J. Appl. Cryst. (1992). **25**, 26–30

A Combined Isotropic and Multiple *s*-Shell Anisotropic Scaling Method for Multiple Data Sets

BY FUSAO TAKUSAGAWA

Department of Chemistry, University of Kansas, Lawrence, KS 66045, *USA*

(*Received* 6 *June* 1990; *accepted* 19 *July* 1991)

Abstract

A method to improve the scaling of multiple intensity data sets is presented. A general scaling function $K(x, s)$, which uses the direction cosine of the diffraction vector (x) and $(\sin\theta)/\lambda$ (s) as scaling parameters, is developed by combining an isotropic scaling function, $K(s) = A\exp(Bs^2)$, and a multiple s-shell anisotropic scaling function, $K(x)_s = (\sum\sum c_{ij}x_{ij})_s$. This combined scaling function can greatly reduce the systematic differences in intensities among multiple data sets measured independently. This scaling method for the multiple data sets consists of three steps. In the first step, the individual isotropic scaling functions, $K(s)$, are determined by an indirect least-squares method. Then the weighted mean intensity, $\langle I\rangle$, is calculated by applying the $K(s)$ to the individual data sets. In the second step, the data in each data set are divided equally into 20 thin shells of $(\sin\theta)/\lambda$ (s). The anisotropic scaling functions, $K(x)_s$, of each s shell are determined by using the weighted mean intensity, $\langle I\rangle$, obtained in the first step as the target quantity in a least-squares minimization, *i.e.* $\sum w_i\{\langle I\rangle_i - K(x)_s[K(s)I_i]\}^2$. In the final step, the new weighted mean intensity, $\langle I\rangle$, is calculated by applying the combined scaling function, $K(x, s) = K(s)K(x)_s$, to the individual data sets. The new multiple s-shell anisotropic scaling functions are determined using the new weighted mean intensity, $\langle I\rangle$, as the target quantity in another least-squares minimization. By repeating this procedure three to five times, the least-squares minimization will converge. The method was successfully used to scale and merge 27 sets of *S*-adenosylmethionine synthetase data into a single data set. It was also used to scale the isomorphous replacement data sets of the enzyme.

Introduction

Data scaling and merging become major problems when a large number of crystals are used for intensity measurements. The determination of scaling factors between overlapping sets of data is a non-trivial problem which has been discussed by a number of authors (Kraut, 1958; Dickerson, 1959; Rollett & Sparks, 1960; Hamilton, Rollett & Sparks, 1965; Fox & Holmes, 1966; Matthews & Czerwinski, 1975; Rossmann, Leslie, Abdel-Meguid & Tsukihara, 1979). The method of Hamilton *et al.* (1965) has been widely used to combine multiple data sets into a single data set. In this method, each data set has one scale parameter. If no data set contains systematic errors, the method of Hamilton *et al.* is useful and reliable. However, if systematic differences exist between individual data sets, one scaling parameter for each data set is not enough to combine the individual data sets correctly into one reasonable data set. In this paper, we will describe a general scaling function $K(x, s)$ which uses the direction cosine of the diffraction vector (x) and $(\sin\theta)/\lambda$ (s) as scaling parameters. This general scaling function is formulated by combining an isotropic scaling function, $K(s) = A\exp(Bs^2)$, and a multiple s-shell anisotropic scaling function, $K(x)_s = (\sum\sum c_{ij}x_{ij})_s$.

Method

Individual data sets measured with different crystals or different equipment, such as an oscillation camera with film or a single-detector diffractometer or an area-detector diffractometer, often contain some systematic differences. Although we cannot completely eliminate these differences with a relatively simple mathematical method, it is possible to reduce these differences greatly using least-squares methods such as

$$\sum\sum w_i(I_h - K_iI_i)^2 \text{ and } \sum\sum w_i(I_i - G_iI_h)^2$$

where I_h is the best estimate intensity for the reflection h and K_i is a scaling function for the ith data set. G_i is the reciprocal function of K_i, *i.e.* $K_iG_i = 1$. It should be noted that the w_i in $\sum\sum w_i(I_h - K_iI_i)^2$ is not $1/\sigma(I_h)^2$ but $1/\sigma(I_i)^2$. Thus, when the quantity $\sum\sum w_i(I_h - K_iI_i)^2$ is minimized, the I_i and w_i must be updated after every least-squares iteration, *i.e.* the new I_i and w_i are K_iI_i and $1/[K_i\sigma(I_i)]^2$, respectively. K_i should be one at convergence of the least-squares minimization. The simplest scaling function, K_i, is a single parameter. However, if some systematic differences exist between individual data sets, such a simple scaling function will not be able to reduce these differences. The differences between the individual data sets often have strong correlations with the direction cosine of the diffraction vector (x), the

