JOURNAL OF COMPUTATIONAL PHYSICS 37, 56–69 (1980)

Use of Caley–Menger Determinants in the Calculation of Molecular Structures

MICHAEL H. KLAPPER AND DAVID DEBROTA

Department of Chemistry, The Ohio State University, Colombus, Ohio 43210

Received Novermber 14, 1978; revised October 2, 1979

An algorithm has been developed for the direct calculation of molecular structure in the semimetric. distance table representation. The algorithm is based on the Caley–Menger determinant, and an imbeddability theorem taken from distance geometry. An an example, the excluded volume of the tripeptide alanyl-alanyl-alanine has been computed.

Molecular structures are generally computed by generating atomic coordinates from bond lengths, bond angles, and dihedral angles. In turn, the distances d_{ij} between distinct atoms p_i and p_j may be computed given the atomic coordinates. Hence, the distance table **D**, and the coordinate table from which it has been derived represent the same molecular structures:

$$\mathbf{D} = \begin{matrix} 0 & d_{12} & d_{13} & \cdots & d_{1n} \\ d_{12} & 0 & d_{23} & \cdots & d_{2n} \\ d_{13} & d_{23} & 0 & \cdots & d_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ d_{1n} & d_{2n} & d_{3n} & \cdots & 0 \end{matrix} \tag{1}$$

But the two representations do differ in important aspects. For a molecule of n atoms the coordinate table contains $3n$ entries, and the distance table $(n^2 - n)/2$ distinct entries. (The factor of 2 arises from the symmetry of **D**.) On the basis of economy **D** appears to suffer because of the $O(n^2)$ more variables. However, molecular representation in terms of **D** does offer potentially important advantages.

The computation of a molecular structure requires that no two atoms overlap. In a hard sphere model this requirement is expressed by the relationship that the interatomic distance $d_{ij} \geqslant r_i + r_j$, the sum of the noncovalent radii. Molecules may have rings involving four or more atoms. Such rings appear in **D** as d_{ij} cycles with $d_{ij} = r'_i + r'_j$, the two covalent radii, $r'_i < r_i$, $r'_j < r_j$. To compute the structures of large molecules which are self-avoiding and which may also contain cyclic substructures is a major problem. The fact that the distance table explicitly contains the d_{ij} in which overlap and ring constraints are expressed suggests that direct computation of **D** without the intervening construction of a coordinate table may simplify the computation of large molecular structures.

0021-9991/80/100056–14$02.00/0

Crippen [1], who proposed that the direct calculation of **D**, utilizing an imbedding theorem due to Menger [2], would avoid the problems otherwise encountered in the removal of atomic overlaps and the construction of closed rings, has described numerical methods applicable to large molecules [3, 4]. Mackay [5] has solved the problem analytically for structures of only five and six points on the basis of a restricted version of the Menger theorem and has discussed other potential uses of the **D** array. Crippen's method cannot, however, be applied in Monte Carlo calculations, since the numerical procedure forces the solution to an arbitrary structure and randomness is not ensured. The analytic solution of Mackay while of potential usefulness in Monte Carlo calculations must be expanded to large numbers of atoms. In this report we present an algorithm for the direct and analytic computation of **D** for a molecule of any size using the protocol of atom-by-atom construction. In addition, we consider the further problem of molecules with chiral centers. As an example, we present the results of an alanyl-alanyl-alanine (33 atoms) calculation.

THEORY

We consider first those definitions needed for stating the theorem upon which the algorithm is based.

A space *is a set of elements* $P = \{p_1, p_2, ..., p_i, ...\}$ *which we shall call points, and which have an associated topology.*

The topology we require entails distance and is constructed as follows. All ordered point-pairs (p_i, p_j)—the cartesian product—are assigned a distance d_{ij}. While any arbitrary assignment is possible, we are interested specifically in a semimetric space.

The set of points P, and the associate distances d_{ij} *is called a* semimetric space *provided*

$$\begin{array}{ll} \text{(i)} & d_{ij} > 0, \quad i \neq j; \\ \text{(ii)} & d_{ij} = 0, \quad i = j; \\ \text{(iii)} & d_{ij} = d_{ji}. \end{array} \tag{2}$$

The first two conditions are self-explanatory; the third requires that the distance from point p_i to p_j be the same as that from p_j to p_i. The distance table **D** (1) is an example of a semimetric space. The symmetry of **D** is a consequence of condition (iii), the zeros along the diagonal of condition (ii).

A semimetric space is also a metric space *provided the triangle inequality holds for any three distinct points.*

The triangle inequality is a well-known geometric relationship between the three distances associated with the arbitrary points p_i, p_j, p_k

$$d_{ij} + d_{jk} \geqslant d_{ik}. \tag{3}$$

The particular metric space of interest to us is, of course, euclidean space.

