Acta Cryst. (1975). B31, 802

Least-Squares Determination of Idealized Molecular Dimensions and Orientations from Crystallographic Positional Coordinates

By Charles J. Fritchie Jr

Stern Chemical Laboratories, Tulane University, New Orleans, La. 70118, U.S.A.

(*Received* 27 *June* 1974; *accepted* 25 *October* 1974)

An iterative least-squares procedure is described which can be used to optimize the fit of an idealized molecule of any point-group symmetry simultaneously to one or more data sets consisting of experimental positional coordinates. Least-squares parameters include up to three rotational and three translational parameters for each data set, and $3n-6$ model coordinates for a model of n atoms. Variation of the model point-group symmetry permits use of the statistical F-test in discriminating among possibilities and reduces the influence of uncertainties in absolute data variances on the final choice. The procedure is used to derive a best model for the lumiflavin molecule from two data sets and to confirm the point-group symmetries and derive best models for several metal cluster compounds, including the non-heme ferroprotein model complex $[Fe_4S_4(SCH_2C_6H_5)_4]^{2-}$.

Introduction

A procedure for fitting a partially constrained molecular model to experimental atomic positional coordinates would clearly have a number of uses in crystallography and those fields requiring derived structural information such as quantum chemistry. Some uses have been described by McLachlan (1972) and Nyburg (1974), who report least-squares procedures for the matching of a completely rigid model to an experimental data set (or, equivalently, of two data sets to one another). The Diamond (1965) procedure for fitting linked rigid groups in macromolecular chemistry is well known. Dollase (1974) has recently described a general method for fitting a variable model when the constraints may be described by isotropic or anisotropic dilation of atomic coordinates in a single rigid framework.

The method described below is equivalent in result to Dollase's procedure, except that it permits optimization of a symmetry-constrained model to several data sets simultaneously and is thus slightly more powerful. It locates the least-squares fit to one or more data sets of a model whose point-group symmetry and, where required, whose orientation is fixed but which is otherwise completely flexible. The Fortran program, *MATCH*, which performed the calculations described, is rather flexible in that any set may contain extra atoms or omit some model atoms. It calculates root-mean-square deviations between the model and each data set; the use of these and related quantities in identifying approximate point-group symmetry is described.

Algorithm

The function D to be minimized, the weighted sum of squares of deviations between model and data parameters in an orthonormal model coordinate system is

$$D=\sum_{jk} w_{jk}(\mathrm{V}_j\mathbf{m}_{jk}+\mathbf{W}_j-\mathrm{S}_{bm}\mathbf{r}_b)^2, \qquad (1)$$

where j and k denote respectively the data set and data atom within that set, w_{jk} is the appropriate (possibly anisotropic) weight, V_j and $\mathbf{W}_j$ are a rotation matrix and a translation vector which transform the jth data set to the model, the vector $\mathbf{m}_{jk}$ denotes the coordinates of the kth data atom in a convenient orthonormal reference frame related to the crystallographic unit cell, S_{bm} is a fixed symmetry matrix which generates the kth model atom from an appropriate basis atom, and $\mathbf{r}_b$ gives the parameters of the basis atom in the orthonormal basis-model coordinate system. For reasons discussed below, an isotropic weight is used for each atom.

Symmetry matrices are specified by the user. For a model of C_1 symmetry, each model atom is related to a separate basis atom and each symmetry matrix is a 3×3 unit matrix. For a planar molecule of C_s symmetry, there is again a basis atom for every model atom, but the matrices are all $\begin{pmatrix}100\\010\\000\end{pmatrix}$.

For a cube, a single basis atom and matrices such as $\begin{pmatrix}100\\100\\100\end{pmatrix}$, $\begin{pmatrix}-100\\100\\100\end{pmatrix}$, *etc.*, are required (where the first parameter of the basis atom serves as the single variable parameter).

The transformation matrix V_j clearly introduces a nonlinearity into the least-squares normal equations, however it is constructed. *MATCH* chooses one of a pair of product matrices corresponding to rotations κ, μ and ν about specified axes depending on the relative orientation of model and data set, determines initial values of κ, μ, and ν from a trial orientation, and proceeds to solve the angular normal equations by the linearized Taylor series approximation commonly used in structure-factor refinement (Hughes, 1941). Because of the iterative nature of this process, the basis parameters are held fixed at initial values until orientational convergence is reached (four–six cycles to a tolerance of 10^{-6} radians in reasonable cases), then the linear normal equations constructed from the components of the various r_b's are solved, and the whole process is repeated until shifts in all basis parameters fall below a tolerance which is typically 10^{-4} Å. Seldom are more than three or four cycles of basis refinement needed.

