various contrasts. Once these quantities are known it is possible to determine \( I \) for any contrast. The method provides estimates of the variances and covariances for the calculated quantities. In the case of diffraction, the calculations presented here permit the direct determination of the moduli and the cosines of the phase differences between \( F, F_0 \) and \( F_1 \).

2. Scaling

Diffraction data in different contrasts are usually measured from different crystals. The first problem to solve is therefore the scaling of different series of data. The scaling method proposed here can equally well be used for scaling solution scattering data. In this case, the whole intensity curve, or any selected regions of it, can be used, unlike the methods relying on extrapolation to \( I(0) \) which use only the limited Guinier region (this region can also suffer from interparticle interference in concentrated solutions). Furthermore, knowledge of the concentrations of the macromolecule at the different contrasts is not required.

For scaling data sets measured at different solvent contrasts \( c \) we have a set of intensities \( I(hkl) \) or \( I(q) \), measured for the same values of \( hkl \) or \( q \) for the \( n \) different contrasts. Once the data are on a common scale the intensities should fulfill
\[
S_i I(j,i) = I_0(j) + c_1 I_1(j) + c_2 I_2(j). \tag{5}
\]
In this and the following expressions the index \( i \) refers to the solvent contrast \( (i=1, 2, 3, ..., n) \) and \( j \) is the index of the value of \( hkl \) or \( q \) under consideration \((j=1, 2, 3, ..., N_2, \text{ where } N_2 \) is the total number of reflections or \( q \) values used). The factor \( S_i \) is the scale factor for intensities from contrast \( i \). In the case of centrosymmetric reflections in diffraction experiments, or for centrosymmetric molecules in solutions where smearing effects are weak or have been corrected for, (5) can be simplified to become linear:
\[
\pm s_i [I(l,i)]^{1/2} = F_0(l) + c_1 F_1(l), \tag{6}
\]
where \( l=1, 2, ..., N_1 \) refers to the reflections \( hkl \) (or \( q \) for solution scattering) in question, \( s_i \) is the scale factor of the amplitudes \( F \) \( (s_i^2 = S_i) \) and \( F \) is the modulus of \( F \). The determination of scale factors thus becomes a problem of solving the system of equations formed by (5) and (6), or derived from them in the case of a least-squares solution. At the same time one obtains the values of \( I_0(j), I_1(j) \) and \( I_2(j) \) or \( F_0(l) \) and \( F_1(l) \), which is, in fact, the final aim of the calculation. The determination of scale factors, however, is done using only certain reflections, i.e. those measured with good accuracy. The determination of the values of the \( I_0, I_1 \) etc. is therefore performed for the complete data set after scaling, in a separate pass.

2.1. Prescaling

A first estimate of scale factors can be obtained by solving the following system of equations, derived from (5):
\[
S_i I(j,i) - (c_i - c_1) I_1(j) - (c_1^2 - c_2) I_2(j) = I(j,1) \tag{7}
\]
with \( i=2, ..., n \) and \( j=1, ..., N_2 \) and taking the first contrast as a reference \((S_1 = 1)\). This is an overdetermined system of \((n-1)N_2 \) equations for \( n-1 + 2N_2 \) unknowns. Similarly one can solve:
\[
e_i S_i [I(l,i)]^{1/2} - (c_i - c_1) F_1(l) = e_1 [I(l,1)]^{1/2} \tag{8}
\]
with \( i=2, ..., n \) and \( l=1, ..., N_1 \), which is a system of \((n-1)N_1 \) equations and \( n-1 + N_1 \) unknowns. In the latter system of equations there is a problem of sign which is designated in (8) by the symbol \( \epsilon_i = \pm 1 \) and \( \epsilon_1 = \pm 1 \). In order to solve the system we adopt the convention of \( F_1(l) > 0 \). We thus obtain
\[
\epsilon_1 \leq \epsilon_2 \leq \ldots \leq \epsilon_n
\]
if
\[
c_1 < c_2 < \ldots < c_n
\]
which limits the possible number of sign combinations for each \( hkl \) or \( q \) value to \((n+1)\). The method we use to find the correct sign combination for each possible combination is to retain, for each reflection, the combination that minimizes the sum of residuals \( R_i(l) \) for this reflection:
\[
R_i(l) = \sum_{i=1}^{n} \{ s_i \epsilon_i [I(l,i)]^{1/2} - F_0(l) - c_1 F_1(l) \}^2 / s_i^2 \sigma_i^2 \tag{9}
\]
with
\[
F_0(l) = (1/n) \sum_{i=1}^{n} \{ s_i \epsilon_i [I(l,i)]^{1/2} - c_1 F_1(l) \}. \tag{10}
\]
The quantity \( \sigma_i^2 \) is the standard deviation of \([I(l,i)]^{1/2} \), obtained from the experimental variance \( v_{II} \) of \( I(l,i) \) using (Rees, 1976)
\[
\sigma_i^2 = (1/2) \left[ v_{II} / [I(l,i)] \right]^{1/2}. \tag{11}
\]