The n-dimensional euclidean space E_n *is the set of points described by the ordered, real coordinates* $x_1, x_2, \ldots, x_n$. *The distance associated with the arbitrary point-pair* (p_x, p_y) *is defined by*

$$d_{x,y}^2 = \sum_{i=1}^{n} (x_i - y_i)^2. \tag{4}$$

It is easily shown that $d_{x,y}$ satisfies conditions (2) and (3); hence, E_n is a metric space, as are all subspaces of E_n.

It is the case that numerous spaces may be envisioned. This leads to a reasonable concern over their classification. Based on the definitions given above, a limited classification scheme is possible. While all metric spaces are also semimetric, the converse assertion is not true, since semimetric spaces in which (3) does not apply are well known. Hence, all metric spaces form a proper subset of all semimetric spaces. Similarly, the set of all euclidean spaces is a proper subset of all metric spaces. An ordering on the basis of proper subsets establishes one type of classification. At this point we introduce a definition which is based on the idea of such an ordering and is required later in the discussion, but which also constitutes a brief digression.

A set of $n + 1$ *points from a euclidean space is called* independent *provided it is not also a subspace of* E_{n-1}; *otherwise, it is called dependent.*

For example, four coplanar points are a subspace of E_2, and, therefore, are dependent. The vertices of a tetrahedron are independent.

A problem in any classification scheme is how to determine whether an arbitrary example is contained within any one of the delineated subsets. For example, given an arbitrary distance table **D** how does one determine whether the semimetric space it describes is also a space in E_3. Classification from the more general to the more specific is a problem of imbedding.

If an arbitrary space is an element of a set, and also of a proper subset, then the space is imbeddable *in that subset.*

Thus, a plane described in E_3 is imbeddable in E_2, but not in E_1; the semimetric space described by an arbitrary n by n distance table is imbeddable in E_3 provided the proper $(n/2)(n-7)+6$ constraints are met. Imbedding of an independent set of points warrants an additional definition.

A semimetric space with $n + 1$ *points is* irreducibly imbeddable *in* E_n *provided it is independent in* E_n.

There must be one or more criteria to decide whether a space is imbeddable within a subset. The criterion to be used here is that of a similarity defined in terms of congruence.

Assume the ordered point-pair (p_i, p_j) *in the space* P, *and the ordered point-pair* (q_i, q_j) *in* Q. *The two point pairs are* congruent *if* $d_{ij}^P = d_{ij}^Q$.

Congruence implies the conservation of distance and is usefully applied to spaces.

Two spaces P and Q are congruent provided there exists a function f which maps $P \to Q$, *such that each point-pair of P is mapped onto a congruent point-pair in Q.*

By definition, symmetry operations on a geometric figure conserve distance. Thus, the two mirror images of any three-dimensional figure are congruent; similarly, the translation of a set of points in E_n generates a congruent space. Congruence suggests a similarity between spaces, and for this reason serves as a reasonable classification criterion. Hence, we shall use the term congruently imbeddable. Note that two congruent spaces need not be identical; for example, mirror images.

As mentioned in the introduction, a semimetric distance table **D** is easily generated from a coordinate table, but for an arbitrary **D** to be imbeddable in E_3, $n(n-7)/2+6$ constraints must be met. The explicit formulation of these constraints is given in a theorem of Menger [2] which is based upon the Caley–Menger determinant. This determinant is defined next, followed by a statement of the theorem.

The Caley–Menger determinant *associated with the points* $p_1, p_2, \ldots, p_n$ *is given by*

$$\mathrm{CM}(p_1, p_2, \ldots, p_n) = \begin{vmatrix} 0 & 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & 0 & d_{12}^2 & d_{13}^2 & d_{14}^2 & \cdots & d_{1n}^2 \\ 1 & d_{12}^2 & 0 & d_{23}^2 & d_{24}^2 & \cdots & d_{2n}^2 \\ 1 & d_{13}^2 & d_{23}^2 & 0 & d_{34}^2 & \cdots & d_{3n}^2 \\ 1 & d_{14}^2 & d_{24}^2 & d_{34}^2 & 0 & & \\ \vdots & \vdots & \vdots & \vdots & & & \vdots \\ 1 & d_{1n}^2 & d_{2n}^2 & d_{3n}^2 & & \cdots & 0 \end{vmatrix} \tag{5}$$

This determinant for n points will also be abbreviated as $\mathrm{CM}(n)$.

