Deriving Three-Dimensional Representations of Molecular Structure from Connection Tables Augmented with Configuration Designations Using Distance Geometry

JACQUES C. WENGER and DENNIS H. SMITH*

Department of Chemistry, Stanford University, Stanford, California 94305

Received September 21, 1981

A method for obtaining three-dimensional representations of molecular structures, from connection tables augmented with designations of configurational stereochemistry, is described. This method employs distance geometry to obtain X, Y, and Z coordinates for all atoms under constraints which preserve the specified configurations and interatomic distances, the latter being derived from tables of standard bond lengths and angles supplemented with any additional information which serves to constrain possible distances. The method is embodied in a computer program, called BUILD3D, which produces structures suitable for presentation and interpretation on standard graphics terminals. Several examples are given to illustrate the method.

INTRODUCTION

Structure elucidation is one important area of application of computer techniques to chemical problems. Ongoing work of the DENDRAL group at Stanford University has led to the design of structure elucidation programs CONGEN[1] and GENOA.[2] These programs are able to generate all constitutional isomers which satisfy specified constraints inferred by a program or a chemist from available analytical data. A recent addition, the STEREO program,[3] can generate, again within specified constraints, all possible configurational stereoisomers. These programs actually generate within the computer tabular representations of molecular structure. These representations are in the form of connection tables which define atom connectivities and various atom and bond properties, augmented with a "parity" bit ("0" or "1") specifying the configuration of chiral centers or the cis or trans orientation of double bonds. These representations convey nothing about the actual (or potential) geometry of a structure. Thus, bond lengths, bond angles, and other geometric properties are not included in the connection tables.

Another important area of application of computers to chemistry includes several numerical methods, whose purpose is to refine the geometry or determine electronic properties of structures, using molecular mechanics[4] or quantum mechanical[5] calculations. These methods require a three-dimensional representation of the structure under investigation; in other words, they begin with a given conformation of a structure.

The missing link between these two areas is a program capable of producing a reasonable three-dimensional representation of a structure beginning with a connection table augmented with stereochemical descriptors. Given such representations, further processing of the structures to yield optimized geometries, etc., can proceed by using existing programs.

Our current structure elucidation programs CONGEN, GENOA, and STEREO incorporate modules for both teletype and graphical drawings of structures. Both methods of presentation begin with a drawing module capable of producing a molecular diagram on a teletype (i.e., character-based) terminal.[6] This is achieved by computing approximate three-dimensional coordinates which are passed directly to a module for graphical (i.e., vector drawing) output or are transformed to two-dimensional coordinates suitable for subsequent representation in a rectilinear grid.[6] These drawings convey molecular connectivity but do not incorporate any stereochemical information. There is no simple method to modify such programs to include even configurational stereochemical information. Molecular configurations obtained by STEREO are presented using special teletype-oriented drawings.[3] Configurations are conveyed by using special symbols for "up" or "down" or cis or trans bonds. However, these drawings also convey no detailed geometric information and become difficult to interpret for bridged systems with several stereocenters. Examples of such teletype-oriented drawings are presented in Figure 1 for $\Delta^{1,2}$-3α-hydroxy-8β-amino-*trans*-decalin (**1**).

The central problem is, then, obtaining three-dimensional representations of structure (e.g., in the form of X, Y, and Z coordinates) from connection tables augmented with configurational designations. *These representations must at least preserve the molecular constitution and configurational stereochemical designations.* An approach which is independent of sets of templates is important due to the wide variety of structural types and complex ring systems obtained from the structure generating programs CONGEN and GENOA. In addition, a method for solution should be able to incorporate geometrical constraints in order that bond lengths, bond angles, and torsional angles be properly represented when such information is available. Although such a method may produce an arbitrary conformation within such constraints, any reasonable conformation can be passed directly to other programs for subsequent optimization.

Several procedures for three-dimensional model building have been discussed in the literature. For reasons summarized below, none is capable of meeting the criteria discussed in the preceding paragraph. One method is due to Pople,[7] where the information input to the model builder is in the form of a connection table. However, geometrical information is conveyed implicitly in the ordering of entries within the connection table in order to specify the conformation of chains and rings; arbitrary ordering and the presence of heteroatoms usually results in failure to achieve ring closure in structures possessing even simple ring systems. In essence, this model builder is useful when one is able to specify the rotameric state of each bond in a desired conformation, especially for rings. We do not have this information for the arbitrary structural types which may be produced by our structure generators.

Another method to obtain Cartesian coordinates has been mentioned briefly by Wipke.[8] This method begins with a structure, input to the computer by a chemist, in the form of a structure drawing, e.g., on a graphics terminal. Special bond symbols are used to represent "up" and "down" with respect to the plane of the screen. Approximate X, Y, and Z coordinates can be derived directly from such input and used as the starting point for geometry optimization.[8] A related method, which uses two-dimensional screen coordinates to-

0095-2338/82/1622-0029$01.25/0

gether with distance geometry to obtain a three-dimensional representation, has recently been presented.[9] These methods are not applicable to our problem because we have no general way of "perceiving" the two- or three-dimensional relationships among atoms in our connection tables from the connection tables alone.