This procedure seems to be less efficient than ones based on alternative formulations of the rotation matrix such as those given by Nyburg (1974) and Dollase (1974), but the total computation time is negligible in any case. The range of convergence seemed generally to be quite large. Shifts of over 30° from the trial orientation have been observed. As Nyburg (1974) warns, very slow convergence may reveal inconsistent model and data chiralities.

Constraints

Allowance has been made in the program for κ, μ, ν and the components of **W** to remain fixed. In most cases this procedure permits one to constrain data sets in such a way that symmetry elements in a crystallographically determined data set coincide with their counterparts in an appropriately oriented model. One or a pair of vectors and possibly some components of r_b in models of some symmetries must be constrained during model refinement to prevent 'drift' of the model. For most rapid convergence, constraints of this sort are imposed by unconstrained adjustment of the model, followed by a 'nudging' of the model to its constrained orientation and position in each cycle where necessary.*

Weights

The most appropriate weights would be those which permitted a variable component, $w_r = 1/\sigma_r^2$, to be assigned to each coordinate r, s, or t of each data atom. However, the information needed to perform this calculation is lost in publication of crystallographic papers, which rarely include covariance matrices for each atom. An estimate of the correlation of x, y, and z parameters in nonorthogonal crystal lattices is available (Templeton, 1959), but only for isotropic variance. In general, positional standard deviations derived from diffraction experiments are nearly isotropic for atoms which do not lie on special positions. *MATCH* uses individual isotropic weights $1/\sigma^2$ for each data atom, where $\sigma^2 = \frac{1}{3}(a^2\sigma_x^2 + b^2\sigma_y^2 + c^2\sigma_z^2)$.

Significance tests

The ratio $(\sum_i w\Delta D_{i1}^2 - \sum_i w\Delta D_{i0}^2)/\sum_i w\Delta D_{i0}^2$ for a given refinement 0 and a more constrained refinement 1 can be tested against $[g/(o-v)]\ F_{g,o-v,\alpha}$ where $\Delta D_i = D_{io} - D_{im}$ [equation (1)], g is the number of additional constraints, o is the number of observations, v the number of parameters varied in refinement 0, and $F_{g,o-v,\alpha}$ is the rejection point for the F distribution at significance level α (Hamilton, 1964, pp. 139, 208). Because the F test is a ratio, it minimizes the effect of an absolute scale error in the weights w_i. Such errors are common in crystallographic investigations, and great judgment is required in placing reliance on results which are marginally significant by the more usual comparison of.discrepancies from a mean with 2σ or 3σ.

Interpretation of results

Interpretation of a model obtained by this procedure as being representative of the free molecule assumes the lack of systematic error in the overall experimental molecular configuration. One common problem especially important in this regard is the apparent shrinkage of bonds due to anharmonic motion of the type commonly described as 'riding' in which the mean path of a given atom is a curve. Molecular parameters derived from uncorrected atomic positions will of course reflect this error. In the cases described below, this error is minimized by the large size of the molecules, which results in most atoms having nearly linear paths of mean vibration.

As noted by Dollase, bond lengths and angles obtained from the least-squares model are not identical with those obtained by averaging values of individual bonds or angles which would be equivalent under the deduced symmetry. The differences are generally small, however.

Examples

Lumiflavin

This compound, $C_{13}H_{12}N_4O_2$, is an aromatic heterocycle and is one of the simpler molecules containing the redox-active isoalloxazine component of flavocoenzymes (Wang & Fritchie, 1973). It is nearly but not exactly planar in most crystals and has been widely studied. It thus serves as an excellent example of a molecule of low symmetry whose coordinates are to be optimized for purposes of quantum mechanical calculations. The set of r, s, and t coordinates in Table 1 is the result of simultaneously optimizing a planar model and two accurate data sets, one of which carries an extra atom. Root-mean-square deviations generally average about 0·01 Å in r and s.

