Computer Simulation of Physical-Chemical Properties of Organic Molecules. 1. Molecular System Identification

G. KLOPMAN* and M. McGONIGAL

Chemistry Department, Case Western Reserve University
Cleveland, Ohio 44106

Received August 4, 1980

A computer program which generates molecular geometries and performs molecular mechanical calculations for organic molecules is presented. A linear coding methodology used for inputting the molecules is also described. This is the foundation of a program aimed at the simulation of the physical-chemical properties of organic molecules.

The organic chemist is often confronted with the need to retrieve physical-chemical information of known compounds. This need has traditionally been met by the publication of data tables in which these chemical properties are tabulated in a more or less easily retrievable form. More recently, these properties have been stored in computer files that can be updated and maintained as, for example, in the CAS registry or the Cambridge Crystallographic Data File[1] or the UPDATE information system described by Razinger.[2]

With our current knowledge of the factors responsible for observed chemical properties, it is often possible "a priori" to calculate or sometimes simulate these properties with a reasonable degree of accuracy.[3] For some, this task can be performed merely by examining the functional groups present in the molecule (e.g., IR spectra, NMR spectra, heats of formation, mass spectra, log P, etc.); for others it is necessary to perform sophisticated quantum mechanical calculations (e.g., UV spectra, dipole moments, heats of formation). The growing use of computers in our world and other current trends seem to indicate that, in the future, simulated properties may be used in place of experimentally determined ones in many instances, and this simulation will undoubtedly be automated. Although such techniques would never replace nor attain the accuracy of the experimental data files, they nevertheless offer some definite advantages over them. Indeed, simulation procedures can be used for predicting the properties of yet unknown compounds and, in this respect, can be enormously valuable in determining directions for future research. Coupled with modern instrumentation, these techniques can also become an invaluable tool for the chemist interested in structure elucidation of unknown chemicals. For example, the spectrum of a postulated structure can be drawn by the computer and compared with the observed one. The comparison may provide a clue as to whether the postulated structure is indeed that of the unknown. Such techniques can also be automated, and attempts in this direction are now beginning to appear in the literature.[4]

It is thus becoming increasingly apparent that the possibility now exists to write such a cluster of interacting programs capable of simulating many chemical properties. We have undertaken the design of such a program and wish to report here the results of our initial efforts.

The program logic is represented in Figure 1.

The program consists of three parts. The "driver" directs the computer into a logical exploitation of the subroutines. It is connected to the Molecular System Development (MSD) program which is a cluster of three programs that provide complete knowledge of the geometry, stability, and charge distribution of the molecular entity. The decoder accepts the code, interprets it, and returns the connectivity matrix. The latter is input into the model builder which returns the coordinates in space of all atoms. Finally, the quantum mechanical program uses these coordinates to determine molecular orbitals, charge densities, and other useful properties. In the property simulator section, various property simulators calculate prescribed properties by using the MSD results and ad hoc tables.

Table I. Character Codes

(a)		(b)		(c)	
atom	letter code	group	letter code	numbers	code
H	H	CH	D	1	1
C	C	CH_2	R	2	2
F	F	CH_3	M	3	3
Cl	G	NH	E	4	4
Br	B	NH_2	A	5	5
I	I	OH	K	6	6
O	O	SH	J	7	7
S	S	NO	U	8	8
N^{3+}	N	NO_2	X	9	9
N^+	Q	SO	V	10	(
P^{3+}	P	SO_2	Y	11	*
P^{5+}	Z	CO	T	12	=
		PO (P^{5+})	W	13	$
		PO (P^{7+})	L	14	'
				15	,
				16	ƀ

In order to achieve these goals, there is a need for a simple and adequate method for computer identification of compounds, i.e., a molecular coding method suitable for use with data processing equipment. Such an endeavor has already attracted the attention of other authors. Examples are the Wiswesser line formula notation,[5] ADAPT,[6] and MOLY[7] systems. However, Wiswesser's notation was designed for data retrieval; each molecule had to have a unique code, making coding more difficult than necessary for our purposes. On the other hand, the graphic entries such as used in ADAPT and MOLY are not general enough and often require special equipment. Therefore, we designed our own coding method, such that a molecule may be input a number of ways, but will still be perceived by the program as representing the same molecular structure.