2.2. Scaling by least-squares minimization

A more accurate determination of the scale factors can be obtained using a least-squares approach. Such
a method will at the same time allow a critical estimate of the accuracy and of the reliability of the results, based on a statistical analysis.

From the point of view of a statistical analysis, the scale factors are considered as defined functions of the observed data. The relationship is given by least-squares minimization equations where the random variables are the data. The approach chosen by Taupin & Luzzati (1982), on the other hand, uses the a posteriori probability formalism to determine the parameters, which are also treated as a set of random variables. The advantages of our method are: the handling (inversion) of large matrices is easily avoided; it does not require Gaussian random variables; and that, finally, it permits a simple evaluation of correlations between scale factors and the measured intensities.

The least-squares method consists of finding the scale factors which minimize the sum of residuals:

\[ R = \sum_{i=1}^{N_1} R_1(l) + \sum_{j=1}^{N_2} R_2(j) \]  

with \( R_1(l) \) given by (9) and

\[ R_2(j) = \sum_{i=1}^{n} [S_i I(j, i) - l_0(j) - c_1 I_1(j) - c_2 I_2(j)]^2 / \sigma_i^2 \]  

\[ \text{(13)} \]

Formally this is equivalent to solving a system of equations obtained by setting the partial derivatives of \( R \) with respect to various parameters to zero. This is written

\[ \frac{\partial R_2(j)}{\partial l_k} = 0 \]  

\[ \text{with } j = 1, 2, ..., N_2 \text{ and } k = 0, 1, 2. \]  

\[ \frac{\partial R_1(l)}{\partial F_k(l)} = 0 \]  

\[ \text{with } l = 1, 2, ..., N_1 \text{ and } k = 0, 1. \]

\[ \sum_{i=1}^{N_1} \delta R_1(l) / \delta s_i + \sum_{j=1}^{N_2} \delta R_2(j) / \delta S_i = 0 \]  

\[ \text{with } i = 1, 2, ..., n. \]

To simplify the resolution of this system of equations we divide the data set into two subsets which contain the non-centrosymmetric and the centrosymmetric reflections, respectively. This is possible because the two subsets are statistically independent, since all measured intensities are statistically independent (see also §2.3). This subdivision is made by imposing on (16)

\[ \sum_{j=1}^{N_2} \delta R_2(j) / \delta S_i = 0 \]  

\[ \text{(17)} \]

\[ \sum_{i=1}^{N_1} \delta R_1(l) / \delta s_i = 0. \]  

\[ \text{(18)} \]

Solving separately system (I), consisting of (14) and (17), and system (II), composed of (15) and (18), we obtain two different estimates of scale factors, from which we take a weighted mean (see §2.6 Merging of scale factors).