Although not precisely illustrated in this example, it should be obvious that the process of optimizing a planar model to *one* data set is equivalent to a least-squares plane calculation for the data set.

$[Fe_4S_4(SCH_2C_6H_5)_4]^{2-}$

Averill, Herskovitz, Holm & Ibers (1973) have recently reported the crystal structure of this ion as the tetraethylammonium salt. It is the best structural

* Separation of orientational and basis refinement insures nonsingularity even in the absence of constraints in the basis matrix.

Table 1. *Least-squares lumiflavin molecule*

All *t* coordinates are zero. * These parameters fixed the origin and orientation. Uncertainties are approximately 0·01 Å.

Atom	*r*	*s*	Atom	*r*	*s*
N(1)	2·298 Å	0* Å	C(9)	−2·434 Å	0·006 Å
C(2)	3·480	−0·685	C(9a)	−1·211	−0·667
N(3)	3·488	−2·094	N(10)	0*	0*
C(4)	2·373	−2·873	C(10a)	1·180	−0·680
C(4a)	1·092	−2·126	O(2)	4·542	−0·108
N(5)	−0·023	−2·796	O(4)	2·430	−4·089
C(5a)	−1·186	−2·075	C(7*m*)	−4·899	−2·875
C(6)	−2·413	−2·776	C(8*m*)	−4·914	0·073
C(7)	−3·614	−2·117	C(10)	0·019	1·471
C(8)	−3·616	−0·704			

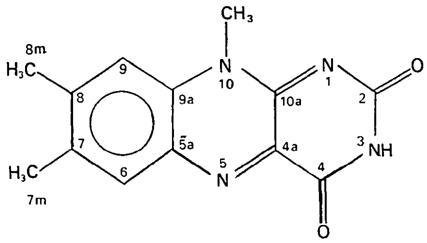
The least-squares transformation equations from crystallographic fractional coordinates **x** to model coordinates **r** are **r** = (XR)**x** + **T**, where

$$\mathrm{XR}=\begin{pmatrix} 5{\cdot}9325 & -6{\cdot}1327 & 4{\cdot}6035 \\ 4{\cdot}3917 & 7{\cdot}4116 & 4{\cdot}0518 \\ -13{\cdot}6400 & -0{\cdot}2810 & 6{\cdot}4015 \end{pmatrix} \text{ and } \mathbf{T}=\begin{pmatrix} -1{\cdot}299 \\ -5{\cdot}627 \\ 1{\cdot}941 \end{pmatrix}$$

for 3-methyllumiflavin (Norrestam & Stensland, 1972), and

$$\mathrm{XR}=\begin{pmatrix} 2{\cdot}8506 & 0{\cdot}7891 & -8{\cdot}4351 \\ -0{\cdot}5409 & -5{\cdot}8478 & -1{\cdot}9262 \\ -6{\cdot}5665 & 3{\cdot}2259 & -6{\cdot}5584 \end{pmatrix} \text{ and } \mathbf{T}=\begin{pmatrix} -1{\cdot}846 \\ -1{\cdot}638 \\ -1{\cdot}652 \end{pmatrix}$$

for 10-methylisoalloxazine (Wang & Fritchie, 1973).



model currently available for the cuboid iron–sulfur cluster in non-heme iron proteins. These authors have shown by techniques similar to those reported here, but restricted to models of D_{2d} symmetry, that D_{2d} rather than T_d is the proper choice of point-group symmetry. This ion is considered here to illustrate the use of the *F*-test and χ^2 tests (Hamilton, 1964) in choosing the model of best fit. The variable parameters are $r_{\rm Fe}$ and $r_{\rm S}$ for T_d symmetry, and $r'_{\rm Fe}, t'_{\rm Fe}, r'_{\rm S}$, and $t'_{\rm S}$ for D_{2d} symmetry. These are found to be 0·971, 1·275, 0·981, 0·950, 1·289 and 1·246 Å respectively.

