Acta Cryst. (1987). A43, 426-428

Superimposing Several Sets of Atomic Coordinates

BY PAUL R. GERBER AND KLAUS MÜLLER

F. Hoffmann-La Roche & Co. Ltd, Central Research Units, CH-4002 Basel, Switzerland

(*Received* 17 *September* 1986; *accepted* 9 *December* 1986)

Abstract

A procedure is described which determines the best rotations to superimpose M rigid n-point objects, such that the weighted sum of mutual squared deviations is minimized. Apart from providing an easy and rigorous way for the least-squares superposition of any number of similar structures, this procedure can also be used to obtain a set of mean atomic positions for a substructural fragment that is contained in different structures with slight deviations from its mean geometry or to symmetrize distorted structures.

1. Introduction

The problem of optimally superimposing two given sets of atomic coordinates $(\mathbf{x}_l)$ and $(\mathbf{y}_l)$ such that the weighted quadratic deviation of the rotated set $(O\mathbf{x}_l)$ from $(\mathbf{y}_l)$

$$E = \tfrac{1}{2}\sum_l w_l(O\mathbf{x}_l - \mathbf{y}_l)^2 \qquad (1)$$

is minimized (McLachlan, 1972, 1982) can be solved analytically by its reduction to the diagonalization of a 3×3 symmetric matrix (Kabsch, 1976).

However, there are often situations, particularly in computer-aided drug design, where more than just two molecules need to be superimposed. In general, such tasks have been handled by sequences of pairwise superpositions without overall optimization. Since the number of pairs grows quadratically with the number of molecules, such a procedure may be quite tedious without leading to a truly optimized result. Furthermore, when matching all molecules onto a single target molecule, the quality of the result may depend strongly on the arbitrary choice of the target.

We outline a procedure for the simultaneous optimization of the superposition of M rigid molecules given by their coordinate sets $(\mathbf{x}_l^m)$, $m = 1,\ldots,M$. The criterion for an optimal match is a minimal value of the sum

$$E = \tfrac{1}{2}\sum_{n<m}^{M} v^{mn}\sum_l w_l(O^m\mathbf{x}_l^m - O^n\mathbf{x}_l^n)^2, \qquad (2)$$

where v^{mn} is a weight matrix that specifically weights the matching of molecule m with molecule n. In the following section we set v^{mn} equal to unity to simplify the notation. However, no additional difficulties arise for non-equal v^{mn}. The modification of the formulae is straightforward.

For the case of two molecules ($M = 2$), (2) reduces to (1). For the sake of simplifying the notation, we assume that all sets $(\mathbf{x}_l^m)$ are centered at the origin, *i.e.*

$$\sum_l w_l\mathbf{x}_l^m = \mathbf{0}, \quad m = 1,\ldots,M. \qquad (3)$$

2. Solution by linearization

In the first step we perform $M-1$ pairwise matches onto a single target molecule. This molecule is selected from the whole set by its largest deviation from linearity. To obtain a measure for this deviation, we start from the second-moment tensor

$$S^{mm} = \sum_l w_l\mathbf{x}_l^m\hat{\mathbf{x}}_l^m, \qquad (4)$$

where $\hat{\mathbf{x}}$ denotes the transpose of the column vector $\mathbf{x}$. The characteristic polynomial of this tensor describes the extensions of the molecule, *i.e.* a vanishing constant coefficient occurs for planar molecules, a vanishing linear coefficient for linear molecules and a vanishing quadratic coefficient for point objects. Thus the molecule with the largest linear coefficient is selected as target molecule m_t.

The first step then minimizes the sum (1) for each pair (m, m_t). After this step the sum (2) has in general not reached its minimum value.

The second step towards minimizing (2) consists of applying an additional orthogonal transformation O^m to each of the first $M-1$ molecules, keeping the last molecule M fixed in space with $OM = \mathbb{1}$ (identity matrix). Equation (2) then reads

$$E = \tfrac{1}{2}\sum_{n<m}\sum_l w_l(O^n\mathbf{x}_l^n - O^m\mathbf{x}_l^m)^2 = \text{minimum}. \qquad (5)$$

Since the traces of the second moments (4) remain unchanged under the orthogonal transformations O^m,

0108-7673/87/030426-03$01.50

we can rewrite (5) as

$$\sum_{n \neq m} \mathrm{Tr}\,(O^m S^{mn} \hat{O}^n) = \text{maximum}, \tag{6}$$

where we define the mixed moment tensors

$$S^{mn} \equiv \sum_l w_l \mathbf{x}_l^m \hat{\mathbf{x}}_l^n \tag{7}$$

in accordance with (4).