In this paper, we describe such a coding method as well as a model builder[8] that determines the geometry of the molecule from its code and the interface with a quantum mechanical program, i.e., MINDO/3, that calculates charge densities, dipole moments, heats of formation, and molecular orbital energies.

Coding of Molecules. In order to input molecules into the program, we have designed a linear coding method by which both simple molecules and complex molecules with rings and branched chains may be entered. Letters of the alphabet are used to denote atoms of the different elements and also some of the most commonly encountered organic groups. These atoms and groups are shown in Table Ia,b.

A linear molecule is entered merely by making a string of the character codes of the atoms or groups of which the molecule is composed in the order in which they are bonded.

0095-2338/81/1621-0048$01.25/0 © 1981 American Chemical Society

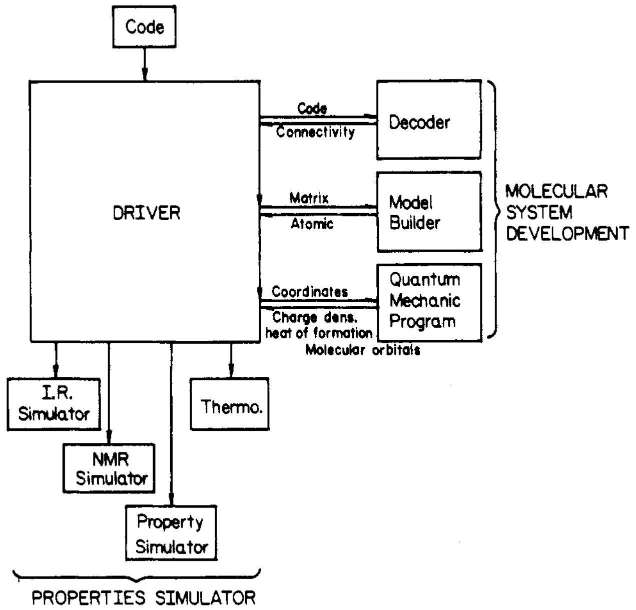
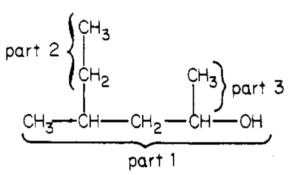


Figure 1. Program logic.

For example, pentane can be entered as MRRRM where M represents a methyl group and R a CH_2 group but not as MRMRR.

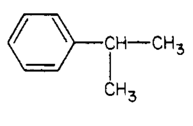
The program interprets all bonds in a molecule to be single bonds unless they have been otherwise specified. For a double or triple bond specification the numeral 2 or 3 must be placed between the respective atoms or groups. For example, acetylene can be entered as D3D or as HC3CH.

In order to enter more complicated molecules, it is first necessary to understand the logic upon which the interpretation of the bonding of a molecule is based. The program interprets an atom or group to be bonded to the previous atom or group unless all its valences have been filled. In this case the group is bonded to the first open valence that was left in the character string. This can be illustrated by the code of the following molecule:



One enters the main chain (part 1), followed by the branch to the first open valence (part 2), and then the branch to the next first open valence (part 3): MDRDK, part 1; RM, part 2; M, part 3; thus yielding MDRDKRMM. Alternate notations are MDRMRDMK, MDKRDMRM, etc. The program will print an incorrect coding message if the valencies have not been properly filled.