A third approach uses a library of templates, each a molecular fragment, along with a set of rules governing the assembly of these fragments into complete molecules. Cohen et al.[10] recently described such a program where a limited set of rings of size 4–8, along with all their possible conformational patterns, are coded in special routines. The generality of such a method is, however, limited by the size and the nature of the sets of templates and rules. No reasonably sized set can be expected to provide adequate building blocks for difficult cases like bridged polycyclic systems.

In this report we describe a new approach to the design of a template-free model builder. The program requires only the connection table of a molecule, with or without parity designations specifying the configurations of chiral centers and double bonds. If such designations are present, the resulting three-dimensional model preserves the given stereochemistry. This program, in conjunction with an interactive graphics package, allows a chemist to produce perspective or ball-and-stick drawings of the molecule under investigation. The resulting X, Y, and Z coordinates are saved on a file and can be used by other programs which require coordinate data as input.

METHOD

At the heart of our method is a variation of the method of distance geometry described by Crippen[11,12] and Crippen and Havel.[13] This method is designed to derive a set of Cartesian coordinates for a molecule based on a given set of interatomic distances among all atoms in the molecule. The distances are defined in an $n \times n$ matrix, called the *distance matrix* where n is the number of atoms in the molecule and each entry i, j ($i, j = 1, 2, \ldots, n, i \neq j$) corresponds to the interatomic distance between atom i and atom j. An interatomic distance may potentially be a *range* of distances rather than a particular value. Therefore, the actual form of the distance matrix used in the procedure stores the *maximum* allowed distances ($i < j$ entries) in the upper half of the matrix and the minimum allowed distances in the lower half ($i > j$ entries) of the matrix. The method constructs a three-dimensional representation of the molecule which (a) satisfies the distance constraints expressed by the distance ranges in the matrix and (b) satisfies the configuration constraints expressed by any configuration designation in a connection table suitably transformed.

It is well-known that a given set of interatomic distances, or distance ranges, does not necessarily specify uniquely a three-dimensional structure. However, because the procedure is capable of preserving any specified configurational stereochemistry, any and all conformations obtained at least maintain that stereochemistry. As we illustrate below, the method thus fulfills our objectives outlined in the Introduction and provides a coordinate-based representation which can then be passed to other programs which require coordinates as a starting point. In our applications, the distance geometry procedures are part of a larger computer program, called BUILD3D, which both processes input structures and displays the results in a variety of ways on graphics terminals.

Structural Input. In our applications, structures are obtained from the CONGEN or GENOA programs as connection tables representing only molecular constitution. Subsequent analysis by the STEREO program augments the connection tables with configuration designations. For the purposes of this presentation, the actual source of such connection tables is unimportant (for example, using our programs a single structure can be defined with or without configurations and processed by BUILD3D).

Table I. Connection Table for $\Delta^{1,2}$-3α-Hydroxy-8β-amino-*trans*-decalin (1)

atom	type	neighbors[a]			config
1	C	10	2	2	1
2	C	3	1	1	0
3	C	11	4	2	1
4	C	5	3		
5	C	10	6	4	1
6	C	7	5		
7	C	8	6		
8	C	12	9	7	0
9	C	10	8		
10	C	5	1	9	0
11	O	3			
12	N	8			

[a] Each entry in the neighbors list represents an atom to which the given atom is bonded. Multiple entries mean multiple bonds. For example, atom 1 is connected to atom 10 and twice to atom 2, the 1,2-double bond.

As an illustrative example, consider the connection table presented in Table I for structure **1** (Figure 1). The configuration designations are referenced to the numbering shown for **1** and correspond to the convention specified previously.[14,15] Such connection tables are passed directly to BUILD3D, which constructs the three-dimensional model.

Placing Bounds on the Distance Matrix. The next step in the procedure is to set distance ranges in the distance matrix by using the connectivity of the molecule as specified in a connection table (e.g., Table I) as a starting point. Because there is no geometrical information in the connection table, we obviously cannot determine exactly all the elements of the distance matrix. However, we can establish 1,2 and 1,3 distances and set limits on possible 1,4 distances as follows:

(1) 1,2 distances are set equal to selected standard bond lengths;[16]

(2) 1,3 distances (d_{ik}, between atoms i and k bonded to a common atom j) are computed from 1,2 distances and standard bond angles,[16] using the law of cosines (eq 1).

$$d_{ik} = (d_{ij}^2 + d_{jk}^2 - 2d_{ij}d_{jk}\cos\theta)^{1/2} \quad (1)$$

At this point, small rings of size 3 and 4 are identified by using the Welsh–Gibbs–Assembly algorithm described by Dyott.[19] 1,3 Distances are determined for these rings by using the nonstandard bond angles of these systems. For both (1) and (2) $d_{ij} = l_{ij} = u_{ij}$ stands for the distance between the ith and jth atom and l_{ij} and u_{ij} stand for minimun and maximum distances, respectively, between the ith and jth atoms in the distance matrix.