THEOREM. *An arbitrary semimetric space S is congruently and irreducibly imbeddable in* E_n *if and only if* (i) *S contains at least one set of points* $p_1, p_2, \ldots, p_r$ *with* $r \leqslant n+1$ *such that the signs of* $\mathrm{CM}(p_1, \ldots, p_k)$ *are given by*

$$\operatorname{sgn} \mathrm{CM}(p_1, p_2, \ldots, p_k) = (-1)^k, \tag{6}$$

for all $k = 2, \ldots, r$; (ii) *for every pair of points* p_i *and* p_j *distinct from* $p_1, \ldots, p_r$, *and from each other*

$$\begin{aligned} \mathrm{CM}(p_1, \ldots, p_r, p_i) &= \mathrm{CM}(p_1, \ldots, p_r, p_j), \\ &= \mathrm{CM}(p_1, \ldots, p_r, p_i, p_j) = 0. \end{aligned} \tag{7}$$

The first condition (6) implies that the semimetric space must contain at least one set of ordered points such that $\mathrm{CM}(p_1, p_2) > 0$, $\mathrm{CM}(p_1, p_2, p_3) < 0$, $\mathrm{CM}(p_1, p_2, p_3, p_4) > 0$, etc. The second condition requires that all $\mathrm{CM}(r+1)$ and

CM($r+2$) constructed with the initial $p_1,\ldots,p_r$ vanish. Rules for an algorithm which constructs imbeddable semimetric spaces can be derived from this theorem.

Assume a set of $m-1$ points imbeddable in E_3. To this set add one point such that all the distances between this last point and the original set but one are known. The two points with the mutual, unknown distance shall be called the test set, and the remaining $m-2$ points the reference set. Since there is only one unknown distance, and since the points may be numbered in any convenient order (renumbering is a symmetry operation), the Caley–Menger determinant may be written as (for the case $m=5$):

$$\mathrm{CM}(5)=\begin{vmatrix} 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & d_{12}^2 & d_{13}^2 & d_{14}^2 & d_{15}^2 \\ 1 & d_{12}^2 & 0 & d_{23}^2 & d_{24}^2 & d_{25}^2 \\ 1 & d_{13}^2 & d_{23}^2 & 0 & d_{34}^2 & d_{35}^2 \\ 1 & d_{14}^2 & d_{24}^2 & d_{34}^2 & 0 & x \\ 1 & d_{15}^2 & d_{25}^2 & d_{35}^2 & x & 0 \end{vmatrix}. \tag{8}$$

Equation (8) is the example for 5 points with x representing the unknown d_{45}^2. For the point set to be irreducibly imbeddable in E_3, $\mathrm{CM}(5)=0$. It can then be shown by expansion that

$$\mathrm{CM}(5)=Ax^2+Bx+C=0, \tag{9}$$

where the coefficients A, B, and C can be evaluated. This quadratic equation has the following properties:

(I) If the reference set is dependent, x is indeterminate.

(II) If the reference set is independent, and either of the sets of $m-1$ points not involving x dependent, then (9) has two real, positive roots which are identical.

(III) Otherwise (9) has two real, positive roots which are not identical.

The derivation of these properties as given here is based upon expansion of (8). Only the case of CM(5) will be treated explicitly; the proof for the arbitrary CM(n) should then be obvious. The expansion of (8) yields the values of A, B, and C [5].

$$A=-\begin{vmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & d_{12}^2 & d_{13}^2 \\ 1 & d_{12}^2 & 0 & d_{23}^2 \\ 1 & d_{13}^2 & d_{23}^2 & 0 \end{vmatrix} =-\mathrm{CM}(3), \tag{10}$$

$$B=-2\begin{vmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & d_{12}^2 & d_{13}^2 & d_{14}^2 \\ 1 & d_{12}^2 & 0 & d_{23}^2 & d_{24}^2 \\ 1 & d_{13}^2 & d_{23}^2 & 0 & d_{34}^2 \\ 1 & d_{15}^2 & d_{25}^2 & d_{35}^2 & 0 \end{vmatrix}, \tag{11}$$

$$C = \begin{vmatrix} 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & d_{12}^2 & d_{13}^2 & d_{14}^2 & d_{15}^2 \\ 1 & d_{12}^2 & 0 & d_{23}^2 & d_{24}^2 & d_{25}^2 \\ 1 & d_{13}^2 & d_{23}^2 & 0 & d_{34}^2 & d_{35}^2 \\ 1 & d_{14}^2 & d_{24}^2 & d_{34}^2 & 0 & 0 \\ 1 & d_{15}^2 & d_{25}^2 & d_{35}^2 & 0 & 0 \end{vmatrix}. \tag{12}$$

The coefficient A is the negative of CM(3). The coefficient B is obtained from (8) by deletion of the last column, and the next to last row (or vice versa), replacement of the one remaining x by zero, and multiplication of the resultant, unsymmetric determinant by -2. Finally, C is obtained by replacing both x's in (8) with zeros.