Values of $\sum_i w_i(\Delta D_i)^2$ are 8762 and 498* for the two structures. We thus compare [(8762 − 498)/498] = 16·6 with $(\frac{1}{17})\,F_{1,17}$. At a confidence α of 0·005, $F_{1,17}$ is 10·38, and the D_{2d} model represents a clear improvement. The χ^2 test (Averill *et al.*, 1973) is performed in this case with 17 or 16 degrees of freedom, and we have rejection levels of 33·4 and 32·0 respectively for $\sum_i w_i(\Delta D_i)^2$, under the assumption that correct values have been used for $\sigma_{D_{io}}$. Both models are statistically rejectable, but as Averill *et al.* note, acceptance of the D_{2d} hypothesis implies that the average σ is underestimated by a factor of at least $(498/33{\cdot}4)^{1/2} \simeq 3{\cdot}9$. Underestimation by a factor of 2 is perhaps reasonable, and the remaining discrepancy can be considered possible distortion due to crystal forces.

Fe_4C_4 *clusters*

Two species are considered, neutral $[(C_5H_5)_4Fe_4(CO)_4]$ (Neuman, Toan & Dahl, 1972) and monocationic $[(C_5H_5)_4Fe_4(CO)_4]^+$ (Toan, Fehlhammer & Dahl, 1972). Following Averill *et al.*, we have obtained three models of D_{2d} symmetry, which yield $\sum_i w_i(\Delta D_i)^2$ of 221, 300, and 320 for the neutral cluster. The T_d model gives 341. The *F*-test quotient (341−221)/221 or 0·543 is to be compared with $(\frac{1}{17})F_{1,17,\cdot 01} = 0{\cdot}589$, and the improvement of the best D_{2d} model over the T_d model is found to be not significant at the 99% level. The T_d model is therefore considered the best description. For this model, $r_{\rm Fe} = 0{\cdot}891$ (5) Å, and $r_{\rm C} = 1{\cdot}08$ (1) Å, giving Fe–Fe = 2·521 Å and Fe–C = 1·99 Å, with uncertainties of approximately 0·010 Å and 0·011 Å respectively.

For the cation $[(C_5H_5)_4Fe_4(CO)_4]^+$, the three D_{2d} models yield $\sum_i w_i(\Delta D_i)^2 = 129$, 473, and 473, compared with 589 for the T_d model. Although the T_d error sum for the cation is not much larger than that in the neutral cluster, the *F*-test yields an improvement ratio of 3·57 which is highly significant at the 99% level, and greatly exceeds even $(\frac{1}{17})F_{1,17,005} = 0{\cdot}610$. The cation is thus considered to have symmetry D_{2d}.

The least-squares parameters for the best D_{2d} model are $r_{\rm Fe} = 0{\cdot}882$, $t_{\rm Fe} = 0{\cdot}866$, $r_{\rm C} = 1{\cdot}05$ and $t_{\rm C} = 1{\cdot}14$, with uncertainties of approximately 0·004 Å for Fe and 0·02 Å for C. The resulting bond lengths are Fe–Fe = 2·493 Å (2 bonds) and 2·472 Å (4 bonds); Fe–C = 1·96 Å (4 bonds) and 2·02 Å (2 bonds).

This work was supported by the National Institutes of Health through research grant GM 14189.

References

Averill, B. A., Herskovitz, T., Holm, R. H. & Ibers, J. A. (1973). *J. Amer. Chem. Soc.* **95**, 3523–3534.

Diamond, R. (1965). *Acta Cryst.* **19**, 774–789.

Dollase, W. A. (1974). *Acta Cryst.* A**30**, 513–517.

Hamilton, W. C. (1964). *Statistics in Physical Science.* New York: Ronald Press.

Hughes, E. W. (1941). *J. Amer. Chem. Soc.* **63**, 1737–1752.

McLachlan, A. D. (1972). *Acta Cryst.* A**28**, 656–657.

Neuman, M. A., Toan, T. & Dahl, L. F. (1972). *J. Amer. Chem. Soc.* **94**, 3383–3388.

Norrestam, R. & Stensland, B. (1972). *Acta Cryst.* B**28**, 440–447.

Nyburg, S. C. (1974). *Acta Cryst.* B**30**, 251–253.

Templeton, D. H. (1959). *Acta Cryst.* **12**, 771–773.

Toan, T., Fehlhammer, W. P. & Dahl, L. F. (1972). *J. Amer. Chem. Soc.* **94**, 3389–3397.

Wang, M. & Fritchie, C. J. Jr (1973). *Acta Cryst.* B**29**, 2040–2045.

* This value is slightly smaller than the quantity $21\sigma^2$ calculated by Averill *et al.* because a weighted centroid is used here.