0021-8898/92/010026-05$03.00

magnitude of the intensity (I) and $(\sin\theta)/\lambda$ (s). Therefore, if we can develop a scaling function which uses the direction cosine of the diffraction vector, the magnitude of the intensity and $(\sin\theta)/\lambda$ as scaling parameters, such a scaling function will be able to reduce the systematic differences between the individual data sets greatly.

Very roughly speaking, the magnitude of the intensities can be represented as the simple reciprocal function of $(\sin\theta)/\lambda$, *i.e.* strong reflections are in the low $(\sin\theta)/\lambda$ region whereas the weak reflections are in the high $(\sin\theta)/\lambda$ region. Thus, a simple but reasonable scaling function for the individual data sets can be represented as combined functions of the direction cosine of the diffraction vector (x) and $(\sin\theta)/\lambda$ (s) such as

$$K(x, s) = (c_{11}x^2 + c_{22}y^2 + c_{33}z^2 + 2c_{12}xy + 2c_{12}xy + 2c_{23}yz + 2c_{31}zx)\exp(Bs^n)$$

where x, y, z are the direction cosines of the diffraction vector h,k,l, *i.e.* $x^2 + y^2 + z^2 = 1$, and $n = 2$. Although several different n's in s^n were tested using a simple isotropic scaling function, $K(s)$, described below, $n = 2$ was found to be the most adequate in this kind of approach. The form of $\exp(Bs^2)$ is also supported by the fact that the intensity distribution is a function $\exp(-2Ts^2)$ of the overall temperature factor T. The exponential part can be extended to an anisotropic function using the diffraction vector h,k,l as shown below:

$$\exp(Bs^2) = \exp(d_{11}h^2 + d_{22}k^2 + d_{33}l^2 + 2d_{12}hk + 2d_{23}kl + 2d_{31}lh).$$

If such anisotropic scaling functions, $K(x, s)$, for the individual data sets are correctly determined, the functions will greatly reduce the systematic differences among the individual data sets. The coefficients c_{ij} and d_{ij} in the equations described above can be determined by minimizing $\sum\sum w_i(I_h - K_iI_i)^2$ or its reciprocal form $\sum\sum w_i(I_i - G_iI_h)^2$ using a procedure similar to that described by Hamilton, Rollett & Sparks (1965). However, this approach converges very slowly and, in many cases, the least-squares minimization never converges. Therefore, the $K(x, s)$ function must be modified to a form which can be determined by a simple least-squares method. If the data are divided equally into many thin shells of $(\sin\theta)/\lambda$ (about 20 shells), the scaling function for the data in each thin shell will be independent of $(\sin\theta)/\lambda$. Thus, the $K(x, s)$ function can be replaced with the multiple s-shell functions, $K(x)_s$, as shown below:

$$K(x)_s = (c_{11}x^2 + c_{22}y^2 + c_{33}z^2 + 2c_{12}xy + 2c_{23}yz + 2c_{31}zx)_s.$$

In this case, each $K(x)_s$ represents the scaling function of one of the 20 thin s shells.

If a relatively good 'best estimate I_h' for the reflection h is available, the coefficients c_{ij} in the above equation can be obtained in a straight-forward manner since the least-squares equation is linear. To obtain the initial best estimate I_h, the individual data sets must be appropriately scaled using a simpler isotropic scaling function, such as $K(s) = A\exp(Bs^2)$.

Although the coefficients A and B in the above equation can be obtained using a procedure similar to that described by Hamilton, Rollett & Sparks (1965), the procedure is rather slow because many iterations are required to reach convergence. Thus, an indirect two-step least-squares method is developed to evaluate quickly the coefficients A and B.