(3) The minimum and maximum distances between atoms i and l in a dihedral (1,4) relationship in a set of consecutively bonded atoms i, j, k, and l are determined next by using the relevant bond angles and the previously computed distances. For 1,4 substituents on double bonds, the 1,4 distances can be set precisely by using the standard bond lengths and angles for such systems together with specified configuration, cis or trans. These precise 1,4 distances act to preserve the configuration about double bonds during subsequent construction of a three-dimensional model.

For single and triple bonds, the minimum distance l_{il} and maximum distance u_{il} correspond to a dihedral angle φ of 0 and 180°, respectively, according to eq 2,

$$d_{il} = (p_1 - p_2\cos\varphi)^{1/2} \quad (2)$$

with

$$p_1 = d_{ij}^2 + d_{jk}^2 + d_{kl}^2 - 2d_{ij}d_{jk}\cos\theta_1 - 2k_{jk}d_{kl}\cos\theta_2 + 2d_{ij}d_{kl}\cos\theta_1\cos\theta_2$$

$$p_2 = 2d_{ij}k_{kl}\sin\theta_1\sin\theta_2$$

where θ_1 and θ_2 are the i,j,k and j,k,l bond angles, respectively.

1,4 Distances on appropriately substituted four-membered rings are determined by using the nonstandard bond angles for such systems. In the absence of any additional information on interatomic distances, all remaining minimum distances are set to 2.0 Å to insure that nonbonded atoms do not overlap. The corresponding maximum distances are set up $10n^{1/3}$ Å, where n is the number of atoms in the molecule.[17]

Hydrogen atoms are added automatically by the program to trisubstituted chiral centers and, as required, to 1,2-substituted double bonds whose cis or trans configuration has been specified. This is done to completely define the substitution pattern as part of the procedure to satisfy configurational constraints (see next section). Other hydrogen atoms are not explicitly represented by our program, although they certainly could be for other applications of our method.

For the example structure **1**, the distance matrix shown in Table II is obtained by the above procedure. Hydrogen atoms H-13–H-18, bonded to C-1–C-3, C-5, C-8, and C-10, respectively, have been added by the program.

For the purposes of this presentation, we continue with the distance matrix as specified in Table II. We note, however, that this procedure can make use of any geometrical information which can be translated into minimum, maximum constraints on interatomic distance, for example, torsional angle restrictions (e.g., from ^{1}H NMR coupling constants) or criteria on the spatial relationships of potential pharmacophores. Several examples of additional constraints have been described in a study of macromolecular conformations.[18]

The last step in refinement of the distance matrix consists of the exhaustive application of the triangle inequality, iterating until no further contraction of the (l_{ij},u_{ij}) intervals can be made.[11]

Determination of Initial Coordinates. Given the distance matrix, we use the method of Crippen et al. to determine an initial set of atom positions consistent with the set of interatomic distances (ranges). This problem is solved by computing first the matrix of the distances d_{io} to the center of mass O of the molecule (assuming unit mass for all the atoms) ([13], eq 3). Next the metric matrix G is determined by using the center of mass distances ([13], equation 4).

G is a real symmetric matrix, by definition, and therefore can be resolved into its eigenvalues and eigenvectors. A trial set of atomic coordinates can be computed from the first three largest eigenvalues and their corresponding eigenvectors.[13]

Truncation to the subspace with the three largest eigenvalues corresponds to the projection of the trial molecule into the three-dimensional Euclidean space. This is a very drastic simplification and usually leads to coordinates which violate the distance and configuration constraints.

Configuration and Distance Refinement. The most convenient method to bring back the computed coordinates within the distance and configuration constraints is the conjugate gradient procedure of Fletcher and Reeves.[20] The error function which is then to be minimized, E_{err}, comprises two terms (eq 3).

$$E_{err} = F + C \tag{3}$$

The first term F is a function of the distances only (eq 4).

$$F = \sum_{i<j}{}' (d_{ij}^2 - u_{ij}^2)^2 + \sum_{i<j}{}' (d_{ij}^2 - l_{ij}^2)^2 \tag{4}$$

the summation being made only for the distances outside the l_{ij}, u_{ij} intervals of the distance matrix.