To enter a ring, the special character / or) is used. The character / indicates that the atom or group which it follows is bonded both to the atom immediately preceding it and the first open valence that was left in the character string. The character) indicates, in a similar fashion, that the atom or group which it follows is bonded to both the atom immediately preceding it and the last open valence that was left in the character string. For example, benzene can be coded as D2DD2DD2D) and the molecule



as MDC2DD2DD2D)M. If the parenthesis was replaced by a slash, i.e., MDC2DD2DD2D/M, the molecule would have the structure

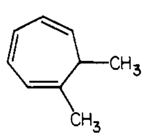


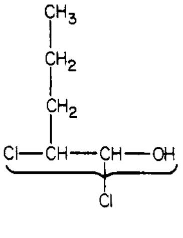
Table II. Special Character Codes

code	significance
/	close ring with first open valence
)	close ring with last open valence
*	monosubstituted phenyl ring
(n	monosubstituted aliphatic ring of size *n*
= n	linking aliphatic chain of *n* carbon atoms
= nƀ	terminal aliphatic chain of *n* carbon atoms

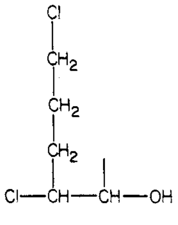
For further facilitation of data input, the program contains three functions designed to simplify the input of benzene rings, aliphatic rings, and aliphatic chains. The monosubstituted phenyl ring may be entered by using an asterisk followed by the string that is to represent the substituent. For example, the previous example can be shortened by recoding it as *DMM.

To enter an aliphatic ring with one hydrogen replaced, the character (followed by a character code corresponding to the number of carbons (Table Ic) in the ring followed by the string replacing the hydrogen can be used. Using this method, methylcyclohexane can be represented by (6M. Rings containing less than three carbons are unacceptable.

To enter an aliphatic chain, the special character = followed by the character code indicating the number of carbons (Table Ic) may be used. Although the maximum number of carbons is 16, this function may be used sequentially if more are needed. If this function is positioned first or last in the character string, the first or last aliphatic group, respectively, will be interpreted as a methyl group. If it is positioned in the middle of the character string, there are two possible situations. The function may be followed by one blank and then the rest of the character string, in which case the last group in the chain will be interpreted as a methyl group indicating the end of a branch, or the function may be used with no intervening space, causing all groups in the chain to be methylene groups. As an example $Cl - CH_2 - CH_2 - Cl$ may be input as G=2G and



may be input as GDDK=3 G. If the latter code were to be input without the space, the molecule would be interpreted as

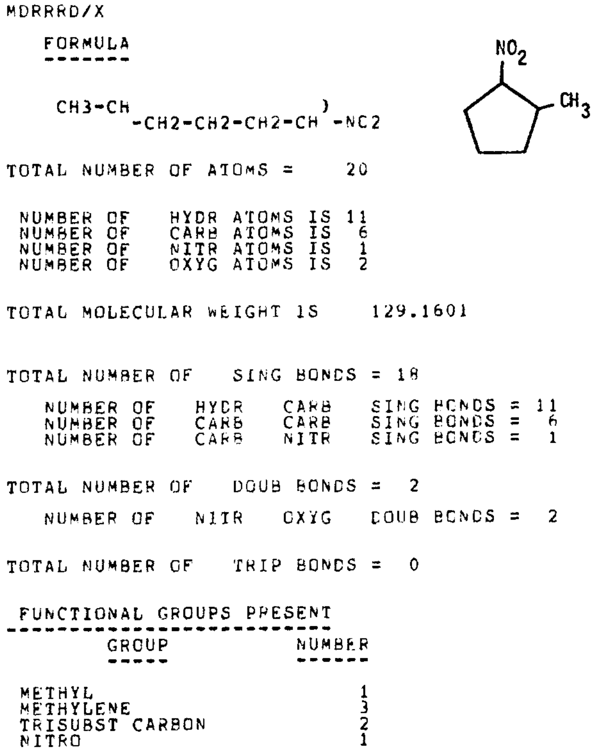


and an incorrect coding message would be received. Table II lists the active special characters.