We shall now describe rapidly one way of solving the equations of system (I). System (II) is solved in exactly the same way. Differentiating (17) we can write

\[ \sum_{j=1}^{N_2} [S_i I(j, i) - l_0(j, i)] A(j, i) = 0 \]  

\[ \text{with } \]  

\[ I_0(j, i) = I_0(j) + c_1 I_1(j) + c_2 I_2(j) \]  

\[ \text{(20)} \]

and

\[ A(j, i) = I_0(j, i) / \sigma_i^2 v_{ji} \]  

\[ \text{(21)} \]

for \( i = 1, 2, ..., n. \) The \( 3N_2 \) values of the parameters \( I_0(j), I_1(j) \text{ and } I_2(j) \) are obtained by solving separately the \( N_2 \) subsystems of three equations of three unknowns formed by equations (14) for each of the \( N_2 \) values of \( j. \) This operation is in fact a parabolic regression of the measured intensities \( I(j, i) \) (for given \( j \)) with respect to concentrations \( c, \) the intensities having been multiplied by the supposedly known scale factors. In the case of system (II) (linear regression) it is at this stage that the choice of a correct sign is made by trial and error. Starting from a first estimate of scale factors (e.g. from prescaling or from a previous iteration) the correct combination of signs is found by selecting the one which gives the smallest sum of residuals \( R_1(l). \)

The values of \( I(j, i) \) need not be known for all values of \( c \) for any given \( hkl \) or \( q \) value. The minimum necessary number of contrasts, \( c, \) for which intensities must be known is four for the parabolic case and three for the linear case. [For a number of contrasts smaller than these minima the sums \( R_2(j) \text{ and } R_1(l) \) are identically equal to 0, whatever the value of \( S_i \) or \( s_i. \)]

The results of the regression is finally an expression of \( l_0(j) \) as a function of the scale factors \( S_i: \)

\[ l_0(j, i) = \sum_{k=1}^{n} a_k(j, i) S_k, \]  

\[ \text{(22)} \]

where the coefficients \( a_k(j, i) \) are known expressions of intensities \( I(j, i), \) variances \( v_{ji}, \) and concentrations \( c. \) We thus end up with a system of equations

\[ \sum_{k=1}^{n} S_k B(k, i) = 0 \]  

\[ \text{(23)} \]

with

\[ B(k, i) = \sum_{j=1}^{N_2} [I(j, k) \delta_{ik} - a_k(j, i)] A(j, i) \]  

\[ \text{(24)} \]

\( i = 1, 2, ..., n \) and \( \delta_{ik} = 1 \text{ or } 0 \text{ if } i = k \text{ or } i \neq k, \) respectively. The system (23) is in fact a system of \( (n - 1) \) independent equations of \( n \) unknowns. The reason is that the residuals \( R_1 \text{ and } R_2 \) are homogenous functions of the scale factors and of the contrasts. The system is therefore solved choosing one contrast as a reference,
for which \( S = 1 \) and discarding one of the equations. It can then be solved as a linear system with constant coefficients \( B \), iterating on the values of \( B \). The calculations usually converge quite rapidly.

The scaling method proposed here can be considered as a generalization of the method of Hamilton, Rollett & Sparks (1965) to the problem of scaling different sets of data. Each set is then characterized by the value \( c_i \) of a parameter \( c \) and each element of these sets varies independently of others as a function of the parameter \( c \). The expressions given by Hamilton, Rollett & Sparks (1965) are found by suppressing the dependence on the parameter \( c \) in our relationships. In this case, the index \( i \) will refer to a given data set rather than to a contrast. The necessary equations are obtained from (9) and (10) by setting \( F_i(l) = 0 \) and \( \sigma_i = 1 \). Since in practice a complete data set for one contrast may in fact be composed of measurements from different samples, we have written the program to treat this kind of scaling as well. We used the iterative method of solving the equations as indicated above and the evaluation of the variance–covariance matrix of the scale factors as explained below.

2.3. Variance–covariance and bias on scale factors

The scale factors determined by this method are formally functions of the intensities and contrasts. Because of statistical errors on experimentally determined intensities, the estimation of scale factors is subject to two errors:

- a systematic error (bias) whose magnitude is given by
  \[
  \bar{dS}_i = \frac{1}{S_i} \sum (\partial^2 S_i / \partial I^2) \nu_i; \tag{25}
  \]
  a statistical error of variance–covariance given by
  \[
  \bar{dS}_i dS_k = \sum (\partial S_i / \partial I) (\partial S_k / \partial I) \nu_i. \tag{26}
  \]

