# Analytical Algorithm for Molecular Modeling

Attilio Immirzi†

Dipartimento di Chimica, Università di Salerno, Fisciano (SA) I-84084, Italy

Received June 26, 2007

A molecular modeling procedure, based on internal coordinates and strictly analytical even in the most intricated cases, is described. Internal coordinates, always nonredundant, become mutually independent and can be varied without constraints. Structural refinement from diffraction data (Least-square method, LS) can be done using the classical Gauss−Newton approach and avoiding Lagrange multipliers. A comparative test done using published data has shown that while the new method gives rise to a structural refinement in perfect agreement with the known structure, the traditional methods (**z**-matrix and constraints based) does not work.

## INTRODUCTION

Molecular modeling is generally performed using the so-called "internal coordinates" (i-c): bond lengths (b-l), bond angles (b-a), and torsion angles (t-a), rather than Cartesian coordinates (c-c). Use of Eyring's[1] algorithm (also known as the **z**-matrix method) to translate i-c into c-c is almost universal. Eyring's method is used in spectroscopy,[2] molecular dynamics,[3,4] and diffraction analysis.[5,6] The benefit of using the i-c is that the rules of stereochemistry restrict a number of i-c to definite values, thereby decreasing the number of effective variables. In the analysis of crystal structure by diffraction, b-l are substantially known a-priori, at least at a crude level, b-a span narrow ranges, and some t-a may also be known (see, e.g., conformations about double bonds). Imposing such restrictions in c-c is possible but tortuous; in i-c restrictions simply implies keeping a number of coordinates constant. Going from c-c to i-c reduces the unknowns to about one-third, at a coarse level of analysis (with b-l and b-a fixed) and to about two-thirds at a more refined level (with only b-l fixed). Reducing the number of structural variables in the trial-and-error stage of the analysis certainly increases the chances of success; in the refinement stage it decreases the size of the standard errors.

The important question remains: how many i-c have to be used in building a model? Since a molecule (of arbitary orientation) is described by $3N - 6$ c-c, the same number of i-c should be used (including the ones kept constant!). If more i-c are used, coordinates are said to be *redundant*.[2,7] Modeling molecules with redundant coordinates leads to a number of complications, at least during structural analysis. A modeling strategy based strictly on $3N - 6$ i-c is definitely preferable. A second question is: what should be used as i-c: b-l, b-a, or t-a? There are of course many solutions since the total number of b-l, b-a, and t-a may exceed $3N - 6$. If one defines the i-c in such a way that an analytical relationship for translating i-c into c-c exists, the problem is altogether straightforward.

Eyring's method satisfies the above requirements, giving rise to a strictly analytical process, by modeling both chemically linear structures and branched structures. However there are problems with cyclic molecules (rings, condensed rings, and molecules with internal bridges, etc.) and in polyhedrical structures. The main problem is that the modeling based strictly on the **z**-matrix method requires the use of a number of i-c that exceed $3N - 6$. As a consequence the i-c are *no longer independent among them* and molecular modeling is no longer an analytical procedure. In diffraction studies, structural refinement using the Least-square method (LS) makes the use of constrained variables (e.g., using the Lagrange method) necessary.

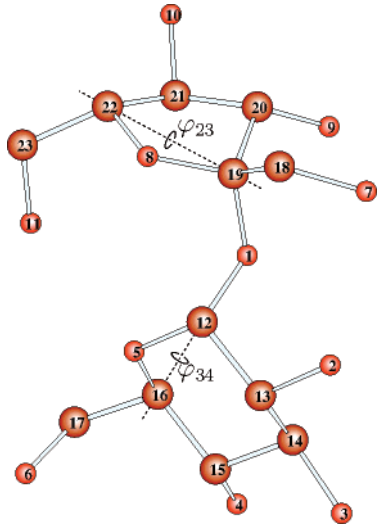



**Figure 1.** Sucrose. The orientation is arbitrary. Larger circles are C atoms; smaller circles, O atoms. H atoms are omitted. The bending angles $\varphi_{23}$ and $\varphi_{34}$, used for closing the rings, are indicated.

So, a new algorithm for handling the above cases is highly desirable. The solution described here is simple, and the results are altogether satisfactory.

## NEW BUILDING ALGORITHM

The above difficulties can be overcome by introducing a quite simple rule: always use $3N - 6$ i-c and include *all* b-l. A method to ensure that molecular building is always analytical involves introducing two new contructions, in

† E-mail: aimmirzi@unisa.it.