Table III. Connectivity Matrix for 2-Butanone[a]

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1			1									
2		1		1									
3			1	1									
4	1	1	1	6			1						
5					1		1						
6						1	1						
7				1	1	1	6		1				
8								8	2				
9							1	2	6				1
10										1			1
11											1		1
12												1	1
13									1	1	1	1	6

[a] All blank spaces are zeros.



```
MDRRRD/X

   FORMULA
   -------

    CH3-CH                    )
          -CH2-CH2-CH2-CH -NC2

TOTAL NUMBER OF ATOMS =     20

 NUMBER OF   HYDR ATOMS IS 11
 NUMBER OF   CARB ATOMS IS  6
 NUMBER OF   NITR ATOMS IS  1
 NUMBER OF   OXYG ATOMS IS  2

TOTAL MOLECULAR WEIGHT 1S    129.1601


TOTAL NUMBER OF   SING BONDS = 18

   NUMBER OF   HYCR   CARB   SING HCNDS = 11
   NUMBER OF   CARB   CARB   SING BONDS =  6
   NUMBER OF   CARB   NITR   SING BCNDS =  1

TOTAL NUMBER OF   DOUB BONDS =  2

   NUMBER OF   NIIR   OXYG   COUB BCNDS =  2

TOTAL NUMBER OF   TRIP BONDS =  0

 FUNCTIONAL GROUPS PPESENT
---------------------------
        GROUP          NUMBER
        -----          ------

 METHYL                   1
 METHYLENE                3
 TRISUBST CARBON          2
 NITRO                    1
```

Figure 2. Sample output for 2-methylnitrocyclopentane.

Decoding the Input String. The decoding process is performed in two stages. The first step is the identification of each character in the input string and the creation of an array of numbers corresponding to the characters in this string. This array consists of five-digit numbers containing the following information about each group: the atomic number of the main atom, the number of open valencies, the number of hydrogens, and the number of oxygens bonded to the main atom within the group.

The second step is the interpretation of the array created in the previous step to create a connectivity matrix representing all the bonds in the molecule. This matrix can be easily referenced later in the program to find the information necessary to calculate various molecular properties. An example of such a matrix is the matrix for 2-butanone, coded in as MRTM; the elements of this matrix are shown in Table III. The rows and columns in the matrix represent all the atoms in the molecule. The atoms are numbered in the order that they were coded in, the main atom always being numbered last within a group. Each diagonal element $L(I,I)$ in the matrix holds the atomic number of atom I. Each nondiagonal element $L(I,J)$ in the matrix holds the multiplicity of the bond between atoms I and J; this is zero if there is no bond.

Table IV. Standard Bond Lengths, Å

		Single Bonds					
H–H	0.74	C–B	1.56	O–F	1.42	B–P	1.8
H–C	1.09	C–Si	1.86	O–B	1.36	B–S	1.61
H–N	1.00	C–P	1.87	O–Si	1.64	B–Cl	1.72
H–O	0.96	C–S	1.81	O–P	1.61	Si–Si	2.30
H–F	0.92	C–Cl	1.71	O–S	1.44	Si–P	2.20
H–B	1.21	N–N	1.46	O–Cl	1.70	Si–S	2.15
H–Si	1.48	N–O	1.36	F–F	1.42	Si–Cl	2.03
H–P	1.42	N–F	1.36	F–B	1.29	P–P	2.21
H–S	1.34	N–B	1.42	F–Si	1.56	P–S	1.86
H–Cl	1.27	N–Si	1.57	F–P	1.54	P–Cl	2.03
C–C	1.53	N–P	1.49	F–S	1.58	S–S	2.04
C–N	1.47	N–S	1.30	F–Cl	1.63	S–Cl	1.99
C–O	1.43	N–Cl	1.75	B–B	1.59	Cl–Cl	1.99
C–F	1.33	O–O	1.48	B–Si	1.90		
		Double Bonds					
C=C	1.34	C=S	1.71	N=S	1.20	P=P	1.89
C=N	1.29	N=N	1.25	O=O	1.27	P=S	1.70
C=O	1.22	N=O	1.14	O=P	1.45	S=S	1.90
C=P	1.70	N=P	1.30	O=S	1.30		
		Triple Bonds					
C≡C	1.20	C≡N	1.17	N≡N	1.10		