The sums in these expressions are over all intensities \( I \) used in the calculation. The statistical errors on the intensities are assumed to have zero covariances and variance \( \nu_i \). These expressions are obtained by a formal Taylor expansion of the \( S_i \) as functions of the intensities in the vicinity of their expectation values. The expansions are limited to the first order in the variances. The quantities on the left-hand side of (26) form the variance–covariance matrix of the scale factors. In the linear case, similar expressions are obtained if \( S_i \) is replaced by \( S \) and \( I \) by \( I^{1/2} \) in (25) and (26). In practice the partial derivatives in (25) and (26) are evaluated numerically by finite increments to the intensities and recalculating each time the new values for the scale factors. The results of these calculations (see §4, Discussion: tests of the method) are reliable only if the data are good, i.e. without systematic errors on the intensities, with correctly estimated variances and, finally, with small deviations of the values of \( I \) with respect to their expectation values. This method of determination of variance and bias is similar to that used in the statistical physics of fluctuations theory (Landau & Lifshitz, 1978).

2.4. Statistical distribution

Another method for evaluating the quality of the determination of scale factors is the analysis of the distribution of residuals. We consider the \( N_2 \) and \( N_1 \) values of the sums of residuals over the \( n \) contrasts per reflection \( hkl \) or \( q \) value, i.e. the quantities \( R_2(k) \) and \( R_1(l) \) for \( j = 1, \ldots, N_2 \) and \( l = 1, \ldots, N_1 \). These quantities measure the accuracy of the fit of the scaled data to the theoretical parabolic or linear variation given by (5) and (6) with respect to the contrast, for each reflection or \( q \) value. We compare the distributions of the \( R_2 \) and \( R_1 \) values with the theoretical \( \chi^2 \) distributions with \((n-4)\) and \((n-3)\) degrees of freedom, respectively, assuming that the intensities are Gaussian random variables. This comparison is strictly valid only if the scale factors are known exactly. Since the scale factors are in fact determined from the same set of intensities as the residuals, however, these \( \chi^2 \) distributions must be considered only as approximations which come closer to the true distributions as more reflections are used. The validity of this approximation was proved qualitatively by numerous tests on model data (Figs. 1a, b). From these distributions one can judge the quality of the estimates of variances of the measured intensities and detect any anomalies in the intensity values. Thus it is possible to use the \( \chi^2 \) distribution to remove from the calculation those reflections whose values of \( R_1 \) or \( R_2 \) are abnormally high due to systematic errors of measurement.

2.5. Effect of the accuracy of the contrasts

The derivatives of scale factors with respect to contrasts, \((1/S_i) \partial S_i / \partial c_k (i = 1, \ldots, n, k = 1, \ldots, n)\), are easy to calculate numerically using a method of finite increments. If the contrasts are expressed in %, the values of these derivatives are typically of the order of 0.01 to 0.1. The value of the contrasts must therefore be known with very good accuracy for a good estimate of scale factors. One can ask if it is possible, on the other hand, by minimizing the sums of residuals with respect to contrasts, to improve the accuracy of the determination of scale factors by a refinement on the contrasts. The answer to this question depends on the value of the quantities \( \partial \Delta / \partial c \), which are easy to evaluate. In the parabolic case, for example,

\[
\partial R_2 / \partial c_i = -2 \sum_{j=1}^{N_1} \left( I_i(j) + 2c_i I_2(j) \right) \left( S_i I(j,i) - I_4(j,i) / (S_i^2 \nu_i) \right). \tag{27}
\]

One finds that for the minimum number of contrasts \( m \), \( m = 3 \) or 4 in the linear and parabolic case, respec-
tively, these derivatives are identically equal to zero and refinement is impossible. If \( n > m \), however, it is possible to minimize the value of \( R \) by changing the values of \( n - m \) contrasts while keeping the other \( m \) constant. This operation is only valid (see §4. Discussion: tests of the method) if applied to a contrast for which one has good reason to suspect an error. A relatively high value of \( \partial R / \partial c_i \) is a possible indication of such an error on \( c_i \), as the value of \( \partial R / \partial c_i \) is a weighted sum, for the contrast \( c_i \), of the differences \( (I_{\text{calc}} - I_{\text{obs}}) \) for all data corresponding to this contrast [see (27)].