10.1021/ci700225x CCC: $37.00 
Published on Web 10/31/2007

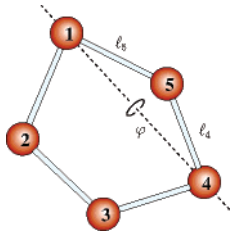
**Table 1.** Structural Refinement of Sucrose Using Internal Coordinates and the New Algorithm[a]

| i-c[b] | definition | this work | ref 11 |
|---|---|---|---|
| $\tau_1$ | C16−C15−C14 | 114.8(15) | 110.7 |
| $\tau_2$ | C15−C14−C13 | 107.8(13) | 107.2 |
| $\tau_3$ | C14−C13−C12 | 117.0(16) | 112.1 |
| $\tau_4$ | O5−C12−O1 | 109.9(15) | 110.1 |
| $\tau_5$ | C15−C14−O3 | 111.4(18) | 107.7 |
| $\tau_6$ | C14−C15−O4 | 110.5(17) | 112.5 |
| $\tau_7$ | C14−C13−O2 | 115.1(17) | 109.8 |
| $\tau_8$ | C15−C16−C17 | 110.2(18) | 112.2 |
| $\tau_9$ | C16−C17−O6 | 115.3(19) | 111.2 |
| $\tau_{10}$ | C12−O1−C19 | 112.0(16) | 113.8 |
| $\tau_{11}$ | O1−C19−C20 | 112.0(16) | 108.2 |
| $\tau_{12}$ | C19−C20−C21 | 101.3(14) | 102.2 |
| $\tau_{13}$ | C20−C21−C22 | 100.1(14) | 102.8 |
| $\tau_{14}$ | C19−C20−O9 | 113.3(16) | 115.8 |
| $\tau_{15}$ | C20−C21−O10 | 122.3(18) | 111.8 |
| $\tau_{16}$ | O8−C22−C23 | 109.3(19) | 110.0 |
| $\tau_{17}$ | C22−C23−O11 | 113.0(19) | 112.8 |
| $\tau_{18}$ | C20−C19−C18 | 113.9(16) | 114.7 |
| $\tau_{19}$ | C19−C18−O7 | 109.5(16) | 111.4 |
| $\vartheta_{21}$ | C16−C15−C14−C13 | −49.4(24) | −56.8 |
| $\vartheta_{22}$ | C15−C14−C13−C12 | 45.5(26) | 57.3 |
| $\varphi_{23}$ | C15−C16−C12−O5 | 126.1(23) | |
| $\vartheta_{24}$ | C16−C5−C12−O1 | −58.4(25) | −68.2 |
| $\vartheta_{25}$ | C16−C15−C14−O3 | −171.7(18) | −177.2 |
| $\vartheta_{26}$ | C13−C14−C15−O4 | −178.2(18) | −174.1 |
| $\vartheta_{27}$ | C15−C14−C13−O2 | 179.8(19) | 178.5 |
| $\vartheta_{28}$ | C14−C15−C16−C17 | 168.9(17) | 173.4 |
| $\vartheta_{29}$ | C15−C16−C17−O6 | −62.0(28) | −64.,1 |
| $\vartheta_{30}$ | C13−C12−O1−C19 | 133.6(16) | 130.2 |
| $\vartheta_{31}$ | C12−O1−C19−C20 | 159.1(15) | 160.2 |
| $\vartheta_{32}$ | O1−C19−C20−C21 | −84.4(17) | −87.6 |
| $\vartheta_{33}$ | C19−C20−C21−C22 | −41.5(16) | −35.2 |
| $\varphi_{34}$ | C21−C22−C19−C8 | 166.9(24) | |
| $\vartheta_{35}$ | O8−C19−C20−O9 | 154.7(18) | 158.1 |
| $\vartheta_{36}$ | C19−C20−C21−O10 | −155.2(18) | −155.6 |
| $\vartheta_{37}$ | C19−O8−C22−C23 | −132.7(18) | −132.9 |
| $\vartheta_{38}$ | O8−C22−C23−O11 | 64.3(27) | 70.5 |
| $\vartheta_{39}$ | C21−C20−C19−C18 | 151.1(15) | 148.8 |
| $\vartheta_{40}$ | C20−C19−C18−O7 | 75.9(22) | 72.1 |
| $R_x$ | | 65.4(2) | |
| $R_y$ | | −197.0(3) | |
| $R_z$ | | −54.8(2) | |
| $t_x$ | | 2.454(4) | |
| $t_z$ | | 2.104(5) | |