The initial best estimate I_h for the reflection h is obtained as a weighted mean intensity as shown below:

$$\langle I\rangle = \sum w_iK(s)I_i/\sum w_i$$

where the w_i is $1/[K(s)\sigma(I_i)]^2$. Then the anisotropic scaling function, $K(x)_s$, of the s shell in each data set is determined by the linear least-squares method using the reflections within the s shell:

$$\sum w_i[\langle I\rangle_i - (c_{11}x^2 + c_{22}y^2 + c_{33}z^2 + 2c_{12}xy + 2c_{23}yz + 2c_{31}zx)K(s)I_il^2.$$

The new best estimate I_h is calculated as a weighted mean intensity after applying the combined scaling function, $K(x, s) = K(s)K(x)_s$:

$$\langle I\rangle = \sum w_iK(x, s)I_i/\sum w_i$$

where the w_i is $1/[K(x, s)\sigma(I_i)]^2$. Then the new multiple s-shell anisotropic scaling functions are determined using the new weighted mean intensity, $\langle I\rangle$, as the target quantity in another least-squares minimization:

$$\sum w_i[\langle I\rangle_i - (c_{11}x^2 + c_{22}y^2 + c_{33}z^2 + 2c_{12}xy + 2c_{33} + 2c_{31}zx)K(x, s)I_i]^2.$$

After the new multiple s-shell anisotropic scaling functions have been obtained by least-squares minimization, the combined scaling functions are updated by multiplying by the new $K(x)_s$, *i.e.* $K(x, s) = K(x)_sK(x, s)$. By repeating this procedure three to five times, the coefficients in the $K(x)_s$ function c_{11}, c_{22} and c_{33} become one and c_{12}, c_{23} and c_{31} become zero suggesting convergence of the least-squares minimization.

It should be noted that the weighting scheme is one of the important parameters in a least-squares procedure. The initial weight, w, for the intensity, I, is taken as $1/\sigma(I)^2$. In the final stage, the weight adjustment factor, K_w, is determined so that the new weight, $w_{new} = K_ww_{old}$, gives equal magnitudes for $\sum w_{new}(\Delta I)^2$ through the entire region of $(\sin\theta)/\lambda$.

In conclusion, the $K(x, s)$ function can be represented by combining the $K(s)$ and $K(x)_s$

functions. The combined function, $K(s)K(x)_s$, can greatly reduce the systematic differences in intensities among the multiple data sets measured independently. The details of the mathematical algorithm of the entire method are described, along with the step-by-step procedures, in the next section.

It should be noted that similar ideas and methods have been published, such as 'local scaling' by Matthews & Czerwinski (1975) and 'anisotropic scaling for film data' by Rossmaan *et al.* (1979).

Description of the algorithm and procedure

Step 1

The relative isotropic scaling function between the ith and jth data sets is defined as $a_{ij}\exp(b_{ij}s^2)$, where s is $(\sin\theta)/\lambda$. The coefficients a_{ij} and b_{ij} are determined by minimizing the following quantity:

$$\sum_{l}^{L} w_l[K_{i_l}I_{i_l} - a_{ij}\exp(b_{ij}s^2)K_{j_l}I_{j_l}]^2$$

where w is $1/\{[K_i\sigma(I_i)]^2 + [K_j\sigma(I_j)]^2\}$ and L is the number of reflections which overlap with the other set. The $\sigma(I_i)$ are $\sigma(I_j)$ are standard deviations of the intensity I_i and I_j, respectively. The K_i and K_j are the scaling factors for intensity I_i and I_j, respectively. The initial K_i and K_j are 1.0 and then the new K_i (or K_j) is updated as a product of the old K_i and $K(s)$ determined in step 2.