### 2.6. Merging of scale factors

When scaling crystal data, one usually has two sets of scale factors, \( s_i \) and \( S_i \) obtained with centrosymmetric and non-centrosymmetric reflections, respectively. In order to make use of both (independent) determinations one has to merge the two sets of scale factors using least squares. In matrix notation we have

\[
S' = (M_1^{-1} + M_2^{-1})^{-1} (M_1^{-1} S_1 + M_2^{-1} S_2),
\]

where \( S' \) is the vector of least-squares averaged scale factors.

---

**Fig. 1.** Distribution of residuals \( R_1 \) and \( R_2 \) corresponding to the data in Table 1. (a) and (b) Histograms of residual distributions from (a) linear \( (R_1) \) and (b) parabolic \( (R_2) \) scaling compared with the theoretical histograms (broken lines) deduced from the \( \chi^2 \) distributions with three and two degrees of freedom, respectively (see §2.4). (c) and (d) Distribution of residuals \( R_2 \) from the calculation of phase differences (§3) on the same data, (c) for 50 centrosymmetric and ten pseudo-centrosymmetric reflections and (d) for 40 non-centrosymmetric reflections, compared with the theoretical \( \chi^2 \) histograms with three and two degrees of freedom, respectively.
factors, $S_1$ and $S_2$, the vectors of the first and second set of scale factors and $M_1$ and $M_2$ the corresponding variance-covariance matrices. The set $S_1$ is composed of values of $s_1^2$. The elements $(i, k)$ of the corresponding variance-covariance matrix are given by $4s_i s_k m_{ik}$, where $m_{ik}$ is the element $(i, k)$ of the variance-covariance matrix of the $s_i$. The variance-covariance matrix of $S'$ is given by

$$M_s = (M_1^{-1} + M_2^{-1})^{-1}. \tag{29}$$

3. Phase difference and interpolation

The ultimate aim of this data treatment in a series of contrast variation crystallographic experiments is the determination of $F_0$ and $F_1$ and the phase difference $\phi$ between $F_0$ and $F_1$ for every $hkl$. This can be achieved using a weighted least-squares solution of (5) for $i = 1, 2, ..., n$, with respect to the unknowns $l_0, l_1$ and $l_2$, i.e. by solving (14) for every value of $j$. The $(n, n)$ variance-covariance matrices involved are those corresponding to the quantities $S_1 I(j, i)$. The elements $M(k, m)$ of this matrix can be derived from

$$M(k, m) = d[S'_k I(j, k)] d[S'_m I(j, m)], \tag{30}$$

where $S'_k$ is the $k$th element of $S'_1$, and $i = 1, 2, ..., n$, and $m = 1, 2, ..., n$. This may be rewritten

$$M(k, m) = M_s (k, m) I(j, k) I(j, m) + S'_k S'_m \delta_{km} \delta_{jk} + I(j, k) S'_m [\partial S'_k / \partial I(j, m)] v_{jm} + I(j, m) S'_k [\partial S'_m / \partial I(j, k)] v_{jk}. \tag{31}$$

In this equation, the derivatives $\partial S'/\partial I$ will be equal to zero except in the cases where the intensities $I(j, k)$ or $I(j, m)$ were used in the calculation of the scale factors $S$ or $s$. The values of these derivatives can then be easily obtained from (28) and from the values of $\partial S/\partial I$ or $\partial s/\partial I^{1/2}$ that have already been calculated, to obtain the matrices $M_1$ and $M_2$ (26).

