(14) K. L. H. Ting et al., *Science*, **180**, 417 (1973).
(15) S. A. Hiller et al., *Comput. Biomed. Res.*, **6**, 411 (1973).
(16) J. T. Clerc, P. Naegeli, and J. Seibl, *Chimia*, **27**, 639 (1973).
(17) B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.*, **96**, 916 (1974).
(18) C. L. Perrin, *Science*, **183**, 531 (1974).
(19) K. C. Chu, *Anal. Chem.*, **46**, 1181 (1974).
(20) R. O. Mathews, *J. Am. Chem. Soc.*, **97**, 935 (1975).
(21) K. C. Chu et al., *J. Med. Chem.*, **18**, 539 (1975).
(22) A. J. Stuper and P. C. Jurs, *J. Am. Chem. Soc.*, **97**, 182 (1975).
(23) L. A. Cox, Jr., R. H. Pritchard, and C. F. Bender, "RECOG: A Polyalgorithm for the Analysis of Generalized Data Sets. An Operator's Manual," UCID-16443, Rev. 1, Lawrence Livermore Laboratory, 1975.
(24) J. R. Koskinen and B. R. Kowalski, *J. Chem. Inf. Comput. Sci.*, **15**, 119 (1975).
(25) J. T. Tou and R. C. Gonzalez, "Pattern Recognition Principles", Addison-Wesley, Inc., Reading, Mass., 1974.
(26) W. S. Meisel, "Computer-Oriented Approaches to Pattern Recognition", Academic Press, New York, N.Y., 1972.
(27) N. J. Nilsson, "Learning Machines", McGraw-Hill, New York, N.Y., 1965.
(28) H. C. Andrews, "Introduction to Mathematical Techniques in Pattern Recognition", Wiley-Interscience, New York, N.Y., 1972.
(29) R. O. Duda and P. E. Hart, "Pattern Classification and Scene Analysis", Wiley-Interscience, New York, N.Y., 1973.
(30) P. C. Jurs and T. L. Isenhour, "Chemical Applications of Pattern Recognition", Wiley-Interscience, New York, N.Y., 1975.
(31) W. E. Brugger and P. C. Jurs, *Anal. Chem.*, **47**, 781 (1975).
(32) E. A. Sussenguth, *J. Chem. Doc.*, **5**, 36 (1965).
(33) G. S. Zander, Thesis, Department of Chemistry, The Pennsylvania State University, 1974.
(34) A. Bondi, *J. Phys. Chem.*, **68**, 442 (1964).
(35) J. E. Williams, P. J. Stang, and P. v. R. Schleyer, *Annu. Rev. Phys. Chem.*, **19**, 531 (1968).
(36) E. M. Engler, J. D. Andose, and P. v. R. Schleyer, *J. Am. Chem. Soc.*, **95**, 8005 (1973).
(37) G. S. Zander, A. J. Stuper, and P. C. Jurs, *Anal. Chem.*, **47**, 1085 (1975).

Generation of Descriptors from Molecular Structures

W. E. BRUGGER, A. J. STUPER, and P. C. JURS*

Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802

Received December 11, 1975

To apply pattern recognition methods to studies of structure–activity relations, the molecular structures must be decomposed into numerical descriptors. The descriptors generated are of two types: topological and geometrical. Topological descriptors include: fragments, which code the atom and bond types; substructure descriptors, which code the presence or absence of particular, explicitly defined, substructures; and environmental descriptors, which code the immediate surroundings of an atom center of interest. The geometrical descriptors are derived from a strain-energy minimized three-dimensional structure, and they code the shape and size of the molecule. These descriptors can be used in conjunction with pattern recognition programs to investigate relationships between molecular structure and biological activity.

INTRODUCTION

A major problem in any pattern recognition system is the generation of informative pattern vectors which permit the correct classification of the objects or system under investigation.

