An easily fixed error in the Nyburg algorithm for discovering the best fit between molecules

Daniel P. Dolata and James Arnold

Department of Chemistry, University of Arizona, Tucson, AZ

The extremely popular Nyburg algorithm for discovering the best fit between two molecules or fragments has an error that can give a false best fit under some circumstances. This error is described, and a simple fix is provided. The original Nyburg program (BMFIT) is compared to Sippl's program of 1991.

Keywords: best fit, Nyburg

The ability to find the "best fit" between selected atoms of two molecules or fragments has many uses; among them are discovery of duplicate structures, comparison of different molecules, and aligning of molecular templates for use in model building. There are many methods that find the best fit based on atom coordinates, internal coordinates, intramolecular distances, or other approaches.[1] In 1974, Nyburg[2] described the algorithm used in his *Best Molecular Fit program* (BMFIT). This algorithm has been used by numerous groups and in numerous programs, as witnessed by the fact that there are well over 150 citations to this paper in *Science Citation Index*. Another indication of the strength of this method is the fact that Sippl published an extension of the same algorithm in 1991[3] (although he failed to reference Nyburg's original paper).

We have been using a number of different academic and commercial programs based on variations of this algorithm for several years, with varying degrees of success. When the fragments are in rough alignment before comparison, the algorithm works very well. Even in cases where the initial alignment is poor, the algorithm generally works. However, if the fragments are close to being rotated 180° with regard to each other, the algorithm often discovers a false best fit. This fact might have been discovered by Nyburg himself, since Kabasch[4] gives a reference to a private communication from Nyburg in which he discusses the problem of a false minimum. However, no details were given in Kabasch's paper about the problem or any possible fix to Nyburg's algorithm.

Address reprint requests to Dr. D.P. Dolata at the Department of Chemistry, University of Arizona, Tucson, AZ 85721, USA.
Received 28 April 1992; revised 1 July 1992; accepted 7 July 1992

When utilizing graphical programs, the false fit is not much of a problem. Generally, the user roughly aligns the fragments to act as a visual guide while selecting the atoms. This rough alignment obviates the problem we are about to describe. In those cases where the initial alignment still gives an occasional false minimum, this is easily seen and then corrected. However, without a visual check these false solutions can be a problem. For example, we have been using a Monte Carlo method to search the conformational space of large rings. After perturbation and minimization, we used the Nyburg algorithm (code which was a direct descendant of BMFIT) to find duplicate structures. When we rechecked our results with a duplicate discovery algorithm based on internal coordinates, we found that the program had failed to detect duplicates in as much as 10% of the cases. Following this and other difficulties in using programs based on the Nyburg algorithm, and considering how important this algorithm is, we decided to investigate the problem and search for a general solution to the problem.

Nyburg's algorithm begins by calculating the weighted centroids of the atoms, and translating them to the origin. Sippl's algorithm also begins with this step, although he omits the possibility of using weighted values. Then, according to Nyburg's paper: "One of the molecules is rotated to give a reasonably good first fit. This approximate fit makes use of only three points associated with each of the two molecules: the (weighted) centroids and the positions of the first two atoms listed for matching. It causes the normals to the planes formed by these two sets of three points to be parallel and for the lines joining the centroids (already coincident) to the first atoms to be collinear."

In most cases, this will bring the molecules into good alignment. However, if the molecules are essentially inverted, as shown in Figure 1, then the alternative alignment shown in Figure 1 also satisfies the criteria. The lines connecting the centroids to atoms A and A' are collinear, and the normals are parallel. The exact means used to fulfil Nyburg's criterion of first alignment can lead to this problem. Our copy of BMFIT[5] aligned the first atoms to the x-axis by calculating the arctangent of the rotation needed to bring the atom to lie on the x-axis. This technique did not distinguish between atoms lying in the first and third quadrants, or between the atoms lying in the second and fourth. Thus, these atoms could be aligned along the positive or negative x-axis.