Once the systems of equations (14) have been solved with respect to $l_0, l_1$ and $l_2$, the quantities $F_0$ and $F_1$ are determined using the expressions

$$F_0 = I_0^{1/2}; \quad F_1 = I_1^{1/2}; \quad \cos \phi = 0.5 \left( I_1 / (I_0 I_2) \right)^{1/2}. \tag{32}$$

The variance and covariance of these quantities are determined in the usual way (Hamilton, 1964). An acceptable solution of these equations must fulfil the condition

$$l_0 > 0; \quad l_2 > 0; \quad \cos^2 \phi \leq 1,$$

which can be expressed as

$$l_1^2 \leq 4l_0 l_2. \tag{33}$$

Centrosymmetric reflections can be treated in the same way with the constraint $\cos \phi = \pm 1$. The approach we use is to solve the equations (14) with the additional constraint

$$l_1^2 = 4l_0 l_2. \tag{34}$$

This is performed using the Lagrange multiplier method as given by Hamilton (1964) for the case of a linear constraint: the solution with constraint is a function of the solution without constraint. The corresponding equation has to be solved by iteration, since in our case this function is implicit.

For non-centrosymmetric reflections the condition (33) is not always fulfilled. In such a case it can be demonstrated that the best least-squares estimates of $l_0, l_1$, and $l_2$ are those that fulfil the condition (34). The best solution for these reflections therefore is of the centrosymmetric type. In practice this centrosymmetric solution for non-centrosymmetric reflections occurs either for reflections that in fact are nearly centrosymmetric or for weak reflections whose intensities have a relatively large error and may even be negative. In the latter case the phase information provided by contrast variation will be rather poor.

Here again, $\chi^2$ distributions (as described in §2.4) can be used to test the quality of data and calculations (Figs. 1c, d). The sums of residuals $R_1$ and $R_2$ for every centrosymmetric or non-centrosymmetric reflection, respectively, are calculated with the values obtained for $l_0, l_1$ and $l_2$.

4. Discussion: tests of the method

4.1. Scaling tests

The scaling method was tested on artificially generated data (centrosymmetric, $F$'s, and non-centrosymmetric, $l$'s), comparing results from linear and parabolic scaling procedures for their reliability and their sensitivity to perturbations. Each value of $l$ or $F^2$ was perturbed with a random error according to a Gaussian distribution with a variance equal to $I$ and $F^2$, respectively. We used between 20 and 50 reflections and typically five contrasts for these tests.

Solution scattering test data were generated from a spherical shell model of a molecule, convoluted with instrumental smearing (using an algorithm communicated to us by S. Cusack) and randomized with Gaussian errors. The variance was taken proportional to $I/q$ to simulate experimental data obtained with a two-dimensional position-sensitive detector.

The results of numerous tests can be summarized as follows (an example is given in Table 1).

(1) Prescaling is useful to initiate the iterative solution of scaling by least squares, although in many cases the correct solution can be obtained directly, with starting values for the scale factors equal to $I$.

(2) In the least-squares scale-factor determination: (a) the bias is usually small compared to the calculated standard deviation and can be neglected. Note that the bias given by (25) is the bias due to the non-
change in the contrast can introduce a wrong sign and allowed to refine. The result was within 1% of the correct value. If, however, the displaced contrast was fixed and another, initially correct, allowed to refine, the latter refined to a value 7% off its correct value. The error was accommodated by the scale factors for all contrasts. The R factor and the sums of residuals did not increase significantly, thus giving no indication of a false refinement. In the case of linear scaling, these contrast refinements have to be used with care, because of the sign assignment: a change in the contrast can introduce a wrong sign attribution, thus leading to a wrong overall solution.

4.2. Calculation of the relative phases

The calculation of relative phase differences was tested using data derived in the following way: experimental structure-factor amplitudes from Satellite Tobacco Necrosis Virus (STNV) crystals measured at four different contrasts (0, 25, 68 and 99% D$_{2}$O; G. A. Bentley, A. Lewit-Bentley, L. Liljas, U. Skoglund & T. Unge, unpublished results) were scaled and their phase differences calculated. An ideal data set was derived replacing the $I_{obs}$ by the $I_{calc}$ according to (20). This ideal data set consisted of 857 unique reflections $hkl$, of which 68 were centrosymmetric (STNV belongs to the space group C2). A set of 'model' data was derived by perturbing the ideal data with a Gaussian random error.