We proceed by writing the orthogonal matrices

$$O = \mathbb{1} + \sin\varphi \begin{pmatrix} 0 & -c_z & c_y \\ c_z & 0 & -c_x \\ -c_y & c_x & 0 \end{pmatrix} + (1-\cos\varphi) \begin{pmatrix} c_x^2-1 & c_x c_y & c_x c_z \\ c_x c_y & c_y^2-1 & c_y c_z \\ c_x c_2 & c_y c_z & c_z^2-1 \end{pmatrix} \tag{8}$$

in terms of the rotation angle φ and the axis of rotation given by its unit vector $\mathbf{c} = (c_x, c_y, c_z)$. Of course, all these quantities carry the label m of the molecule to be rotated.

The first step of pairwise matching generally ensures that the remaining rotations O^m involve only small rotation angles φ^m such that (6) may be solved successfully by linearization. The representation (8) facilitates this procedure. The rotations are equivalently characterized by the vectors ε^m, the components of which are given by

$$\varepsilon_i^m = \varphi^m c_i^m, \quad i = x, y, z, \tag{9}$$

and are expected to be small compared to one.

Solution of the extremum problem (6) in this linearized form is now straightforward, if somewhat tedious. For a concise representation, we introduce the 3×3 matrices J_i, K_i and U_i defined by

$$(J_i)_{jk} = \begin{cases} -\mathrm{sig}\,(P), & \text{if } (i,j,k) \text{ is a permutation } P \text{ of } (x,y,t) \\ 0, & \text{otherwise,} \end{cases}$$

$$(K_i)_{j,k} = |(J_i)_{jk}|, \tag{10}$$

$$(U_i)_{jk} = \begin{cases} 1, & \text{for } i=j=k, \\ 0, & \text{otherwise.} \end{cases}$$

The orthogonal transformations then read, to order $O(\varepsilon^2)$,

$$O^m = \mathbb{1} + \sum_k \varepsilon_k^m J_k + \tfrac{1}{2}\sum_k \varepsilon_{k'}^m \varepsilon_{k''}^m K_k - \tfrac{1}{2}\sum_k \varepsilon_k^{m2}(\mathbb{1} - U_k) + O(\varepsilon^3), \tag{11}$$

where k, k' and k'' are all different.

Insertion of (11) into (6) and differentiation with respect to ε_k^m yields the set of linear normal equations

$$\sum_n \{(\varepsilon_k^m - \varepsilon_k^n)\,\mathrm{Tr}\,[(\mathbb{1}-U_k)S^{mn}] - \tfrac{1}{2}\sum_{k'}(\varepsilon_{k'}^m - \varepsilon_{k'}^n)\,\mathrm{Tr}\,(K_{k''}S^{mn})\} = \sum_n \mathrm{Tr}\,(J_k S^{mn}), \quad m = 1, \ldots, M-1,\ k = x, y, z, \tag{12}$$

$$\varepsilon_j^M \equiv 0.$$

Here again k, k' and k'' are all different.

Solution of the set of linear equations (12) yields the $M-1$ triples $(\varepsilon_x^m, \varepsilon_y^m, \varepsilon_z^m)$, $m = 1, \ldots, M-1$, which determine the rotation matrices O^m to order ε^2 through (11). To ensure that the rotation matrices remain strictly orthogonal, they are taken in the form (8), where the required parameter values φ_m and c^m are obtained from ε^m through

$$\varphi^m = (\varepsilon_x^{m2} + \varepsilon_y^{m2} + \varepsilon_z^{m2})^{1/2}, \quad c_i^m = \varepsilon_i^m/\varphi^m, \quad i = x, y, z. \tag{13}$$

3. Computational aspects

As in the pairwise match, the only step which increases with the size of the coordinate sets is the evaluation of the mixed moments. However, because there are $M(M-1)/2$ such matrices the computational effort for this step grows quadratically with the number of involved molecules. The remaining calculations are manipulations of 3×3 matrices and the solution of $3(M-1)$ linear equations (12).

Our procedure has been implemented as a Fortran 77 subroutine package on a VAX-11/780. It typically takes about 0·4, 0·5 and 1·3 s of CPU time for the superposition of ten molecules using respectively 8, 16 and 150 reference atoms in each molecule. On the other hand, superposition of 20 molecules with eight atoms each took 1·6 s.

Furthermore, in all practical applications one single iteration has proved sufficient to achieve alignment accuracy below one tenth of a degree.