[a] The refinement uses 589 reflections ($\sin\theta/\lambda < 0.447$ Å$^{-1}$) and a unique isotropic thermal parameter ($B = 1.00$ Å$^{-2}$). H atoms were neglected. C−C b-l are kept fixed to 1.521 Å and C−O b-l to 1.427 Å (average value found by Hynes and Le Page). The 39 angular i-c, viz., 19 b-a $\tau$, 18 t-a $\vartheta$, and 2 out-of-plane bending angles $\varphi$, are adjusted by LS and converge to the values given below with standard errors in parentheses. Note that there are 23 atoms and 24 bonds and that 39 is just $3 \times 23 - 6 - 24$. The molecule is built up using the **z**-matrix constructions and two flap constructions (see text) for closing the two rings. There are in addition five adjusted variables: three overall rotations $R_x$, $R_y$, *and* $R_z$ about the principal axes and two overall translations $t_x$ and $t_z$ ($y$-translation is origin fixing), *applied to the molecular model referred to its inertial axes*. Atom numbering (see Figure 1) is different from the original one,[12] due to the program requirements. The final $R_1$ index obtained is 0.071 (compare with the one published 0.032), obtained with anisotropic thermal parameters and including H atoms. All correlation coefficients between i-c are less than 0.75. [b] Angular i-c (deg) obtained in this work compared with the published figures.

addition to Eyring's, one for closing rings (*flap*) and another for closing polyhedra (*cage*). In summary, while Eyring uses one b-l and two angles for each atom, flap uses two b-l and one angle and cage uses three b-l and zero angles. The number of variables is the same, but the quality is not, and this simple contrivance, making the number of i-c *always* $3N - 6$, renders the whole procedure strictly analytical, leads to regular LS matrices, and removes the necessity of using Lagrange multipliers.

**Closing Rings.** We consider the case of cylopentane. The bare $C_5$ ring has nine i-c of which five are b-l and four angles. Following Eyring one builds the molecule starting from three atoms (2 b-l + 1 b-a) and add the fourth atom by the **z**-matrix method (1 b-l + 2 angles) using in all three b-l and three angles). For completing the ring one must therefore use two b-l and one angle. It is evident that if one wishes to use *only* the **z**-matrix method (each step needs one b-l and two angles), the use of an extra angle is inevitable and one b-l remains unused. Using the **z**-matrix method necessarily implies *the lack of ring closure*. The five angles used following the Eyring's algorithm are *redundant* by one unit, and the procedure of ring building, with the prescripted values for all b-l, may be completed only imposing a *constraint*: viz., C1−C5 distance = the known b-l.




There is a simple way to overcome the problem, which avoids the use of constraints: to compute the last ring atom (C5 in the example) not as a function of *one* b-l and *two* angles (as **z**-matrix does), but as a function of *two* b-l and *one* angle. The former are the two b-l C5−C1 ($l_5$) and C5−C4 ($l_4$); the latter is neither a b-a nor a t-a but the out of plane bending angle $\varphi$ about the line C1−C4 (the use of bending angles was proposed by Goto and Osawa[8] in studying the conformations of cyclic molecules). Considering the triangle C1−C5−C4 with sides of known lengths ($l_5$, $l_6$ are i-c; C1−C4 is known since C1 and C4 have already been computed), it is easy to express the $x$, $y$, $z$ coordinates of C5 as a function of $l_5$, $l_4$, $\varphi$.

This works regardless of ring size and is also applicable in building multiring molecules (e.g., steroids), molecules with bridging atoms (e.g., norbornane, camphor), and spiro-compounds among others.

**Closing Polyhedra.** The flap procedure illustrated above overcomes many but not all the problems. Polyhedrical molecules, for instance, cannot be closed, avoiding redundance, even in the most simple case, e.g., cubane. The bare $C_8$ molecule has 18 i-c of which 12 are b-l so that only six angles must be used. One can proceed along the following route: (1) define the first three atoms C1, C2, and C3 using one angle; (2) close the four-atom ring (C4) with a flap construction (one angle); (3) use **z**-matrix construction to define the C5 atom (two angles); (4−5) close two four-atom rings C1−C2−C5−C6 and C2−C3−C6−C7 using again flap twice (two angles); (6) define the last C8 atom with the difficulty that *all six angles available have already been used*. Indeed computing C8 *does not need angles*, needing only the known bond lengths to C4, C5, and C7. There is of course only a solution, provided the polyhedron is convex. Several polyhedrical molecules have been built with a similar

procedure, all without redundancies, including the (hypothetical) dodecahedron-shaped hydrocarbon $C_{20}H_{20}$ and last the C60 fullerene.[9]

## STRUCTURAL REFINEMENT BY THE LEAST-SQUARES PROCEDURE

The novel algorithm presented in this paper is particularly useful when performing structural refinement using the LS method based on internal coordinates and diffraction data. When the model building becomes a strictly analytical procedure, and redundant coordinates are avoided, many problems can be resolved simply. The i-c based LS procedure can be programmed using the Gauss−Newton algorithm and normal matrix is nonsingular. No constraints are needed; standard errors for i-c, and correlation coefficients are correctly computed from the inverse normal matrix; convergence is regular. Of course the claimed analytic character applies to the model building, not to the entire refinement process.