In chemical applications of pattern recognition, several types of input data can be used to generate pattern vectors. Early investigations used single source data such as mass spectra,[1-3] infrared spectra,[4,5] NMR spectra,[6,7] and stationary electrode polarograms[8] as input to the pattern recognition system. The object of these investigations was to determine chemical structure information or to elucidate information about a chemical system. In these studies, the transformation of the spectra into pattern vectors was accomplished by first digitizing the spectra and then selecting all or parts of each spectrum to be the pattern vector. The major difficulty arose in obtaining consistent and representative spectra of a large number of compounds. However, with the abundant number of library reference spectra now available, this problem is lessened.

A second type of input consists of a compound's molecular structure alone. The purpose of these studies, where the molecular structure serves as the input, is to search for correlations between structure and the physical, biological, or pharmacological properties of the compound. Examples of investigations of this type have been published including the generation of the low-resolution mass spectrum of a given molecule,[9] structure–activity studies of drugs,[10-14] and studies of cancer chemotherapy agents.[15,16] For studies of this type, the generation of descriptors from the molecule's structures to be used in the pattern vectors constitutes a major problem. Unfortunately, there is no single method available to express all of the chemical information embedded within a chemical structure. Therefore, the pattern vector must be constructed of bits and pieces of information extracted from the molecular structure. It is the intention of this paper to describe the molecular development routines employed in the ADAPT pattern recognition system detailed in the companion paper.

Table I. Connection Table for 2-Methyl-6-bromobenzothiazole

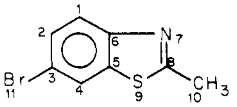
Atom no.	1	2	3	4	5	6	7	8	9	10	11
1	1	4				4					
2	4	1	4								
3		4	1	4							1
4			4	1	4						
5				4	1	4			1		
6	4				4	1	1				
7						1	3	2			
8							2	1	1	1	
9					1			1	4		
10								1		1	
11			1								7

Atom type	Numeric codes	Bond type	Numeric codes
C	1	Single	1
O	2	Double	2
N	3	Triple	3
S	4	Aromatic	4
F	5	Delocalized	5
Cl	6	Ionic	6
Br	7		
I	8		
P	9		

Table II. List of Atomic Descriptors

Atom descriptors	Value
Total number of atoms	11
Number of carbon atoms	8
Number of oxygen atoms	0
Number of nitrogen atoms	1
Number of sulfur atoms	1
Number of fluorine atoms	0
Number of chlorine atoms	0
Number of bromine atoms	1
Number of iodine atoms	0
Number of phosphorus atoms	0

Since all of the molecular descriptor routines to be discussed are computerized, it is necessary for the chemical structure to be in a generalized, computer-compatible format. In the ADAPT system, all molecules are stored in a standardized connection table. Other storage formats are possible,[17] but connection tables were chosen because of their simplicity, numeric nature, and their complete representation of the topology of molecules. The compound 2-methyl-6-bromobenzothiazole will serve as an example to aid in describing the different molecular descriptor routines. The connection table for this molecule is given in Table I along with the numeric codes used for the atoms and bonds. The sequence of atom numbering in the structure is arbitrary and indicates the atom's location in the connection table. Hydrogen atoms are not explicitly listed, but they are assumed to fill all otherwise vacant valencies. The main diagonal of the connection table matrix contains the numeric code for each atom. The off-diagonal elements of the connection table contain the bonding information of the structure. If the matrix element a_{ij} equals zero, a bond does not occur between atoms i and j. A value greater than zero indicates the presence of a bond between the two atoms, and the numerical value of the entry is the bond-type code. Since the matrix is symmetrical, only the diagonal and the upper triangle need to be stored for subsequent use. The connection tables used by the ADAPT system are generated during the input of a molecular structure at the CRT through the routine UDRAW.[18] The only required action taken by the user who is generating a structure file is to draw the structures: the connection table is automatically created by UDRAW. Once the molecule is in its connection table form, molecular descriptors can be derived.



FRAGMENT DESCRIPTORS

Any chemical structure can be broken down into its basic atom and bond components which contain information about the molecule's chemical nature. The total number of atoms and bonds in the structure are two examples of "fragment descriptors". However, further steps can be taken.