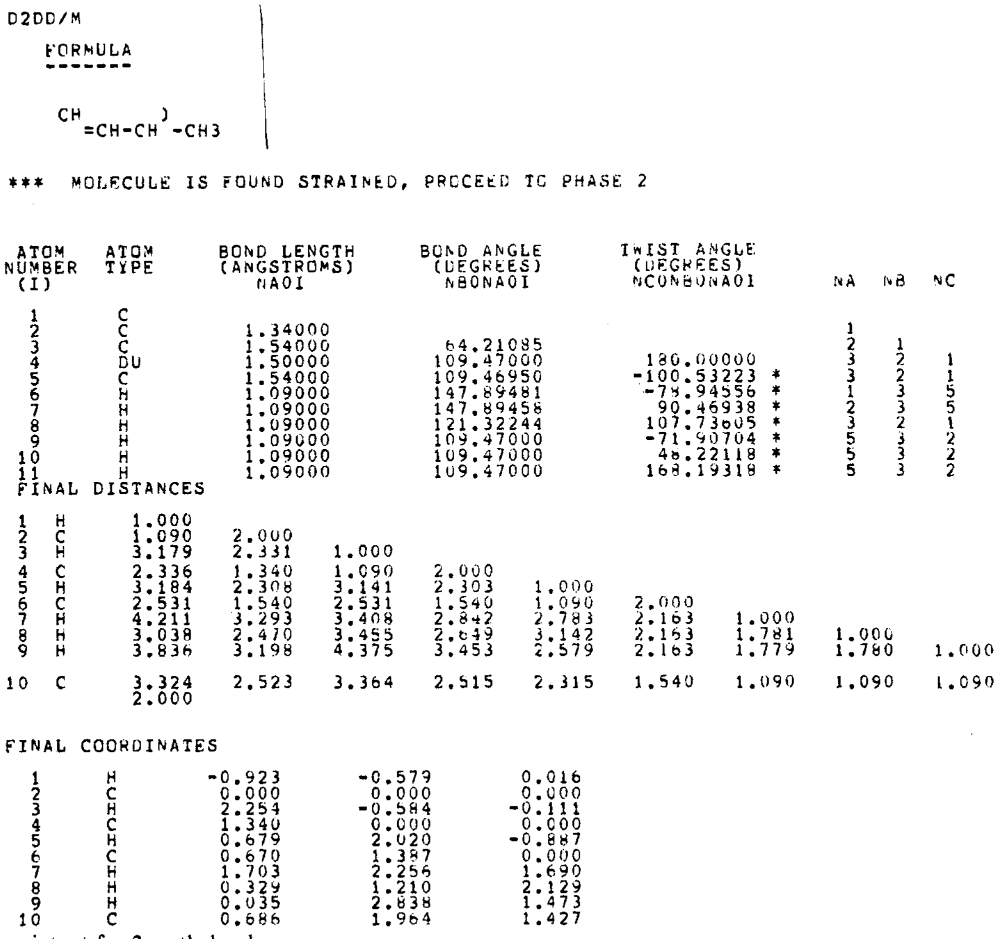
Table V. Standard Bond Angles, Degrees

	single bond	double bond	triple bond
H	109.47	(120.0)	(180.0)
C	109.47	120.0	180.0
N	108.00	128.0	180.0
O	109.00	120.0	(180.0)
F	109.47	(120.0)	180.0
B	120.00	120.0	180.0
Si	109.50	120.0	180.0
P	99.60	120.0	180.0
S	104.60	120.0	180.0
Cl	109.47	(120.0)	(180.0)

At the end, the program prints out the formula of the molecule. This is useful to verify that the correct molecule has been input. A modified version of the conventional molecular formula is used due to limitations imposed by the printer. The symbols of the elements are composed of all capital letters, and numerical subscripts appear as regular numbers. Single, double, and triple bonds between groups appear as the symbols —, =, and $, respectively, and all rings are completed by the symbol). Branches and rings off the main chain appear on different lines from the main chain. A sample formula output is shown in Figure 2 together with other properties identified in the preliminary examination of the molecule MDRRRD/X by the decoder.