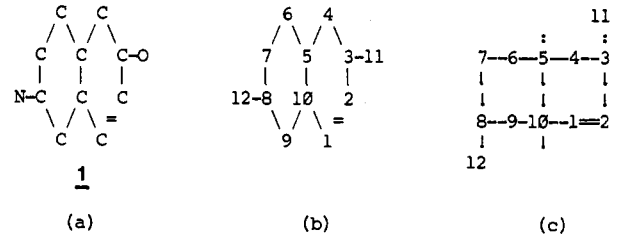


Figure 1. Teletype-oriented drawings illustrating the method of presentation of the constitution of **1** as both atom (a) named and (b) numbered drawings; (c) configuration of **1**. In (c) the colon symbol ":" is used to denote a substituent below the plane of the diagram, the exclamation mark "!" denotes substituents above the plane, and the equal sign "=" denotes cis stereochemistry at the double bond.[3]

The second term C of the error function is used to impose a given configuration on specified chiral centers (eq 5).[21] The

$$C = \sum_a (r_{ia}\cdot(r_{ja} \times r_{ka}) - v_a)^2 \tag{5}$$

C term is the sum of the squares of the differences between the volume of the parallelohedron defined by the vectors r from the central chiral carbon atom a to three of its neighbors i, j, and k and v_a, the oriented volume when the chirality and the bond angles are correct. The three neighbors i, j, and k are ordered so that the spanned volume is always positive. The summation is carried over all asymmetric carbon atoms.

In our initial trials, we found this error function inadequate. Although, strictly speaking, the configurations were preserved, chiral tetrahedral centers whose trial coordinates (see previous section) represented the incorrect configuration always ended up umbrella shaped, with the bond to the fourth substituent directed toward the plane defined by the three other substituents. This results because the procedure outlined above only takes into account three of the four substituents bonded to a chiral center. Therefore, in order to impart an equal weight to each of the four substituents, we extended the above procedure by taking into account all four possible triplets of four substituents about a chiral center. Including four contributions to the error function from each chiral center, rather than just one, has dramatically improved the procedure; it preserves the requisite stereochemistry and produces tetrahedral carbon atoms. An alternative aproach is to use the parallelohedron defined by the four substituents, leaving the correct placement of the central atom to be controlled by the distance constraints. In our trials this approach does not appear to be significantly better than the above, although we have only tried a few small (10–20 atom) molecules.

The conjugate gradient procedure continues until some terminating conditions are reached. Three cases are considered. (1) Success is achieved if the chirality and distance constraints are met within empirical error limits, id est when the error function drops below a preset value. (2) The situation is indeterminate if after some maximum number of iterations the gradient norm is "small" compared to the error function value; in this case, the minimum of the error function is ill-defined, which implies that the optimization procedure reached a "high" plateau. (3) If neither of the above conditions are reached, "failure" is reported.

Although X, Y, and Z coordinates are obtained under all these conditions, under conditions 2 and 3 some constraints are not fully satisfied. Failure to achieve success may be due to poor trial coordinates or a set of constraints which cannot be mutually satisfied. Note that failure to converge has been taken as presumptive evidence that geometric impossibilities exist in the initial distance matrix.

Once coordinates have been obtained, they can be displayed on suitable graphics terminals by using an interactive graphics

Table II. Distance Matrix Obtained for 1, After Setting 1,2 and 1,3 Distances and Establishing Minima and Maxima for 1,4 Distances