The total number of atoms in a chemical structure can be subdivided into smaller groups by first separating the atoms according to elemental type and then counting them. Since there is no advantage including atom types not found in the structures being studied, the number of new descriptors generated is data set dependent. A list of the atom descriptors generated by the DODABN routine of ADAPT is given in Table II along with the appropriate values for the example molecule. The values for these descriptors are easily obtained from the main diagonal of the connection table. Because of the numeric codes employed for the atom types, the sorting and counting of atomic descriptors is quickly accomplished by a digital computer.

Table III. List of Bond Descriptors

Bond descriptor	Value
Total number of bonds	12
Number of single bonds	5
Number of double bonds	1
Number of triple bonds	0
Number of aromatic bonds	6
Number of delocalized bonds	0
Number of ionic bonds	0

An analogous sorting and counting procedure can be carried out on a chemical structure for the various bond types. However, the total number of new descriptors possible is smaller owing to the limited number of different bond types. The six bond fragment descriptors listed in Table III can describe most molecules. Since the number of exceptions found in any given data set is small, there is no advantage to the inclusion of other bond fragments. As a matter of fact, data sets of organic molecules contain so few ionic and delocalized bonds that these descriptors may never be collated into the final pattern vector. Nevertheless, they are included for completeness. Moreover, in generating the bond fragments, a distinction is not made between identical bond types formed by different atom types, (i.e., a single bond is a single bond regardless of the atoms which it joins). These bond fragments are obtained from the off-diagonal elements of the connection table. The values calculated for the example molecule are included in Table III.

The manner in which these descriptors are implemented depends upon the specific data set under investigation. In some situations, the individual fragments can be used as single descriptors, but in other cases, mathematical operations (e.g., ratios, weighted linear summations, summation of squares) using these fragments may yield more meaningful descriptors. An example of a combination descriptor, arbitrarily called a "length" fragment, is given below.

$$\text{Length} = 4*\text{Single} + 3*\text{Aromatic} + 2*\text{Double} + \text{Triple}$$

As can be seen, this is an example of a weighted summation of the bond fragments where single and aromatic bonds are more important than the double and triple bonds. This demonstrates one of several different possible ways of combining fragment descriptors. However, the specific application determines the actual use of the fragment descriptors.

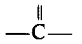
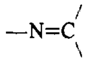
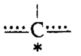
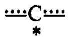
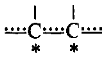
Additional molecular information is indirectly generated along with the general atom and bond information contained within these fragment descriptors. The molecular size and weight are directly related to the number of atoms and bonds in the molecule. Furthermore, if the total number of bonds is greater than or equal to the total number of atoms, a ring system is indicated. The number of hydrogens is implied by the number of unsaturated bonds and the atom types present. However, the information provided by a single descriptor depends upon the decision being sought from the classifier and the particular data set being used.

SUBSTRUCTURE DESCRIPTORS

Searching the molecule for the presence of larger fragments provides an alternative method for generating descriptors. If the "substructure" is found in the molecule, the descriptor can be given a value of one. Otherwise, it has a value of zero. Therefore, to generate substructure descriptors for a given molecular data set, two things are needed: a substructure searching algorithm and a library of appropriate substructures.

Algorithms for substructure searching fall into two general categories. The first, atom-by-atom searching, is the easiest to implement on a digital computer because it simply matches the structure and substructure atoms and associated bonds one

Table IV. Example Substructures and Substructure Search Results

Substructure[a]	Search general	Search specific
1. 	1	0
2. —S—	1	0
3. 	1	0
4. CH_3—	1	1
5. 	3	3
6. 	3	3
7. 	1	1

[a] (····) aromatic bond type; (*) designates a ring atom.

at a time using all possible combinations. However, for large structures and substructures, the time required for a single search becomes prohibitive because of the number of possible combinations increases factorially.