Limitations of the Coding Method. The coding method was tested extensively and found to be very convenient for common organic molecules. The method was not designed for data retrieval, and there was thus no attempt to design a unique formulation as in the Wiswesser notation. Its advantage, for our purpose, is that the coding remains very flexible. As stated above, this is due to the fact that useless "chemical" information does not have to be carried by the formulation to make it unique. However, even here, the procedure becomes somewhat cumbersome for large highly branched and polycyclic molecules. Furthermore, as in previous methods, the problem of stereochemistry still remains unresolved.

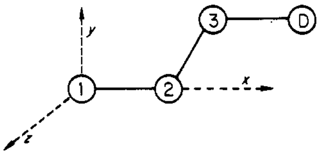
Model Builder. Once the connectivity matrix of the molecule has been established, a molecular geometry is automatically generated by the model builder. Standard bond lengths and bond angles (Tables IV and V) and reasonable twist angles are initially used to represent the heavy atom (i.e., nonhydrogen atoms) skeleton of the molecule.



```
D2DD/M
   FORMULA
   -------

    CH          )
       =CH-CH  -CH3

***  MOLECULE IS FOUND STRAINED, PROCEED TO PHASE 2


 ATOM    ATOM        BOND LENGTH         BOND ANGLE          TWIST ANGLE
NUMBER   TYPE       (ANGSTROMS)          (DEGREES)           (DEGREES)
 (I)                   NAOI               NBONAOI            NCONBONAOI          NA   NB   NC

  1       C
  2       C          1.34000                                                     1
  3       C          1.54000            64.21085                                 2    1
  4       DU         1.50000           109.47000            180.00000            3    2    1
  5       C          1.54000           109.46950           -100.53223  *         3    2    1
  6       H          1.09000           147.89481            -78.94556  *         1    3    5
  7       H          1.09000           147.89458             90.46938  *         2    3    5
  8       H          1.09000           121.32244            107.73605  *         3    2    1
  9       H          1.09000           109.47000            -71.90704  *         5    3    2
 10       H          1.09000           109.47000             48.22118  *         5    3    2
 11       H          1.09000           109.47000            168.19318  *         5    3    2
 FINAL DISTANCES

 1  H      1.000
 2  C      1.090    2.000
 3  H      3.179    2.331    1.000
 4  C      2.336    1.340    1.090    2.000
 5  H      3.184    2.308    3.141    2.303    1.000
 6  C      2.531    1.540    2.531    1.540    1.090    2.000
 7  H      4.211    3.293    3.408    2.842    2.783    2.163    1.000
 8  H      3.038    2.470    3.455    2.649    3.142    2.163    1.781    1.000
 9  H      3.836    3.198    4.375    3.453    2.579    2.163    1.779    1.780    1.000

10  C      3.324    2.523    3.364    2.515    2.315    1.540    1.090    1.090    1.090
           2.000

FINAL COORDINATES

  1     H        -0.923       -0.579        0.016
  2     C         0.000        0.000        0.000
  3     H         2.254       -0.584       -0.111
  4     C         1.340        0.000        0.000
  5     H         0.679        2.020       -0.887
  6     C         0.670        1.387        0.000
  7     H         1.703        2.256        1.690
  8     H         0.329        1.210        2.129
  9     H         0.035        2.838        1.473
 10     C         0.686        1.964        1.427
```

Figure 3. Sample printout for 2-methylcyclopropene.

The molecular axes are defined with respect to the first three heavy atoms of the molecule and a dummy, placed as represented below:



The molecular geometry is then optimized by treating all twist angles as parameters in a nonlinear minimization routine of a function E (eq 1) which tends to reconcile the postulated geometry with a set of target interaction distances deduced from the bonding constraints. The first term of this function,

$$E = \mathrm{EC} + \mathrm{EA} \tag{1}$$

EC, is a weighted summation of the deviation of the constraint distances in the initial geometry from their target values.

$$\mathrm{EC} = \sum_{A}\sum_{B} \mathrm{FACTOR} \times (r_{AB}^{\mathrm{real}} - r_{AB}^{\mathrm{target}})^2$$

The target interaction distances result from the consideration of any bonding or angular constraints that apply to the molecule. These constraints are of three types: (1) distances between two atoms that are bonded, (2) distances between two atoms that are bonded to a common atom, i.e., distances that are imposed by bond angles, and (3) distances between two atoms that are attached to two atoms that are multiply bonded (cis–trans effect). A weighing factor, FACTOR, is used since some constraints are more important than others; constraints are more important when the distances are small, i.e., when the atoms are bonded, and not as important when the distances are large, i.e., when the atoms are constrained to be coplanar due to multiple bonds.