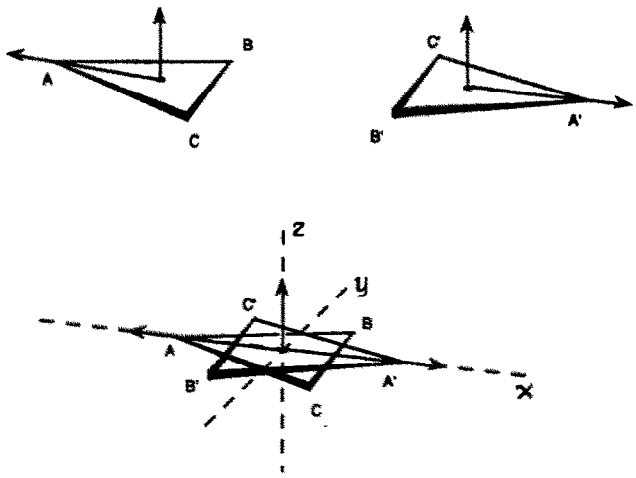

Figure 1. False alignment that satisfies Nyburg's conditions.

Once the initial alignment has been performed, both Nyburg and Sippl find the best fit by successive rotations around the three axes until a termination condition is met. Nyburg gave an example of the rotation matrix around the z-axis as

$$\mathbf{S}_3 = \begin{matrix} \cos\omega_3 & -\sin\omega_3 \\ \sin\omega_3 & \cos\omega_3 \end{matrix} \tag{1}$$

and

$$\omega_3 = \tan^{-1}\left\{\frac{\sum_j^n w_j(Y_{1j}X_{2j} - Y_{2j}X_{1j})}{\sum_j w_j(X_{1j}X_{2j} + Y_{1j}Y_{2j})}\right\}. \tag{2}$$

Similar formulas are used to calculate the best rotations around the x- and y-axes. Sippl's Equation (26c) for rotation around the z-axis is the 3 × 3 form of Nyburg's matrix above. His angle of rotation γ_m is found from his Equation (25c), shown here:

$$\gamma_m = \arctan\left[\frac{(u_{21} - u_{12})}{(u_{11} + u_{22})}\right] \tag{3}$$

His Equation (22) defines u as

$$u_{kl} = \sum_i x_{ik}y_{il} \tag{4}$$

This is essentially the same as Nyburg's equation, which can be seen in the following fashion. Sippl uses X and Y for the names of the two molecules to be matched, and the numbers 1, 2, and 3 as labels for the x-, y- and z-axes. To show that Sippl's formula is the same as Nyburg's, we substitute A for X, and B for Y, and x for 1, y for 2, and z for 3 in Formula (4). When these substitutions are made in Equation (3), this gives us the somewhat odd form

$$\gamma_m = \arctan\left\{\frac{\sum_i(A_{iy}B_{ix} - A_{ix}B_{iy})}{\sum_i(A_{ix}B_{ix} + A_{iy}B_{iy})}\right\} \tag{5}$$

If we rewrite this with $A_{ix} = X_{1i}$, $A_{iy} = Y_{1i}$, $B_{ix} = X_{2i}$, and $B_{iy} = Y_{2i}$, then it can be seen that Equation (5) is equivalent to Equation (2), without Nyburg's weighting scheme.

Having established the identity of the rotational matrix portions of Nyburg's and Sippl's algorithm, we would like to demonstrate symbolically how this algorithm can give false best fits. Figure 2 shows two identical sets of three atoms which are to be matched, but which are rotated by 180° with regard to each other. The initial alignment scheme aligns them as shown. Thus, for each point A, B, and C, the corresponding point in the other set has the opposite values for the Cartesian coordinates. This is shown in Table 1.

If we expand the numerator (*num*) of Equation (2), we obtain the equation

$$num = Y_{11}*X_{21} - Y_{21}*X_{11} + Y_{12}*X_{22} - Y_{22}*X_{12} + Y_{13}*X_{23} - Y_{23}*X_{13} \tag{6}$$

Substituting the data from Table 1 we obtain:

$$num = (Ay* - Ax) - (-Ay*Ax) + (By*Bx) - (-By* - Bx) + (-Cy*Cx) - (Cy* - Cx) \tag{7}$$

which can be reduced to

$$num = 0 \tag{8}$$

for all values of A_x, A_y, B_x, B_y, C_x, and C_y. Plugging this into Equation (2), we find

$$\omega_3 = \tan^{-1}\left\{0/\sum_j w_j(X_{1j}X_{2j} + Y_{1j}Y_{2j})\right\} = \tan^{-1}\{0\} \tag{9}$$