STNV is a virus comprised of sixty protein subunits (21 600 daltons each) and a single strand of RNA (340 000 daltons). The protein density is matched by 40% D$_{2}$O buffer while the RNA is matched by 68% D$_{2}$O buffer. The 40% D$_{2}$O contrast was not measured experimentally since the RNA represents only about 20% of the total scattering matter; because of limited neutron beam time another contrast, 25% D$_{2}$O, was chosen, since here the diffraction from the RNA is stronger. The structure amplitudes at the 40% D$_{2}$O contrast were then obtained by interpolation.

Scaling of the data was rather sensitive to the displacement of one contrast: a 7% displacement produced a 10% change of the scale factor for the highest contrast (99% D$_{2}$O). The scale factor for 0% D$_{2}$O was taken as reference. Table 2 summarizes the analysis of the quality of interpolation for the 40% D$_{2}$O contrast. The value $\phi_{ij}$ is in fact the relative phase difference between the perturbed and interpolated 40% D$_{2}$O contrast. It is interesting to note that a small (up to 5% of the real value) error on the D$_{2}$O contrast does not have a significant effect on the interpolation and relative phase calculation. As could be expected, the relative phase calculation and the interpolation are more accurate with increasing value of $F$.

5. Conclusions

The tests described above show that provided the data are free of systematic errors (e.g. properly corrected for background), the scale factors and their variances are estimated in a reliable way. Crystallographic data must be corrected for absorption and if possible multiple reflection prior to scaling. In solution scattering experiments, care must be taken to eliminate multiple scattering and interparticle effects.

Extinction in crystallographic data is easily detected in the case of centrosymmetric reflections since these will deviate from linearity [[6]] in a characteristic way. For non-centrosymmetric reflections extinction is less easily detected on an individual basis. The effect of extinction is best examined by studying the variation of the scale factor as a function of the upper intensity threshold used in the selection of reflections. We have

| Table 1. Example of scaling using test data as described in text |
|----------------------|--------|--------|--------|--------|--------|
| Contrasts (% D$_{2}$O) | 0      | 39     | 65     | 90     | 100    |
| Theoretical scale factors | 1      | 2.250  | 4.000  | 0.250  | 1.440  |
| Scale factors from 50 centrosymmetric reflections | 1      | 2.272  | 4.064  | 0.254  | 1.470  |
| Standard deviation | 0      | 0.020  | 0.053  | 0.004  | 0.024  |
| Bias correction | 0      | 0.000  | 0.000  | 0.000  | 0.000  |
| Scale factors from 50 non-centrosymmetric reflections | 1      | 2.222  | 3.933  | 0.238  | 1.272  |
| Standard deviation | 0      | 0.046  | 0.133  | 0.011  | 0.069  |
| Bias correction | 0      | 0.026  | 0.081  | 0.005  | 0.041  |
| Merged scale factors | 1      | 2.264  | 4.072  | 0.252  | 1.465  |
| Standard deviation | 0      | 0.018  | 0.048  | 0.004  | 0.022  |
| Correlation matrix | 1.000  | 0.000  | 0.000  | 0.000  | 0.000  |
| 0.000  | 1.000  | 0.942  | 0.925  | 0.908  |
| 0.000  | 0.942  | 1.000  | 0.972  | 0.962  |
| 0.000  | 0.925  | 0.972  | 1.000  | 0.995  |
| 0.000  | 0.908  | 0.962  | 0.995  | 1.000  |

Linearity of the least-squares problem. It does not account for possible bias due to systematic errors in the data;

(b) the estimated standard deviation is a good measure of the effective error on the scale factors;

(c) the accuracy of the scale-factor determination depends on the accuracy of the data and increases with the number of reflections used in the calculation. It is usually of the order of 1 to 10% and is better with linear than with parabolic scaling;

(d) the scale factors are highly correlated (correlation factors of 0·1 to 1;

(e) the distribution of residuals is very well represented by a $\chi^2$ distribution.