Weighing Factors. If $r_{AB}^{\mathrm{target}} < 1.75$ Å, FACTOR = 38 (e.g., bond distances). If 1.75 Å $\leq r_{\mathrm{AB}}^{\mathrm{target}} \leq$ 2.6 Å, FACTOR = 8 (e.g., bond angles). If $r_{\mathrm{AB}}^{\mathrm{target}} > 2.6$ Å, FACTOR = 3 (e.g., planarity of multiple bonds).

The remaining term of the function E, i.e., EA, considers the distances between atoms for which no constraints exist; EA maximizes the distances between these atoms in order to minimize the repulsive forces between them.

$$\mathrm{EA} = \sum_{A}\sum_{B} \frac{1}{(r_{AB}^{\mathrm{real}} + 1)^2}$$

The minimization of E is performed by the Fletcher Powell method described in Dewar's MINDO/3 program.

After minimization, the model builder checks the value of EC. If EC is less than 0.01, it is assumed that an appropriate geometry has been found, and the program proceeds to add the hydrogen atoms. If EC is large, there is strain in the molecule that cannot be relieved by varying the dihedral angles only. This occurs in strained ring systems such as cyclopropane. In these cases, the bond angles become variables also and the function E is again minimized. Once this is done, the target distances are reset to their final values. The hydrogen atoms are included in the minimization process and the program once again minimizes the function E, this time including the dihedral angles for the hydrogen atoms. However, some elements of symmetry are now considered. For example, in an unstrained molecule containing a methyl group, only one of the methyl hydrogen dihedral angles needs to be minimized since its position dictates the position of the other two methyl hydrogens.

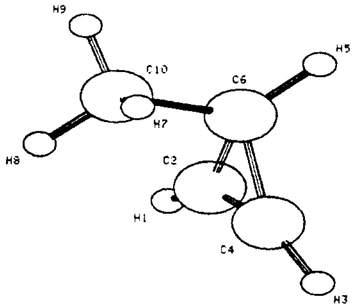


Figure 4. PLUTO plot of 3-methylcyclopropene.

After this step a reasonable geometry for most molecules is obtained. A sample printout for D2DD/M, i.e., 3-methylcyclopropene, is shown in Figure 3.

Plotting of Molecular Structures. After the geometry has been calculated, the program prepares a plotting file so that a three-dimensional ball-and-stick plot of the molecule may be obtained. The molecular plotting program used is a modified version of PLUTO, a program developed by Sam Matherwell at the University Chemical Laboratory at Cambridge, England. The PLUTO plot of 3-methylcyclopropene is shown in Figure 4.

Quantum Mechanics Program. The atomic coordinates generated by the model builder are transferred into a quantum mechanics program to obtain the molecular orbitals, charge densities, and heat of formation of the molecule. These results may then be used later for the calculation of other molecular properties. Currently the quantum mechanical method which is interfaced with the Molecular System Identification routine (MSI) is the MINDO/3 program.[9]

The program is currently executed interactively on a VAX-11 computer.[10] The complete calculation of a 20-atom molecule requires approximately 20 s of CPU time, including geometry minimization and SCF calculations.

In the succeeding papers of this series we will examine the property synthesizers, presently at various stages of development.

ACKNOWLEDGMENT

We thank the National Science Foundation (Grant No. Env 77-74061) and the National Cancer Institute (Contract No. NO1-CP-75927) for supporting our work.