There are two roots for $\tan^{-1}(0)$; one at 0° (the false fit), and another at 180° (the right answer). The version of BMFIT that we obtained[5] uses the FORTRAN ATAN(X) function. ATAN(X) returns the first answer, and thus this program will not rotate such a rotationally disposed pair of identical molecules. The second molecule need not be identical to the first to obtain such a false fit. We have seen the same behavior with molecules that are similar but not identical to each other. We have not analyzed precisely how different the molecules must be to avoid this false minima, but molecules that show a rms fit of 0.15 Å over 4 atoms when correctly aligned have been known to exhibit this problem when the initial alignment is off by an angle between 180° and 270°. However, Sippl

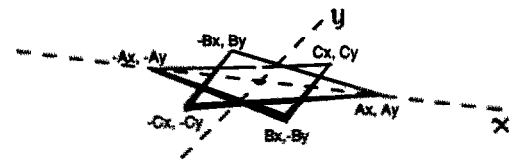

Figure 2. Identical sets of atoms rotated 180°.

Table 1. Symbolic coordinates for the superimposition of the atom sets shown in Figure 2

Molecule 1					Molecule 2		
X_{11}	Ax	Y_{11}	Ay	X_{21}	$-Ax$	Y_{21}	$-Ay$
X_{12}	$-Bx$	Y_{12}	By	X_{22}	Bx	Y_{22}	$-By$
X_{13}	$-Cx$	Y_{13}	$-Cy$	X_{23}	Cx	Y_{23}	Cy

uses the ATAN2(Y,X) routine, which examines both the numerator and the denominator even in those cases where *num* = 0. In our example, ATAN2(0, positive) = 0, and ATAN2(0, negative) = 180, which gives the right fit.

We have been able to overcome these problems with two very simple fixes. In addition to replacing BMFIT's ATAN with ATAN2, we modify the initial fitting routine as follows. Instead of performing an initial rotation as shown in Figure 1, we move the centroids to the origin, and then align the *rays* from the centroid to the first atoms along the *positive x*-axis, and orient the normals to the planes along the *positive z*-axis. This assures that the two molecules fulfil Nyburg's original criteria, but also assures that they will be sitting in the same sense with regard to each other and the coordinate system. We checked this modified Nyburg algorithm for the detection of duplicate structures on several thousand structures obtained from our large ring studies. We compared these results against the results of our old BMFIT based program and the results from a newer program, which is based on internal coordinates. The modified algorithm found all of the duplicates that were found via the newer method, and was much faster than the internal coordinate approach.

The false best fit is not a common problem unless one is performing duplicate searches using the Nyburg algorithm. But considering the number of programs that are based on Nyburg's original algorithm (some of which utilize code originally obtained from BMFIT), and the fact that we have seen this false best fit problem in more than one academic and commercial program,[6] we feel that it is important that the community be aware of the improvements inherent in Sippl's algorithm or our improvement to Nyburg's algorithm.

REFERENCES

1 Bajaj, M. and Blundell, T. Evolution and the tertiary structure of proteins. *Annu. Rev. Biophys. Bioeng.* 1984, **13**, 453–455

2 Nyburg, S.C. Some uses of a best molecular fit routine. *Acta. Crystallogr.* 1974, **B30**, 251

3 Sippl, M.J. and Stegbuchner, H. Superposition of Three-Dimensional Objects: A Fast and Numerically Stable Algorithm for the Calculation of the Matrix of Optimal Rotation. *Comp. Chem.* 1991, **15**(1), 73–78

4 Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr.* 1978, **A34**, 827

5 Obtained from the Computer Graphics Laboratory, OK2, Lunds University, Lund, Sweden, Dr. R.E. Carter and Dr. T. Liljefors

6 A reviewer suggested that we make a list of the better known programs that exhibit this problem. We are loath to do so for several reasons: We cannot test all of the programs, and so our list would be far from comprehensive and might unfairly single out those that we can test while failing to indict those we cannot test. In addition, the list membership should change with time as authors fix the bug in their programs. The reader can test any program by making two copies of a molecule, rotating one copy around the *z*-axis by an angle between 180° and 270°, and performing a best fit on corresponding atoms.