The determinant in (11) can be expanded along the lowest row to give

$$B/2 = b_1 + d_{15}^2 b_2 + d_{25}^2 b_3 + d_{35}^2 b_4, \tag{13}$$

where the coefficients b_i are obtained by Cramer's rule for the solution of the set of equations:

$$\begin{aligned} 0 \; &+ \quad \alpha_2 + \quad \alpha_3 + \quad \alpha_4 = 1, \\ \alpha_1 \; &+ \quad 0 \quad + d_{12}^2 \alpha_3 + d_{13}^2 \alpha_4 = d_{14}^2, \\ \alpha_1 \; &+ d_{12}^2 \alpha_2 + \quad 0 \quad + d_{23}^2 \alpha_4 = d_{24}^2, \\ \alpha_1 \; &+ d_{13}^2 \alpha_2 + d_{23}^2 \alpha_3 + \quad 0 \quad = d_{34}^2. \end{aligned} \tag{14}$$

The coefficient determinant of (14) is CM(3) and, thus, equals $-A$. Therefore,

$$\begin{aligned} -\alpha_i A &= b_i, \qquad i = 1,\ldots, 4; \\ B &= -2A(\alpha_1 + d_{15}^2 \alpha_2 + d_{25}^2 \alpha_3 + d_{35}^2 \alpha_4). \end{aligned} \tag{15}$$

If the three reference points are collinear, they are imbeddable in E_1, and thus dependent. As a result, $A = 0$ on the basis of the theorem, $B = 0$ by Eq. (15), and x is indeterminate, establishing property I.

If the three reference atoms are not collinear, they are not imbeddable in E_1 and $A \neq 0$. Because (9) is the equation of a parabola, the x coordinate at the vertex is

$$\begin{aligned} x_v &= -B/2A, \qquad B = -2Ax_v; \\ x_v &= \alpha_1 + d_{15}^2 \alpha_2 + d_{25}^2 \alpha_3 + d_{35}^2 \alpha_4. \end{aligned} \tag{16}$$

The determinant (12) can be treated similarly.

$$C = c_1 + d_{15}^2 c_2 + d_{25}^2 c_3 + d_{35}^2 c_4, \tag{17}$$

where the coefficients c_i are obtained by Cramer's rule for the solution of the linear equations:

$$
\begin{aligned}
0 &+ \alpha_2' + \alpha_3' + \alpha_4' + \alpha_5' = 1,\\
\alpha_1' &+ 0 + d_{12}^2\alpha_3' + d_{13}^2\alpha_4' + d_{15}^2\alpha_5' = d_{14}^2,\\
\alpha_1' &+ d_{12}^2\alpha_2' + 0 + d_{23}^2\alpha_4' + d_{25}^2\alpha_5' = d_{24}^2,\\
\alpha_1' &+ d_{13}^2\alpha_2' + d_{23}^2\alpha_3' + 0 + d_{35}^2\alpha_5' = d_{34}^2,\\
\alpha_1' &+ d_{14}^2\alpha_2' + d_{24}^2\alpha_3' + d_{34}^2\alpha_4' + 0 = 0.
\end{aligned}
\tag{18}
$$

The coefficient determinant in this case is the transpose of the determinant in (11); hence,

$$
\begin{aligned}
&-\alpha_i'(B/2) = c_i, \qquad i = 1,\ldots, 5;\\
&C = -(B/2)(\alpha_1' + d_{15}^2\alpha_2' + d_{25}^2\alpha_3' + d_{35}^2\alpha_4');\\
&C = A(\alpha_1 + d_{15}^2\alpha_2 + d_{25}^2\alpha_3 + d_{35}^2\alpha_4)(\alpha_1' + d_{15}^2\alpha_2' + d_{25}^2\alpha_3' + d_{35}^2\alpha_4').
\end{aligned}
\tag{19}
$$

Comparing Eq. (14) and (18), it is evident that if $\alpha_5' = 0$, then $\alpha_i = \alpha_i'$; otherwise, $\alpha_i \neq \alpha_i'$. From Eq. (14), and by Cramer's rule,

$$
\alpha_5' = -(2/B)\begin{vmatrix}
0 & 1 & 1 & 1 & 1\\
1 & 0 & d_{12}^2 & d_{13}^2 & d_{14}^2\\
1 & d_{12}^2 & 0 & d_{23}^2 & d_{24}^2\\
1 & d_{13}^2 & d_{23}^2 & 0 & d_{34}^2\\
1 & d_{14}^2 & d_{24}^2 & d_{34}^2 & 0
\end{vmatrix}
\tag{20}
$$