REFERENCES AND NOTES

(1) S. R. Wilson and J. H. Huffman, *J. Org. Chem.*, **45**, 560 (1980).

(2) M. Razinger, J. Zupan, M. Penca, and B. Barlic, *J. Chem. Inf. Comput. Sci.*, **20**, 158 (1980).

(3) See for example (a) S. W. Benson, F. R. Cruickshank, D. M. Golden, G. R. Hangen, H. E. O'Neal, A. S. Rodgers, R. Shaw, and R. Walsh, *Chem. Rev.*, *69*, 279 (1969); (b) J. T. Chou and P. C. Jurs, *J. Chem. Inf. Comput. Sci.*, **19**, 3 (1979).

(4) (a) C. A. Shelley, H. B. Woodruff, C. R. Snelling, and M. E. Munk, *ACS Symp. Ser.* **No. 54**, 92 (1977); (b) H. B. Woodruff, C. R. Snelling, C. A. Shelley, and M. E. Munk, *Anal. Chem.*, **49**, 2075 (1977); (c) H. B. Woodruff and M. E. Munk, *J. Org. Chem.*, **42**, 1761 (1977).

(5) E. G. Smith, "The Wiswesser Line Formula Chemical Notation", McGraw-Hill, New York, 1968.

(6) See, for example, A. J. Stufer and P. C. Jurs, *J. Chem. Inf. Comput. Sci.*, **16**, 99 (1976).

(7) T. M. Dyott, A. J. Stuper, and G. S. Zander, *J. Chem. Inf. Comput. Sci.*, **20**, 28 (1980).

(8) For other model builder routines, see, for example, W. T. Wipke, S. R. Heller, R. J. Feldman, and E. Hyde in "Computer Representation and Manipulation of Chemical Information", Wiley, New York, 1974.

(9) R. C. Bingham, M. J. S. Dewar, and D. M. Lo, *J. Am. Chem. Soc.*, **91**, 1285 (1975).

(10) The program is available for use by others. For copies, contact G. Klopman.

Computer Perception of Topological Symmetry via Canonical Numbering of Atoms

MILAN RANDIĆ,*[1] GREGORY M. BRISSEY, and CHARLES L. WILKINS*

Department of Chemistry, University of Nebraska—Lincoln, Lincoln, Nebraska 68588

Received August 15, 1980

A previously described algorithm for perception of topological symmetry has been programmed for computer use. The algorithm is based on the concept of the smallest binary code for a graph and requires generation of all numberings for a structure which will produce the canonical form for the adjacency matrix. The unique adjacency matrix corresponds to the smallest possible (binary) number representing the structure when its rows are read from top to bottom and from left to right. Operation of the program is illustrated with a selection of polycyclic structures. Typically, molecules with a dozen carbon atoms and several rings produce results in a few hundred tests as compared with $N!$ which would be considered in an exhaustive procedure.

INTRODUCTION

Although the problem of recognition of symmetry for rigid bodies and rigid molecular frames is straightforward, the same task in the case of nonrigid structures and particularly for graphs in which only connectivity is specified is much more difficult. Even the case of constitutional symmetry for polycyclic structures projected in two dimensions may pose difficulties, since oblique projections frequently obscure equivalence of sites. In many problems, in particular in computer manipulations with structures, it is of interest to establish the symmetry of the structure and the equivalence of atoms in the structure. The problem has received attention in recent literature, and several alternative schemes have been advocated.[2] The problem, however, which is closely related to the problem of graph isomorphism, ordering of graphs, and graph construction, is sufficiently involved that it appears desirable to continue to pursue the subject and report on alternative schemes. One justification for such continuing interest is the lack of clear-cut evaluations of one method vs. another and the lack of knowledge about how close existing schemes approach an idealized optimal algorithm. Furthermore, it is entirely possible that certain approaches may be superior to others for various applications. Therefore we report here another approach to the determination of constitutional symmetry (or atom equivalence) in molecules. Our method is quite general and, in fact, is a means of determining the symmetry

0095-2338/81/1621-0052$01.25/0 © 1981 American Chemical Society