Step 2

The coefficients of the individual isotropic scaling function, A and B in $K(s) = A\exp(Bs^2)$, are determined from a_{ij}, b_{ij} and their standard deviations, $\sigma(a_{ij})$ and $\sigma(b_{ij})$, as obtained in step 1. The best A_i are determined by minimizing the following quantity:

$$\Psi_a = \sum_{i}^{M}\sum_{j}^{M} w_{ij}(A_i - A_j a_{ij})^2 \qquad (i \neq j)$$

where w_{ij} is $1/\sigma(a_{ij})^2$ and M is the number of independent data sets. The derivative of A_i is

$$\partial\Psi_a/\partial A_i = 2A_i\sum_{j}^{M}(w_{ij} + w_{ji}a_{ji}^2) + 2\sum_{j}^{M}(-w_{ij}a_j - w_{ji}a_{ji})A_j = 0.$$

The matrix representation of all the derivatives of A's are $[\mathbf{P}_{ij}][\mathbf{A}_i] = 0$ where

$$P_{ii} = \sum_{j}^{M}(w_{ij} + w_{ji}a_{ij}^2)$$

$$P_{ij} = -w_{ij}a_{ij} - w_{ji}a_{ji}.$$

The eigenvectors with the smallest eigenvalue of the matrix $[\mathbf{P}_{ij}]$ should be the best A's.

The best B's are obtained by minimizing the following quantity:

$$\Psi_b = \sum_{i}^{M}\sum_{j}^{M} w_{ij}(b_{ij} - B_i + B_j)^2 \qquad (i \neq j)$$

where $w_{ij} = 1/\sigma(b_{ij})^2$. The derivative of B_i is

$$\partial\Psi_b/\partial B_i = 2B_i\sum_{j}^{M}(w_{ij} + w_{ji}) + 2\sum_{j}^{M}(-w_{ij} - w_{ji})B_j - 2\sum_{j}^{M}(-w_{ij}b_{ij} + w_{ji}b_{ji}) = 0.$$

The matrix representation of all the derivatives of the B's are $[\mathbf{Q}_{ij}][\mathbf{B}_i] = [\mathbf{V}_i]$ where

$$Q_{ii} = \sum_{j}^{M}(w_{ij} + w_{ji})$$

$$Q_{ij} = (-w_{ij} - w_{ji})$$

$$V_i = \sum_{j}^{M}(-w_{ij}b_{ij} + w_{ji}b_{ji}).$$

When the condition that $\sum_i^M B_i = 0$ is imposed on Q_{ii} and Q_{ij}, the results are as follows:

$$Q_{ii} = Q_{ii} + (Q_{ii}/M)$$

$$Q_{ij} = Q_{ij} + (Q_{ii}/M)^{1/2}(Q_{jj}/M)^{1/2}.$$

The best B's are determined by solving the above normal equation $\mathbf{QB} = \mathbf{V}$.

Step 3

The weighted mean intensity, $\langle I\rangle$, is calculated from the following equation:

$$\langle I\rangle = \sum_{i}^{M} w_i K_i I_i \Big/ \sum_{i}^{M} w_i$$

where K_i is a product of old K_i and $K(s)$ determined at step 2, *i.e.* $K_i = K_i K(s)$ and w_i is $1/[K_i\sigma(I_i)]^2$.

The weight adjustment factors, K_w's, are determined so that the new weight, $w_{\text{new}} = K_w w_{\text{old}}$, gives the equal magnitude of $\sum w_{\text{new}}(\langle I\rangle - I)^2$ in 50 shells of $(\sin\theta)/\lambda$. Steps 1 through 3 are then repeated three times.

Step 4

In this step, the multiple s-shell anisotropic scaling functions,

$$K(x)_s = (c_{11}x^2 + c_{22}y^2 + c_{33}z^2 + 2c_{12}xy + 2c_{23}yz + 2c_{31}zx)_s,$$

are determined. The data in each data set are divided equally into 20 thin shells of s. The coefficients of the anisotropic scaling function, c_{ij}, in each shell are determined by minimizing the following quantity:

$$\sum_i^N w_i[\langle I\rangle_i - (c_{11}x^2 + c_{22}y^2 + c_{33}z^2 + 2c_{12}xy + 2c_{23}yz + 2c_{31}zx)K_iI_i]^2$$

where N is the number of reflections in a shell, x, y, z represent the direction cosines of diffraction vector h,k,l and K_i is a scaling factor which is a product of the old K_i and $K(s)$ obtained at step 3 or a product of the old K_i and $K(x)_s$ obtained at the previous iteration. $\langle I\rangle_i$ is the weighted mean intensity obtained at step 3 or step 5.