The determinant in (20) is CM(4). If the four points p_1, p_2, p_3, p_4 are dependent, they are coplanar, or imbeddable in E_2. Hence, CM(4) vanishes by the imbeddability theorem, $\alpha_5' = 0$ and $\alpha_i = \alpha_i'$. Therefore, from (16) and (19),

$$
C = Ax_v^2. \tag{21}
$$

But (16) and (21) imply that the discriminant of the quadratic (9) vanishes, and hence, the two roots are identical and equal to x_v. This establishes property II. If p_1, p_2, p_3, p_4, are independent, CM(4) $\neq 0$, the discriminant of (9) is not zero, and the two roots of (9) are distinct (property III). Using the three properties of (9) together with (6) and (7) assures the construction of a semimetric space imbeddable in E_3. The corresponding geometric proof is indicated in Fig. 1.

Structural information is usually supplied in the form of bond lengths, bond angles, and dihedral angles. Hence, the two angles (defined in Fig. 1) must be converted to distances before beginning computations. The distance between the two atoms about the bond angles is calculated by the law of cosines. The distance between atoms related by a dihedral angle may be calculated as follows. The maximum distance $d_{\max}$ between the two atoms is achieved in the *trans*-planar configuration; the minimum distance $d_{\min}$ in the *cis*-planar configuration. Since four coplanar points are irreducibly imbeddable in E_2, CM(4) vanishes, and the quadratic (9) has two distinct

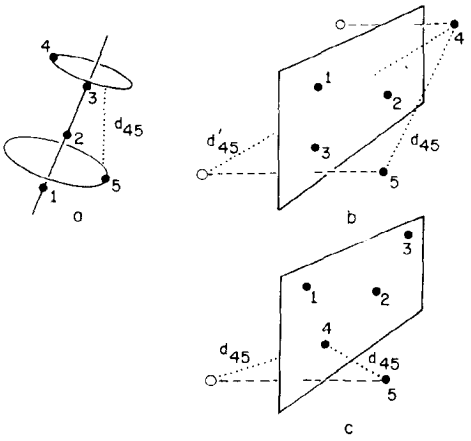


FIG. 1. Illustration of properties I, II, and III of Eq. (9) for the case of five points in E_3. (a) If the reference set (points 1, 2, and 3) is collinear, points 4 and 5 can be anywhere on the two respective circles perpendicular to the line as indicated. Thus, the distance between points 4 and 5 cannot be determined. (b) If the reference set is coplanar, points 4 and 5 can each be placed on either side of the reference plane. Therefore, there are two distinct solutions. (c) If four of the five points are coplanar, then even though the last point can be placed on either side of (or on) the plane, there is only one value of d_{45}.

roots, one of which is $d^2_{\min}$, the other $d^2_{\max}$. Then from simple trigonometric considerations (Fig. 2) it can be shown that

$$d^2 = d^2_{\min} + (d^2_{\max} - d^2_{\min}) \sin^2(\theta/2). \tag{22}$$

One last problem must be faced. A molecule with n chiral centers can be constructed 2^n ways excluding bond rotations and molecular translations. These constructs can, however, be grouped as pairs congruent to one another by some symmetry operation. The 2^{n-1} noncongruent isomers require 2^{n-1} distance tables. For example, a molecule with one chiral center has one of two possible structures related by the symmetry operation of inversion. Since distance is preserved in a symmetry operation, one distance table represent both isomers. In the case of two chiral centers, two distance tables are required to exhaust all congruent structures; in one set the pairs are congruent by inversion, in the other by rotation. For all molecules with $n > 1$ chiral centers $n - 1$ additional constraints are required to ensure computation of the correct **D**. Expressed differently, atoms must be placed around $n - 1$ chiral centers, relative to the placement about one reference center.

In a chain of amino acids the peptide bond is assigned as *trans*-planar. Therefore, the backbone conformation is entirely specified by the dihedral angles around the N–C^α and C_α–C′ bonds (Fig. 2; see [6] for conventions and nomenclature). The chain can, therefore, be visualized as a series of planes linked by C^α atoms acting as swivel points. Identical relative geometry about each C^α is thus established by always