The choice of the most non-linear molecule as a target for the first step generally leads to a good initial alignment, resulting in subsequent rotations by only a few degrees. The situation becomes considerably less favorable for ill chosen (*e.g.* nearly linear) target molecules.

Of course, it is possible to construct pathological examples for which the first step would generate orientations that still necessitate large rotations. In such cases linearization of the problem may not be justified. An example of this sort would be a set of four-point objects which contains two tetrahedra that differ only by interchange of the labels of two vertices. However, while such cases may still be of mathematical interest, they are of little practical relevance.

4. Finding averaged coordinates

A frequent problem in computer-assisted molecular modelling is that of obtaining a weighted average structure for a substructural fragment contained in a number of different molecular structures determined by X-ray diffraction techniques.

There are several ways to obtain such averaged fragment structures. A particularly convenient one is to average atomic positions in a set of appropriately superimposed fragment structures. This section demonstrates that our superposition procedure provides a basis for obtaining weighted average structures.

We require that a weighted average structure ($\mathbf{x}_l^0$) obtained from M molecules ($\mathbf{x}_l^m$) be an average taken after suitable reorientation of the molecules. The reoriented molecules have the coordinates

$$\mathbf{y}_l^m = O^m \mathbf{x}_l^m. \tag{14}$$

The weighted average structure is given by

$$\mathbf{x}_l^0 = \sum_m u^m \mathbf{y}_l^m, \quad \sum_m u^m = 1. \tag{15}$$

The requirement that the average structure be optimal can be formulated as

$$\sum_l w_l \sum_m u^m (\mathbf{y}_l^m - \mathbf{x}_l^0)^2 = \sum_l w_l \sum_m u^m (\mathbf{y}_l^{m2} - \mathbf{x}_l^{02}) = \text{minimum}, \tag{16}$$

where the parameters to be varied are the orientation matrices O^m. Again, $\mathbf{y}_l^{m2}$ is unchanged if the rotation matrix O^m is varied; hence (16) is equivalent to

$$\sum_l w_l \mathbf{x}_l^{02} = \text{maximum}. \tag{17}$$

Inserting the definition (15) of $\mathbf{x}_l^0$, we get

$$\sum_{m,n} u^m u^n \sum_l w_l \mathbf{y}_l^m \mathbf{y}_l^n = \text{maximum}. \tag{18}$$

The term with $m = n$ is again invariant. Thus, after inserting (14), we may equivalently write

$$\sum_{m \neq n}^{M} u^m u^n \operatorname{Tr}(O^m S^{mn} \hat{O}^n) = \text{maximum}. \tag{19}$$

This condition is exactly equivalent to (6) if we take instead of constant pairing weights v^{mn} [see (2)] the values

$$v^{mn} = u^m u^n. \tag{20}$$

Therefore, in order to obtain an optimized structure as a weighted average of M given stuctures, one has to perform the superposition of the M molecules as described in § 2, taking as pairing weights the values (20). The average coordinates ($\mathbf{x}_l^0$) are then obtained through (15) from the matched coordinates.

This procedure can also be applied to symmetrize a structure which is expected to show a certain symmetry, but for which the actual coordinates deviate slightly from the required symmetry. The symmetrized structure is easily obtained as the average structure of all possible symmetry-related orientations superimposed with equal weight factors ($v^{mn} = \text{constant}$).

References

KABSCH, W. (1976). *Acta Cryst.* A**32**, 922-923.
MCLACHLAN, A. D. (1972). *Acta Cryst.* A**28**, 656-657.
MCLACHLAN, A. D. (1982). *Acta Cryst.* A**38**, 871-873.

SHORT COMMUNICATIONS

Contributions intended for publication under this heading should be expressly so marked; they should not exceed about 1000 *words; they should be forwarded in the usual way to the appropriate Co-editor; they will be published as speedily as possible.*

Acta Cryst. (1987). A**43**, 428-430

Sayre's equation is a Chernov bound to maximum entropy. By ROBERT W. HARRISON, *Center for Chemical Physics, National Bureau of Standards, Gaithersburg, MD* 20899, *USA*

(*Received* 26 *June* 1986; *accepted* 30 *September* 1986)

Abstract

Sayre's equation is fundamental to a large part of classical direct methods. In this paper, it is shown that this equation can be derived *via* an integral bound to the entropy integral. While positivity is implicit in this derivation, atomicity is not used.

Introduction

Sayre's equation and similar triplet-based forms have formed the basis for the highly successful direct methods used in small-molecule crystallography (Sayre, 1952; Karle & Hauptmann, 1950). The maximum-entropy method has the potential to extend these methods to larger and more