Step 5

The new scaling factors are updated as a product of the old K_i and $K(x)_s$ obtained at step 4, *i.e.* $K_i = K(x)_sK_i$. The new weighted mean intensity, $\langle I\rangle$, is calculated by applying the updated scaling factors. Steps 4 and 5 are repeated until the $K(x)_s$ are 1.0 for all x, y, z and s, *i.e.* the coefficients c_{11}, c_{22} and c_{33} become one and c_{12}, c_{23} and c_{31} become zero. The weight-adjustment factors described in step 3 are redetermined after every iteration and applied on the individual weights.

Results and discussion

This method has been tested using 27 sets of *S*-adenosylmethionine synthetase data. The intensity data were measured on multiwire area detectors at the University of Virginia and the University of California, San Diego. The R_{sym} of each individual data set ranges from 0.05 to 0.11. The individual data sets contain about 5000 to 8000 unique reflections. Three tests have been carried out:

(1) Application of a single scaling factor, A.

(2) Application of an isotropic scaling factor, $K(s) = A\exp(Bs^2)$.

(3) Application of a combined scaling factor, $K(x, s) = K(s)K(x)_s$.

As shown in Fig. 1, a remarkable improvement in R factors is seen when the combined scaling factors are applied. The final R factors for the combined data up to 3.0 Å resolution (14 274 reflections) are 0.113 for test 1, 0.104 for test 2 and 0.074 for test 3.

The combined scaling functions were also used to scale the isomorphous replacement data sets of *S*-adenosylmethionine synthetase. In this case, the eight sets of UO_2-derivative data were independently scaled to the native data, obtained from the above sets. The following quantity is minimized: $\sum w[I_P - K(x, s)I_{PH}]^2$, where I_P and I_{PH} are intensities of the native and UO_2-derivate data, respectively. After applying the $K(x, s)$ factor to the intensity data, the eight sets of UO_2-derivative data were merged to one data set. The positional, thermal and relative occupancy parameters of U atoms and a relative scale factor (K_r) between the native and derivative data sets were refined using the centric reflections of the native and derivative data by minimizing the following quantity: $\sum w(F_{PH} - K_r|F_P \pm F_H|)^2$, where F_P and F_{PH} are structure factors of native and derivative data and F_H is a calculated structure factor of the U atom. After rescaling the UO_2-derivative data by $1/K_r$, a set of single isomorphous replacement data (h, k, l, F_P, F_{PH}) was created. The phases of reflections were determined by the solvent-flattening procedures developed by Wang (1985) using the refined heavy-atom parameters and the single isomorphous replacement data set. Exactly the same phasing procedures were applied to the data set scaled and merged by using isotropic scaling functions, $K(s) = A\exp(Bs^2)$. The electron density maps of these two data sets were computed using the phases and scaled structure factors (mF_P) with the figures of merit (m) obtained in the solvent-flattening procedure. The overall features of the two electron density maps are quite similar to each other, but the details in many portions of the maps were significantly different. A typical portion of the electron density map is shown in Fig. 2 in order to demonstrate how the combined scaling procedure works well in comparison with a simple isotropic scaling procedure. The data scaled by an isotropic scaling method gave an untraceable electron density map (Fig. 2*a*), whereas the electron density map (Fig. 2*b*) obtained from the data scaled by the combined scaling functions shows clearly several typical β-sheet electron densities, and most portions of the map are traceable without any special difficulty. This suggests that the scaling procedure strongly affects the quality of the electron density map