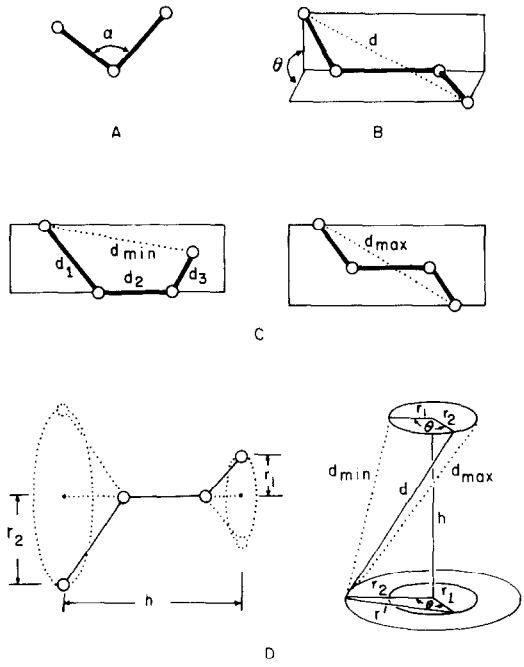


FIG. 2. Illustrations of geometric relationships discussed in the text: (A) bond angle; (B) dihedral angle; (C) minimum and maximum dihedral distances; (D) geometric visualization for the derivation of Eq. (22). r_1, r_2, radii of circles projected by dihedral rotations; h, distance between projected circles. The derivation is based upon the following relationships:

$$d_{\min}^2 = (r_2 - r_1)^2 + h^2 \Rightarrow d_{\max}^2 - d_{\min}^2 = 4r_1r_2,$$
$$d_{\max}^2 = (r_2 + r_1)^2 + h^2$$
$$r'^2 = r_1^2 + r_2^2 - 2r_1r_2\cos\theta \Rightarrow d^2 = d_{\min}^2 + 2r_1r_2(1 - \cos\theta).$$
$$d^2 = r'^2 + h^2$$

placing N and X on the same side of the plane having first "normalized" the rotations about the C_i^{α}–C and N–C_{i+1}^{α} bonds. Since angles are not stored in a distance table, an algorithm which uses distance in this normalization procedure has been developed. This algorithm will be described in the next section.

METHODS AND RESULTS

The algorithm generates polypeptide structures randomly in order to obtain a uniform sampling over all conformational space. Conformation variations are obtained by dihedral rotations, with bond distances and bond angles held fixed. The protocol is to add one atom at a time to a growing structure by calculating its distance to all previous atoms with Eqs. (7) and (22), and properties I through III. Atomic overlaps are then determined for an excluded volume calculation.

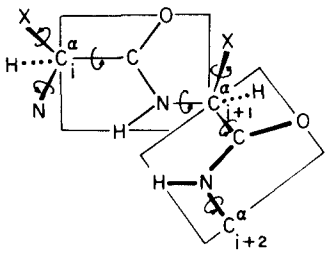


FIG. 3. Relationship of peptide planes.

The distance between covalently bonded atoms, between two atoms bonded to a third, common atom, and between all atoms in the plane of the peptide unit (Fig. 3) are constant over all molecular conformations. These distances, which in the example of Ala_3 constitute approximately 100 of the required 528 entries, are entered into **D** before starting the calculations. All initial structural parameters are identical to those used elsewhere [7].

The distance between the two outside atoms of a quartet defining a dihedral angle (to be called the dihedral distance) falls in the range $d_{min} \leqslant d \leqslant d_{max}$ (Fig. 2). The distances d^2_{min} and d^2_{max} are obtained as the two roots of the quadratic equation $CM(4) = 0$. Generation of a random distance within this range is equivalent to the generation of a dihedral angle in the range $-\pi \leqslant \theta \leqslant \pi$. Uniform sampling over conformational space requires a uniform distribution over θ rather than over the range d_{max}–d_{min}. Since the dihedral distance is not directly proportional to θ, a random sample of d was obtained with (22) and a random, uniformly distributed sample of dihedral angles. The sign of θ, which is lost in Eq. (22), specifies the rotation direction and is used explicitly as discussed below. Had we not been interested in a uniform sampling we could have dispensed with Eq. (22).

Atom-by-atom construction involves the repeated use of a few elementary procedures; these will be described in terms of the specific example Ala_3. (Note that the calculations are general for any polypeptide chain, and that Ala_3 is used as an example only so that the discussion is more concrete.) The carbonyl oxygen, atom 8 (Fig. 4), is placed after all the interatomic distances for atoms 1 through 7 have been determined. The dihedral distance d_{38} is assigned randomly. A CM(5) can then be constructed with one unknown distance, d_{78} using atoms 3, 5, 6 as the reference set.

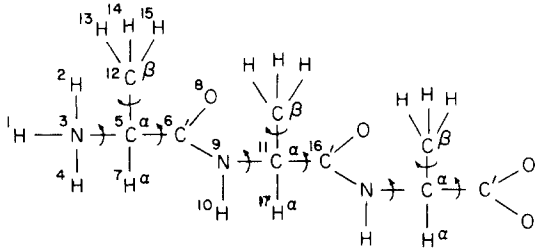


FIG. 4. Numbering scheme used in the calculation of Ala_3.