(f) the sensitivity of the scale factors to the value of the contrasts has already been mentioned (§2.5). The effectiveness of contrast refinement can be illustrated with the following example. With test data, the value of one solvent contrast was displaced by 12% from its true value and allowed to refine. The result was within 1% of the correct value. If, however, the displaced contrast was fixed and another, initially correct, allowed to refine, the latter refined to a value 7% off its correct value. The error was accommodated by the scale factors for all contrasts. The $R$ factor and the sums of residuals did not increase significantly, thus giving no indication of a false refinement. In the case of linear scaling, these contrast refinements have to be used with care, because of the sign assignment: a change in the contrast can introduce a wrong sign attribution, thus leading to a wrong overall solution.
Table 2. The effect of displacing the 68% D₂O contrast on the calculated value of diffraction amplitude (F) and phase difference (φ) at 40% D₂O

<table>
<thead>
<tr>
<th>Data set (D₂O displacement) (%)</th>
<th>Scale factor for 99% D₂O</th>
<th>Scale factor for perturbed contrast</th>
<th>Number of pseudo-centrosymmetric reflections</th>
<th>Mean difference in φ (°)</th>
<th>R.M.S. difference in φ (°)</th>
<th>R factor on F (%)</th>
<th>Total R factor (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1·0</td>
<td>1·0</td>
<td>232</td>
<td>15·7</td>
<td>29·15</td>
<td>6·2</td>
<td>3·3</td>
</tr>
<tr>
<td>68 → 63</td>
<td>1·1</td>
<td>1·0</td>
<td>123</td>
<td>24·7</td>
<td>38·21</td>
<td>11·5</td>
<td>7·2</td>
</tr>
<tr>
<td>68 → 66</td>
<td>1·05</td>
<td>1·0</td>
<td>159</td>
<td>19·2</td>
<td>32·87</td>
<td>7·6</td>
<td>4·4</td>
</tr>
<tr>
<td>68 → 69</td>
<td>0·97</td>
<td>1·0</td>
<td>263</td>
<td>15·4</td>
<td>27·8</td>
<td>6·6</td>
<td>3·8</td>
</tr>
<tr>
<td>68 → 73</td>
<td>0·88</td>
<td>0·97</td>
<td>361</td>
<td>19·6</td>
<td>27·1</td>
<td>11·2</td>
<td>7·8</td>
</tr>
</tbody>
</table>

*Total R factor (R = \( \frac{\sum F_{ij}^\text{model} - F_{ij}^\text{calc}}{\sum F_{ij}^\text{model}} \)) is calculated for all contrasts used.

found that in the case of STNV, where a neutron wavelength of 11·8 Å was used for all data collections, only one or two very intense low-angle reflections were affected in this way.

We find that the scaling procedure is in fact a good tool for verifying to what extent the data may be biased by these systematic effects. In practice one usually has to discard a small amount of the experimental data and here the distributions of R₁ and R₂ are used to operate this selection.

For a crystallographic analysis, the most useful information derived from contrast variation using these calculations is the relative phase. Thus, if a set of centrosymmetric phases is known for one contrast, they may be calculated for any other contrast. In the case of non-centrosymmetric structure factors, there is an ambiguity of sign in the calculation of phases for an unknown contrast from a known one. This is a problem analogous to that of single isomorphous replacement and the ambiguity must be broken by introducing further information. The possible use of the constraint provided by the relative phases in contrast variation in the direct-methods approach to phasing has been discussed elsewhere (Bentley, Finch, Lewit-Bentley & Roth, 1982).

The use of phase information derived from contrast variation has been used in crystallography before, both for X-rays (Bragg & Perutz, 1952; Jack, Harrison & Crowther, 1975) and neutrons (Worcester & Franks, 1976; Bentley, Finch & Lewit-Bentley, 1981). The method we have presented here is, however, very general, and provides statistical analysis of the data and the results.

In conclusion, we draw attention to the fact that (1) is valid to any resolution on the assumption that hydrogen-bond differences are insignificant and that the D/H exchange in the molecule has reached equilibrium. The analysis outlined in this paper can therefore be usefully applied to high-resolution crystal structure determination of biological macromolecules.

The authors are indebted to Dr A. Antoniadis (University of St Etienne) for many helpful discussions and wish to thank Dr S. Cusack (EMBL, Grenoble) who kindly communicated to them his solution scattering data on spherical viruses for testing the scaling procedures.

References