TRY[13] is a novel program for structure analysis by diffraction methods, based on internal coordinates and model building done following the illustrated ideas. The program has numerous options (see ref 12) and allows sructural refinement *also* using constraints (they are mandatory in special cases, viz., linear polymers[10]). It is thus possible to follow, using the same program, both the novel route (use strictly $3N - 6$ molecular i-c) and the traditional route based on **z**-matrix only (use redundant i-c and appropriate constraints). To persuade skeptics that redundant i-c are dangerous, a test was performed using published single-crystal data to compare the two routes. The test is based on the X-ray diffraction data of sucrose for which excellent measurements have been published.[11] Sucrose was chosen for having two ring-closure problems (see Figure 1). With the new method one uses (at fixed b-l) 44 independent i-c and no constraints; with the traditional method (**z**-matrix only) one uses 46 i-c and 2 constraints. To be clear, this test was *not* aimed at finding a better structure for sucrose but only of demonstrating the convergence in the i-c space. The result was excellent, considering the drastic reduction of the number of parameters (from 206 to 44).

Starting from i-c values, which differ from the true ones by 10−20° (the initial *R* index is ∼40%) and using the flap procedure for closing the two rings, the *R* index drops to 7% and the i-c converge to reliable values in six to seven cycles. The differences compared with published figures are modest and the correlation coefficients acceptable (<75%, see Table 1). Alternatively, if one uses the traditional method, computing *also* O5 and O8 atoms using the **z**-matrix and imposing the two constraints (O5−C16 and O8−C19 distances of 1.427 Å), the whole process fails: computed shifts are excessive and divergence occurs. Only starting from a structure very close to the true one, and accepting the shifts cut by some 97%, can the minimum be reached.

As the implementation of the constrained minimization according to Lagrange has been repeatedly verified (see ref 11 and references cited therein), we have concluded that the explanation of this bad result is in the intrinsic ill-conditioning of the normal matrix. The i-c redundance of two units could explain the ill-conditioning.

## PROBLEMS RESOLVED BY TRY AND CONCLUSIONS

The general purpose program TRY is useful both in structure assignment and in structure refinement and has been validated solving numerous problems, most polymeric and of considerable complexity (see ref 10 and references cited therein). In addition TRY has been tested reconsidering known complex structures and using published data. In all the cases considered it demonstrated itself to be a very practical tool both in the trial-and-error stage of structural analyses and in the refinement stage.

**Supporting Information Available:** Input data for running the sucrose test (2 ascii files). This material is available free of charge via the Internet at http://pubs.acs.org

## REFERENCES AND NOTES

(1) Eyring, H. The resultant electric moment of complex molecules. *Phys. Rev.* **1932**, *39*, 746−748.
(2) Califano, S. *Vibrational states*; Wiley: New York, 1974; p 275.
(3) Janezic, D.; Praprotnik, M.; Merzel, F. Molecular dynamics integration and molecular vibration theory. I. New Symplectic integrators. *J. Chem. Phys.* **2005**, *122*, 174101-1−174101-14.
(4) Praprotnik, M.; Janezic, D.; Molecular dynamics integration and molecular vibration theory. II. Simulation of nonlinar molecules. *J. Chem. Phys.* **2005**, *122*, 174102-1−174102-9.
(5) Arnott, S.; Wonacott, A. J. The refinement of the crystal and molecular structures of polymers using x-ray data and stereochemical constraints. *Polymer* **1966**, *7*, 157−166.
(6) Smith, P. J. C.; Arnott, S. LALS: A linked-atom least-squares reciprocal-space refinement system incorporating stereochemical restraints to supplement sparse diffraction data. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1978**, *A34*, 3−11.
(7) Pulay, P.; Fogarasi, G.; Pang, F.; Boggs, J. E. Systematic ab initio gradient calculation of molecular geometries, force constants, and dipole moment derivatives. *J. Am. Chem. Soc.* **1979**, *101*, 2550−2560.
(8) Goto, H.; Osawa, E. Corner flapping: A simple and fast algorithm for exhaustive generation of ring conformations. *J. Am. Chem. Soc.* **1989**, *111*, 8950−8951.
(9) Bond, A. D. C60-thiophene disolvate. *Acta Crystallogr., Sect. E: Struct. Rep. Online* **2003**, *E59*, o1992−o1993.
(10) Immirzi, A.; Tedesco, C.; Alfano, D. New solutions to the problems of chain orientation and chain continuity in structure analysis of fibrous polymers. A reconsideration of the structure of polyisobutene. *J. Appl. Crystallogr.* **2007**, *40*, 10−15.
(11) Hynes, R. C.; Le Page Y. Sucrose, a convenient test crystal for absolute structures. *J. Appl. Crystallogr.* **1991**, *24*, 352−354.
(12) Immirzi, A. submitted for publication in *J. Appl. Crystallogr.*
(13) TRY program can be downloaded from the web site: http://www.theochem.unisa.it/try.html.

CI700225X