	C-1	C-2	C-3	C-4	C-5	C-6	C-7	C-8	C-9	C-10	O-11	N-12	H-13	H-14	H-15	H-16	H-17	H-18
C-1	0.00	1.34	2.48	3.76	2.50	3.86	26.21	3.86	2.50	1.52	3.66	26.21	1.08	2.10	3.36	3.44	26.21	2.15
C-2	1.34	0.00	1.52	2.50	3.76	26.21	26.21	26.21	3.76	2.48	2.41	26.21	2.10	1.08	2.15	26.21	26.21	3.36
C-3	2.48	1.52	0.00	1.54	2.51	3.88	26.21	26.21	26.11	2.86	1.43	26.21	3.47	2.26	1.09	3.46	26.21	26.21
C-4	2.72	2.50	1.54	0.00	1.54	2.51	3.88	26.21	3.88	2.79	2.43	26.21	6.21	3.51	2.16	2.16	26.21	3.46
C-5	2.50	2.72	2.51	1.54	0.00	1.54	2.51	3.88	2.51	1.54	3.77	26.21	3.51	26.21	3.46	1.09	26.21	2.16
C-6	2.56	2.00	2.57	2.51	1.54	0.00	1.54	2.51	3.88	2.51	26.21	3.81	6.21	26.21	26.21	2.16	3.46	3.46
C-7	2.00	2.00	2.00	2.57	2.51	1.54	0.00	1.54	2.51	3.88	26.21	2.46	6.21	26.21	26.21	3.46	2.16	26.21
C-8	2.56	2.00	2.00	2.00	2.57	2.51	1.54	0.00	1.54	2.51	26.21	1.47	6.21	26.21	26.21	26.21	1.09	3.46
C-9	2.50	2.72	2.00	2.57	2.51	2.57	2.51	1.54	0.00	1.54	26.21	2.46	3.51	26.21	26.21	3.46	2.16	2.16
C-10	1.52	2.48	2.86	2.51	1.54	2.51	2.57	2.51	1.54	0.00	26.21	3.81	2.26	3.47	26.21	2.16	3.46	1.09
O-11	2.67	2.41	1.43	2.43	2.53	2.00	2.00	2.00	2.00	2.00	0.00	26.21	6.21	3.41	2.07	26.21	26.21	26.21
N-12	2.00	2.00	2.00	2.00	2.00	2.54	2.46	1.47	2.46	2.54	2.00	0.00	6.21	26.21	26.21	26.21	2.10	26.21
H-13	1.08	2.10	3.47	2.00	2.62	2.00	2.00	2.00	2.62	2.26	2.00	2.00	0.00	2.42	26.21	26.21	26.21	3.12
H-14	2.10	1.08	2.26	2.62	2.00	2.00	2.00	2.00	2.00	3.47	2.57	2.00	2.42	0.00	3.12	26.21	26.21	26.21
H-15	2.56	2.15	1.09	2.16	2.45	2.00	2.00	2.00	2.00	2.00	2.07	2.00	2.00	2.43	0.00	26.21	26.21	26.21
H-16	2.44	2.00	2.45	2.16	1.09	2.16	2.45	2.00	2.45	2.03	2.00	2.00	2.00	2.00	2.00	0.00	26.21	3.06
H-17	2.00	2.00	2.00	2.00	2.00	2.45	2.16	1.09	2.16	2.45	2.00	2.10	2.00	2.00	2.00	2.00	0.00	26.21
H-18	2.15	2.56	2.00	2.45	2.16	2.45	2.00	2.45	2.16	1.09	2.00	2.00	2.43	2.00	2.00	2.27	2.00	0.00

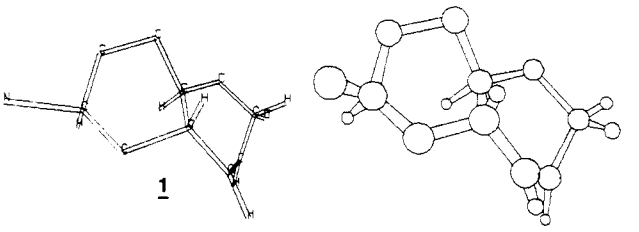


Figure 2. Structure drawings for **1** obtained from the *X*, *Y*, and *Z* coordinates determined by BUILD3D. (a) A perspective line drawing with atom names included; (b) NAMOD[22] ball-and-stick presentation of the molecule in the same orientation.

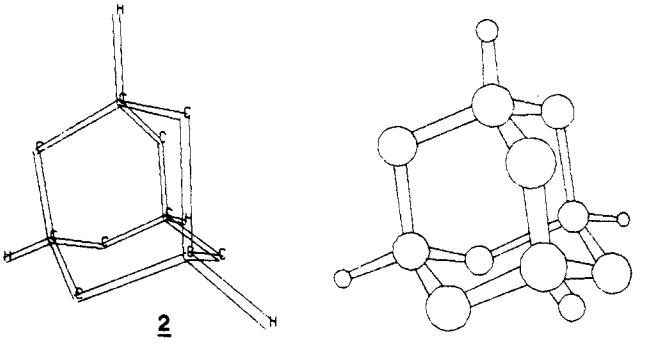


Figure 3. Structure drawings for adamantane (**2**) as perspective line drawing and ball-and-stick figure. The apparent difference in viewing angle is due to the fact that the perspective view represents a true perspective (from a viewing distance of about 5 Å) while the NAMOD drawing achieves a pseudoperspective effect by drawing the atoms and bonds in the foreground larger than those in the background.

package which is part of the BUILD3D program. BUILD3D provides several options for manipulation of the vector drawing including rotations, perspective views, and other utilities. Once an interpretable orientation of the structure has been obtained, it can be plotted offline for subsequent review. Drawings obtained for **1** are shown in Figure 2. Drawing a is the initial presentation of the structure as a line drawing with perspective. Here the trans stereochemistry of the ring junction and the orientation of the hydroxyl and amino substituents are obvious. Often, however, more complex structures are difficult to perceive by using this presentation, so we have included an alternative method of presentation using the NAMOD program of Beppu.[22] This program produces a ball-and-stick drawing with hidden line removal. The result for 1 is drawing b (Figure 2). Note that the conformation of the decalin ring system is not what most chemists would perceive to be "ideal". Because the coordinates are available, however, they can be passed to other programs for subsequent refinement as discussed previously. Additional examples are presented in the Results section.