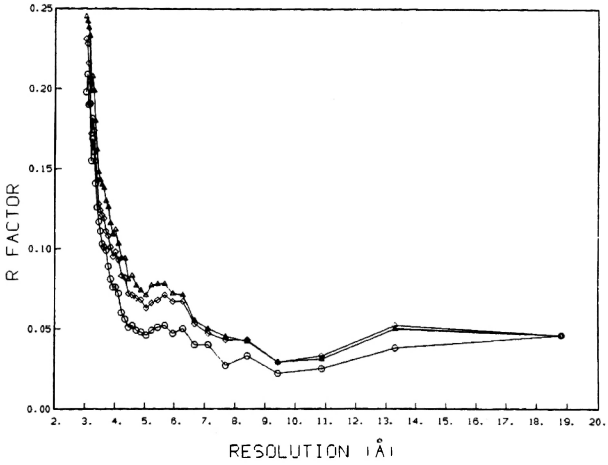

Fig. 1. Plots of resolution *vs* R factor for *S*-adenosylmethionine synthetase. 27 sets of native data were used for the following three tests. The R factor is defined as $R = \sum\sum|\langle I\rangle_h - I_i|/\sum\sum|\langle I\rangle_h|$. Test 1: a single scale factor, A, is applied on each data set (△). Test 2: an isotropic scaling function, $K(s) = A\exp(Bs^2)$, is applied on each data set (◇). Test 3: a combined scaling function, $K(x, s) = K(s)K(x)_s$, is applied on each data set (○).

phased by single isomorphous replacement data. The structure analysis results of *S*-adenosylmethionine synthetase will be published elsewhere.

When the multiple *s*-shell anisotropic scaling functions, $K(x)_s$, were tested without being combined with the isotropic scaling functions, $K(s)$, the calculation did not converge. Therefore, it is important to combine the $K(s)$ and $K(x)_s$ functions in order to carry out scaling and merging of a large number of data sets quickly and successfully.

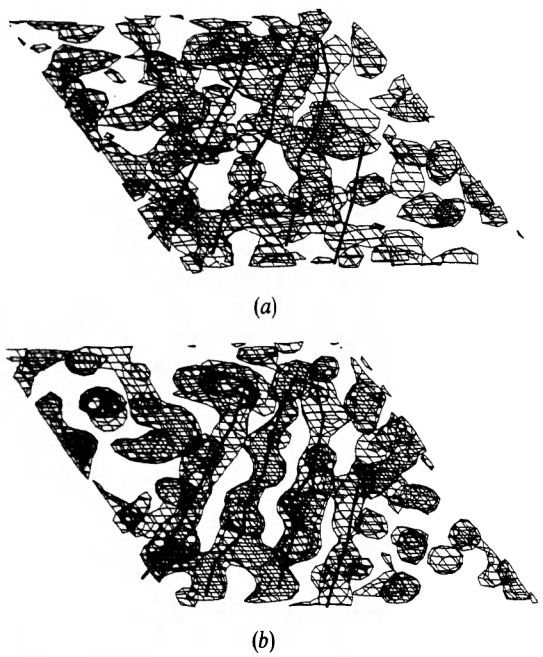

Fig. 2. Electron density maps calculated with the data phased by the solvent-flattening method using a single isomorphous replacement data set. The contour is at the 1.25σ level. The thick lines represent back-bone traces of the peptide chain in map (*b*). (*a*) Map computed with the data scaled and merged by using an isotropic scaling function $K(s) = A \exp(Bs^2)$. (*b*) Map computed with the data scaled and merged by using the combined scaling function, $K(x, s) = K(s)K(x)_s$.

Availability

The program is available on request from the author. Since the program is written in standard Fortran77, it should run on most computers without any modification of the code.

The author expresses his thanks to Professor Grover M. Everett, Dr David Vander Velde and Mr Luis Morales for critical reading of the manuscript. This research is supported by NIH grant GM37233 and by the Biomedical Support Grant from the NIH administered by the University of Kansas. It is also supported by the Wesley Foundation, Wichita, Kansas. The Wesley Foundation is an independent nonprofit organization whose mission is to improve the quality of health in Kansas.

References

DICKERSON, R. E. (1959). *Acta Cryst.* **23**, 610–611.
FOX, G. C. & HOLMES, K. C. (1966). *Acta Cryst.* **20**, 886–891.
HAMILTON, W. C., ROLLETT, J. S. & SPARKS, R. A. (1965). *Acta Cryst.* **18**, 129–130.
KRAUT, J. (1958). *Acta Cryst.* **11**, 895.
MATTHEWS, B. W. & CZERWINSKI, E. W. (1975). *Acta Cryst.* A**31**, 480–487.
ROLLETT, J. S. & SPARKS, R. A. (1960). *Acta Cryst.* **13**, 273–274.
ROSSMANN, M. G., LESLIE, A. G. W., ABDEL-MEGUID, S. S. & TSUKIHARA, T. (1979). *J. Appl. Cryst.* **12**, 570–581.
WANG, B.-C. (1985). *Methods Enzymol.* **115**, 90–112.