Since the reference set is not collinear, and since atoms 3, 5, 6, 7 are not coplanar, the quadratic equation $CM(5) = 0$ has two distinct roots. One corresponds to the O–H^{α} distance after a clockwise rotation about the C^{α}–C′ bond, the other after a counterclockwise rotation. If the value of θ used to generate d_{38} is positive, the larger root is chosen; otherwise, the smaller. Atoms 3, 5, 6, 7 can then serve as the reference set for the three CM(6) determinants used in calculating d_{18}, d_{28}, and d_{48}. The reference set is not coplanar so that three quadratic equations each with one distinct root are obtained. Atom 8 is placed.

Placement of atoms 9, 10, and 11—N, H, and C^{α} in the peptide plane—is an easier task. For example, d_{19} can be calculated with atoms 5, 6, 8 as a reference set. Since atom 9 is coplanar with the reference set, the equation $CM(5) = 0$ has one distinct root. The same reference set is used similarly to calculate all unknown distances to 9, 10, and 11.

Symmetry simplifies the calculations. Placement of atom 6 is one example. The dihedral distance d_{16} is assigned randomly. A CM(5) is constructed with the noncollinear, reference set 1, 3, 5 and the test pair 2, 6. Since neither 1, 3, 5, 6 nor 2, 3, 5, 6 is coplanar, the equation $CM(5) = 0$ has two distinct roots corresponding to clockwise and counterclockwise rotations. But the three ammonium hydrogens are equivalent so that rotation is a symmetry operation; therefore, one root can be assigned do d_{26}, the other to d_{46}, and atom 6 is placed.

Imposing the proper relative geometry about asymmetric centers is the most difficult task, conceptually. Since all amino acids in proteins have the same absolute geometry, the relative geometries about each C^{α} were assigned identically. Begin with the definition of two test constants: t_{max} is the value of d_{38} when atoms 7 and 8 are placed in the *trans*-planar configuration; t_{min} is the value of d_{38} for the *cis*-planar case. Both constants are easily calculated using one CM(5). When in the calculation of d_{38} $\theta > 0$, $d_t = t_{max}$; otherwise $d_t = t_{min}$. The placement of atom 16 begins with a random assignment of the dihedral distance $d_{6,16}$. The CM(5) with atoms 6, 9, 11 as reference set and atoms 7, 16 as test pair yields an equation with two distinct roots. If in the calculation of $d_{6,16}$ $\theta \leqslant 0$ *and* $d_{38} \leqslant d_t$, or if $\theta > 0$ *and* $d_{38} > d_t$, then the smaller root is chosen. Otherwise, $d_{7,16}$ is assigned the larger root. The final step in the chirality determination involves placement of atom 17. The CM(5) with reference set 9, 11, 16 and test set 6, 17 yields a quadratic equation with two distinct roots; the smaller is assigned to $d_{6,17}$ if the value of θ used in the $d_{6,16}$ dihedral distance calculation is positive.

In this algorithm the peptide plane (Fig. 3) serves as reference. Atoms 7 and 16 are placed either on the same side of the plane (smaller root) or on opposite sides (larger root). Whether placement is to be on the same or opposite sides depends upon the directions (angle signs) of rotation around C_i^{α}–C′ and N–C_{i+1}^{α} and the rotation magnitude (comparison with d_t) around C_i^{α}–C′. A complete discussion of the reasoning behind this algorithm is difficult without a physical model. The reader is encouraged to utilize a Labquip or similar molecular model and to rotate for himself.

The four procedures just outlined suffice for the complete construction of any polypeptide, although the descriptions dealt with particular parts of the Ala_3

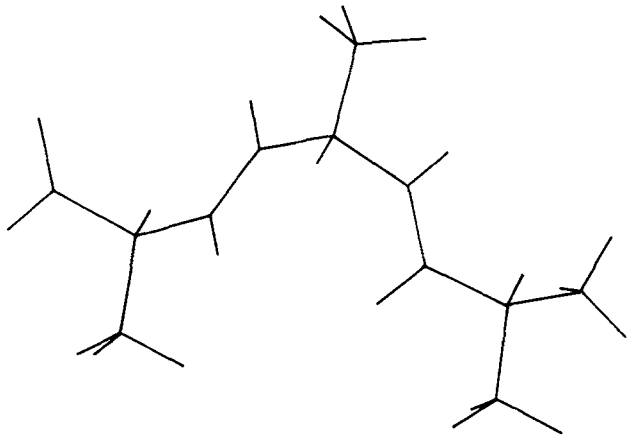
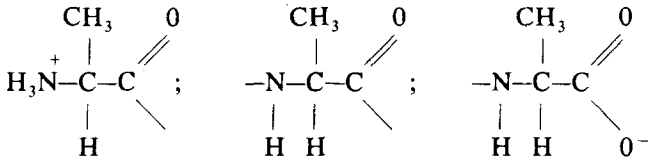


FIG. 5. A structure of Ala_3 first computed as a distance table.

molecule. One successful Ala_3 structure is shown in Fig. 5. The coordinates for the generation of this line drawing were calculated from **D** by Eq. (6) in Ref. [1]. We encountered no numerical instability in the coordinate calculations due to roundoff errors, and thus, an iterative procedure was unnecessary.