RESULTS AND DISCUSSION

In this section we present several examples which indicate the kinds of drawings obtained from BUILD3D for molecules of varying stereochemical complexity. These examples have been chosen to present a small sample of the variety of conformations which can be obtained from different types of ring systems.

Consider first the three-dimensional structure for adamantane (**2**) derived from its connection table. For such a highly fused ring system it turns out that a satisfactory structure can be obtained by BUILD3D without specification of the chirality of the ring juncture atoms. However, the structures for **2** presented in Figure 3 were derived by using configuration designations which specified the normal stereochemistry of the ring junctures. Under even our simple

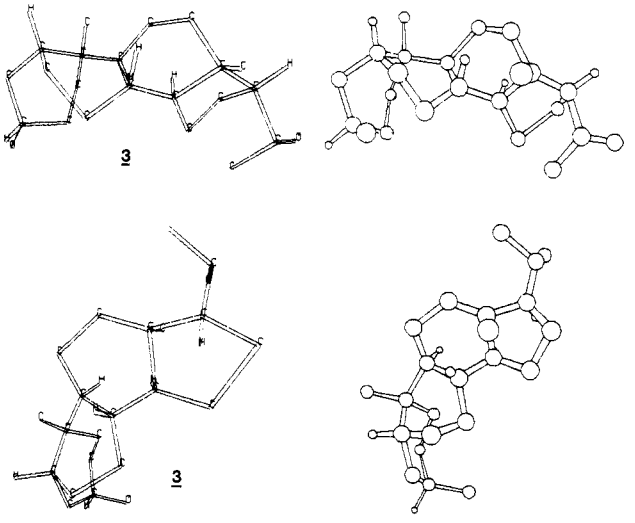


Figure 4. Structure drawings for 3α-hydroxy-5β-pregnan-20-one (**3**). The two pairs of drawings represent two selected conformations by BUILD3D.

distance constraints there is little conformational freedom possible in such a system.

As an intermediate case, two pairs of structure drawings for 3α-hydroxy-5β-pregnan-20-one (**3**) are presented in Figure 4. Each pair represents a different conformation obtained from BUILD3D. Note that in both pairs the specified stereochemistry is retained, but there are considerable differences in conformation, particularly in rings A and D and, of course, the C-17 side chain.

When a structure possesses even more conformational freedom, such as a macrocyclic ring or an acyclic chain, the few distance and configuration constraints imposed on the method from just the connection table allow considerable leeway in the conformations of the structures produced by BUILD3D. An example of the conformational freedom allowed is presented in Figure 5, using two selected conformations of the structure of erythronolide (**4**), the aglycon portion of erythromycin. These conformations are so different that further comment is unnecessary.

In conclusion, there are several attractive features of this approach. First, *X*, *Y*, and *Z* coordinate data can be obtained from simple connection tables even in the absence of configurational information. If stereochemical configurations are available, then the method produces more reasonable structures in which the configurations are preserved. The method admits any additional distance constraints from other sources of data (e.g., ^{1}H NMR[21]) to limit further the conformational variation in the structures produced. Finally, a coordinate representation of structure is a useful result because of the many techniques available for further analysis of such structures.

The method has important limitations. The fact that the computations begin with placement of atoms at random distances from one another (within the distance ranges specified in the distance matrix)[11] means that a random conformation will be obtained (if *any* conformation can be obtained). Thus the method yields only a random sampling of the conformational space and is ill-suited to a systematic exploration of possible conformations. Another important limitation is that there is no energetic information in the procedure. It is in fact possible to construct "in–out" bicyclic systems[3] in small bridged ring systems which obey all distance and configuration constraints. It is even possible to construct "in–in" bicyclic systems (i.e., both substituents at the ring juncture directed to the center of the ring) even though, formally, such systems may possess crossing bonds. Thus, methods for geometry optimization are required to obtain chemically reasonable conformations (or to decide that a structure is so highly strained as to be chemically unreasonable).

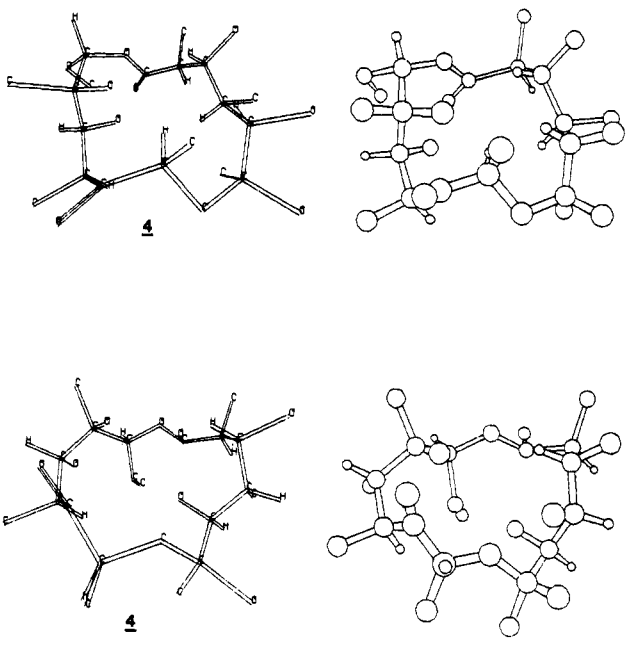


Figure 5. Two pairs of drawings of **4**, each pair representing a different analysis by BUILD3D, to illustrate the variety of conformations, each meeting all constraints, obtained for flexible molecules.

EXPERIMENTAL SECTION

The BUILD3D program is written in FORTRAN, with part of the ring perception algorithm in MACRO10. It runs on a Digital Equipment Corporation PDP-10 at the SUMEX Computer Resource at Stanford University. Execution times for 20–30 atom structures range from 20 to 30 s of central processor time on this machine. Guest access to the program at SUMEX, operating in conjunction with other DENDRAL Project software, is possible via a nationwide computer network. Please contact D. H. Smith for further information.

ACKNOWLEDGMENT

We thank Gordon Crippen, University of Texas, and Irwin Kuntz and Timothy Havel, University of California, San Francisco, for providing us with a copy of their program (EMBED) for carrying out the distance geometry calculations. Special thanks to Timothy Havel for discussions on the revised method for configurational constraints. We thank the Biotechnology Resources Branch, National Institutes of Health, for their generous financial support of this work (RR-00612) and of the SUMEX Resource where the programming was carried out (RR-0785). J.C.W. thanks the Stiftung fuer Stipendien auf dem Gebiete der Chemie (Basel, Switzerland) for their generous financial support.

REFERENCES AND NOTES

(1) R. E. Carhart, D. H. Smith, H. Brown, and C. Djerassi, *J. Am. Chem. Soc.*, **97**, 5755 (1975).
(2) R. E. Carhart, D. H. Smith, N. A. B. Gray, J. G. Nourse, and C. Djerassi, *J. Org. Chem.*, **46**, 1708 (1981).
(3) J. G. Nourse, D. H. Smith, R. E. Carhart, and C. Djerassi, *J. Am. Chem. Soc.*, **102**, 6289 (1980).
(4) (a) E. M. Engler, J. D. Andose, and P. v. R. Schleyer, *J. Am. Chem. Soc.*, **95**, 8005 (1973); (b) A. I. Kitaigorodsky, "Molecular Crystals and Molecules", Academic Press, New York, 1973, Chapter VII; (c) N. L. Allinger, *Adv. Phys. Org. Chem.*, **13** 1 (1976).
(5) (a) "Quantum Mechanics of Molecular Conformations", B. Pullman, Ed., Wiley, New York, 1976; (b) H. F. Schaeffer III in "Critical Evaluation of Chemical and Structural Information", D. R. Lide and M. A. Paul, Eds., National Academy of Sciences, Washington, DC, 1974, pp 591; (c) J. N. Murrell and A. J. Harget, "Semi-Empirical Self-Consistent-Field Molecular Orbital Theory of Molecules", Wiley-Interscience, London, 1972.
(6) R. E. Carhart, *J. Chem. Inf. Comput. Sci.*, **16**, 82 (1976).

(7) M. S. Gordon and J. A. Pople, Program MBLD, *QCPE*, **10**, 135 (1975).
(8) W. T. Wipke, in "Computer Representation and Manipulation of Chemical Information", W. T. Wipke, S. R. Heller, R. J. Feldmann, and F. Hyde, Eds., Wiley, New York, 1974, p 153.
(9) P. K. Weiner and S. Profeta, Jr., "Abstracts", 181st National Meeting of the American Chemical Society, Atlanta, GA, March 29–April 3, 1981, American Chemical Society, Washington, DC, 1981, COMP Division, Abstract No. 15.
(10) N. C. Cohen, P. Colin, and G. Lemoine, *Tetrahedron*, **37**, 1711 (1981).
(11) G. M. Crippen, *J. Comput. Phys.*, **24**, 96 (1977).
(12) G. M. Crippen, *J. Comput. Phys.*, **26**, 449 (1978).
(13) G. M. Crippen and T. F. Havel, *Acta Crystallogr., Sect A*, **A34**, 282 (1978).
(14) J. G. Nourse, *J. Am. Chem. Soc.*, **101**, 1210 (1979).
(15) J. G. Nourse, R. E. Carhart, D. H. Smith, and C. Djerassi, *J. Am. Chem. Soc.*, **101**, 1216 (1979).
(16) *Spec. Publ.—Chem. Soc.*, **No. 11** (1958); **No. 18** (1965).
(17) This formula has been used for test runs of the EMBED program. See Acknowledgment for further information.
(18) T. F. Havel, G. M. Crippen, and I. D. Kuntz, *Biopolymers*, **18**, 73 (1979).
(19) T. M. Dyott, "Utilization of Stereochemistry and Other Aspects of Computer-Assisted Synthetic Design", Ph.D. Thesis, Princeton University, 1973.
(20) R. Fletcher and C. M. Reeves, Comput. J., **7**, 149 (1964).
(21) W. Braun, C. Bosch, L. R. Brown, N. Go, and K. Wuthrich, *Biochim. Biophys. Acta*, **67**, 377 (1981).
(22) Y. Beppu, *QCPE*, **11**, 370 (1978).