The structure-building algorithm described in this report has been used in a Monte Carlo calculation of excluded volume for the case of Ala_3. Each randomly generated Ala_3 molecule was considered in terms of three fragments:



The fraction of conformational space accessible to a particular fragment is estimated by the fraction (f_{obs}) of all generated structures in which there is no atomic overlap in that fragment. The fraction accessible to the total molecule is estimated by the fraction of structures with no atomic overlaps whatosever. When inter- but not intrafragment atomic overlap is allowed, the fraction accessible is estimated as the product (f_{calc}) of fragment f_{obs}'s. For the example of Ala_3, the probability of generating a structure with no intrafragment overlaps is the product of the three individual probabilities associated with the fragments shown above.

For our purposes, the excluded volume is defined as the difference between the volumes in conformational space accessible to a structure with and without allowed intrafragment overlaps. Since volume in conformational space is not well defined in a Monte Carlo calculation, while volume fraction is, the excluded volume is discussed practically in terms of the ratio f_{obs}/f_{calc}, which is also a ratio of volumes. The results obtained with Ala_3 assuming a hard-sphere model are presented in Table I. Approx-

TABLE I

Excluded Volume in Alanyl-alanyl-alanine

Structure	f^{a}_{obs}	f_{calc}
N-terminal alanyl	0.694 (0.0096)[b]	—
Central alanyl	0.657 (0.0099)	—
C-terminal alanine	0.432 (0.0103)	—
Alanyl-alanyl-alanine	0.0023 (0.0012)	0.197 (0.00672)

$f_{obs}/f_{calc} = 0.0121$ (0.0061); $\Delta S_{excluded} = -8.7$ (1.0) eu.

[a] Fraction of structures with no interatomic overlaps; calculated on the basis of 2315 constructions.

[b] Figures in parentheses are standard deviations. For f_{obs} these were estimated by $f_{obs}(1 - f_{obs})/N$, where N is the total number of structures tested. Others obtained by error propagation.

imately 99.8% of conformational space is denied Ala_3 when no overlaps are permitted, while only 80% is denied a structure in which overlaps between fragments are allowed. This represents an entropy loss of approximately 9 eu for assembly of the three fragments. The observed space restriction is increasingly severe for each additional residue and becomes an insurmountable hurdle for Monte Carlo calculations of even moderately sized polypeptide chains.

In conclusion, we have shown that a semimetric representation, the distance table **D**, of a molecular structure can be constructed successfully using Caley–Menger determinants. Two important limitations are the indeterminancy associated with Caley–Menger determinants based on a dependent reference point set and lack of explicit information in **D** concerning inversion. Both limitations can be overcome. The structure-building algorithm developed here as a test of the approach involves atom-by-atom construction. A very cursory examination of the excluded volume problem indicates that the atom-by-atom method cannot be applied to large molecules. Since the distance table contains explicitly all the information needed to define the excluded volume, it should be possible to generate analytically all atomic positions simultaneously within the constraints imposed by the exclusion of atomic overlaps, thus circumventing the excluded volume barrier. Such a procedure, which can be formulated in terms of a set of linear equations coupled to lower boundary conditions, is theoretically possible. When implemented, this approach should allow Monte Carlo calculations that would open areas of structural theory hitherto inaccessible to protein and polymer chemists.

ACKNOWLEDGMENTS

David DeBrota was a participant of the National Science Foundation Undergraduate Research Program at The Ohio State University. Michael H. Klapper was a recipient of an NIH Career Development Award. The authors are especially grateful to the Ohio State University Instruction and Research Computer Center for providing time on an Amdahl 470V/6II.

REFERENCES

1. G. M. CRIPPEN, *J. Computat. Phys.* **24** (1977), 96.
2. L. M. BLUMENTHAL, "Theory and Applications of Distance Geometry," 2nd ed., Chelsea, New York, 1970.
3. G. M. CRIPPEN, *J. Computat. Phys.* **26** (1978), 449.
4. G. M. CRIPPEN AND T. F. HAVEL, *Acta Crystallogr.* **A34** (1978), 282.
5. A. L. MACKAY, *Acta Crystallogr.* **A30** (1974), 440.
6. *J. Mol. Biol.* **52** (1970), 1.
7. J. GELLES AND M. H. KLAPPER, *Biochim. Biophys. Acta* **533** (1978), 465.