Exhaustive Generation of Structural Isomers for a Given Empirical Formula—A New Algorithm

SI-YU ZHU* and JIN-PEI ZHANG
Institute of Elemento-Organic Chemistry, Nankai University, Tianjin, People's Republic of China

Received July 20, 1981

A new algorithm is presented for generating an exhaustive, irredundant list of structural isomers that are consistent with a given empirical formula. The new algorithm is simpler than existing ones, and the required program can be executed on many mini/microcomputers.

INTRODUCTION

Structural isomerism is a classical problem in organic chemistry, and the enumeration of such isomers has received considerable attention.[1,2] Yet, despite these efforts, it is still not possible to determine the number of structural isomers that are possible for an arbitrary empirical formula C_nH_m.

In recent years Smith and co-workers[3] have developed a computer program which aids considerably in this task for molecules which are not too large. However, their algorithm is rather complex and their program, CONGEN, is written in a generally unfamiliar language, INTERLISP. In this paper we described an alternative algorithm. The resulting program, ISOGEN, which is written in FORTRAN-IV can be executed on many mini/microcomputers.

RESULTS

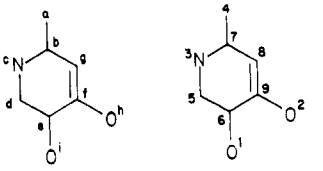
Representation of Structures. Following previous precedent, we have elected to represent molecular structures through the convenience of a connection matrix. Thus, a structure having n nonhydrogen atoms can be represented by a $n \times n$ connection matrix (CM) in which the ith row and the ith column correspond to the ith atom, and the entries CM(i,j) and CM(j,i) represent the bond connecting the ith and jth atoms, while single, double, and triple bonds are represented, respectively, by the digits 1, 2, and 3 and no bond by 0. For example, structure **1b** is represented by the following CM matrix:

	1	2	3	4	5	6	7	8	9
1	0	0	0	0	0	1	0	0	0
2	0	0	0	0	0	0	0	0	1
3	0	0	0	0	1	0	1	0	0
4	0	0	0	0	0	0	1	0	0
5	0	0	1	0	0	1	0	0	0
6	1	0	0	0	1	0	0	0	0
7	0	0	1	1	0	0	0	1	0
8	0	0	0	0	0	0	1	0	2
9	0	1	0	0	0	0	0	2	0

Since it is well-known that there are $n!$ permutations for a structure containing n nonhydrogen atoms, a unique representation under such a system depends upon the ability of the algorithm to designate individual atoms uniquely. Following the examples of several previous authors who employed numbering schemes to effect this end,[4,5] we have chosen as our defining algorithm a numbering sequence based in part upon the earlier algorithm of Ugi and Shubert.[6] However, inasmuch as their method fails to provide a unique numbering for some molecules,[6] we have carried out several modifications. The resulting algorithm and its application to the generation of structural isomers is described below.

I. NUMBERING ALGORITHM

First Level. Sort all the nonhydrogen atoms of a given structure according to their atomic *type* (i.e., C, N, O, S, etc.). Although the ordering of the types of atoms is arbitrary, in this paper we have assigned precedent to atoms according to their respective valence, R. Thus, for example, O (R = 2) and S (R = 2) precede N (R = 3) and P (R = 3) which in turn precede C and Si (R = 4) and so on. Atoms which are dissimilar in type but which have equivalent valences (for example, O and S) are ordered arbitrarily. Application of these first-level criteria to the numbering of structure **1a** is shown in Table I.



1a 1b

Second Level. Atoms which are the same in type (e.g., all carbon atoms) require a second level of numbering, the criteria for which is based upon the *number* and *type* of nonhydrogen atom bonds connected to such atoms. If we allow Vnh to represent the sum of orders of bonds connected to adjacent nonhydrogen atoms, employing the integral values 1 for single bond, 2 for double, and 3 for triple, the resulting index can be used to order these atoms. Atoms with the smaller Vnh index precedes that with a larger one; thus, according to this criterion, atom g in **1a** precedes atom f [Vnh(g) = 3 < Vnh(f) = 4]. However, this still results in some ambiguity.

To further distinguish those atoms which have identical R and Vnh indexes (e.g., atom b and g), it is necessary only to

0095-2338/82/1622-0034$01.25/0