The second category utilizes set reduction techniques to accomplish the substructure search, and factorial calculations are not involved. Although they are more complex than atom-by-atom searching techniques, algorithms implementing set reduction are very attractive because of their searching speed. Several different algorithms have been described which use set reduction.[19-21] In the ADAPT system, a variation of the techniques described by Sussenguth[19] is used for generating substructure descriptors. The modifications allow for greater substructure specificity, a wider variety of substructure types, and numeric instead of binary searches. A discussion of the changes made in the Sussenguth's algorithm has been previously reported[9] and will not be detailed in this paper.

The problem of creating a substructure library is not as easy to solve as obtaining a good substructure searching algorithm. One approach to this problem involves the systematic combining of the basic atom and bond fragments into substructures. However, the final number of substructures generated in this manner would be totally unmanageable. The discrimination between usable and useless substructures would require some type of pattern recognition system, and this approach is not feasible. A more workable approach to the problem is to study the data set of molecules under investigation and allow the human brain, an excellent pattern recognition system, to decide on a collection of substructures to be applied to the data set. The ADAPT system utilizes this second method to generate a substructure library. A set of substructure descriptors can now be generated.

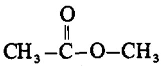
Seven substructures, which describe the example molecule, are given in Table IV along with the numeric results of the substructure searches using the example molecule. As indicated in the table, two types of searches are possible. For a general search, a match is made if the indicated substructure is located anywhere in the molecule; all ring information is ignored. However, during a specific search, ring information is taken into consideration. Therefore, if the substructure is not specified to be in a ring, it cannot possibly be matched to a molecular fragment that is contained in a ring system. Consequently, the first three substructures are not present in the example structure when a specific search is conducted.

The actual information contained in any one substructural descriptor depends highly upon the judgment of the person selecting the substructure library. In some applications, good descriptors can be obtained immediately because sufficient a priori knowledge exists. However, in other cases, a trial-and-error procedure may be warranted where a large number of possible substructures are generated and poor descriptors are eliminated by some prescreening criterion. In general, substructure descriptors serve a very important purpose in that they restore a portion of the structural information lost in the atom and bond fragmentation. Nevertheless, considerable structural information is still missing.

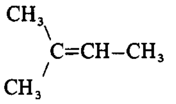
ENVIRONMENT DESCRIPTORS

The description of structures using fragment and substructure descriptors indicates the components of a molecule. However, the manner in which these individual parts are connected is not described. Environment descriptors take into account how different areas of a molecule fit together and provide a measure of the "environment" in which a single atom fragment finds itself.

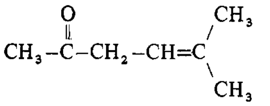
The environment descriptor describes the fragment's surroundings by including its first and second nearest neighbors and their bonds into a single parameter which reflects the atom and bond types connected to it. There may be more than one identical fragment in a molecule but they do not necessarily belong to the same functional group. For example, the fragment >C═ is found once in both structures A and B but twice in structure C. Obviously, the environment seen by this fragment would be different in each of the three cases. Of course, this difference is dependent upon the definition incorporated to calculate the environment descriptor. In the ADAPT system, the three forms most often used are: bond environment descriptors (BED), weighted environment descriptors (WED), and augmented environment descriptors (AED).



A



B



C

The procedure used to calculate these three parameters for a particular environment fragment is as follows:

1. Assign arbitrary values to each type of atom and bond. The values already employed in the connection table will suffice.
2. For "BED", sum the number of bonds connected to the fragment's first and second nearest neighbors.
3. For "WED", sum the values assigned to each bond type instead of merely counting the bonds.
4. For "AED", sum the product of the bond's assigned value and the assigned values for the two atoms which form the bond.

The BED, WED, and AED values for the fragment and structures given above are as follows: for structure A, BÉD = 5, WED = 6, AED = 11; for structure B, BED = 5, WED = 6, AED = 6; for structure C, BED = 12, WED = 15, AED = 17.

Since there may be more than one fragment present, the environment descriptor indicates the sum of all the environments for a given fragment. This feature makes them useful when used in conjunction with substructure descriptors. The substructure descriptors indicate the number of times a particular fragment is found in the molecule and the environment descriptors indicate the context in which the fragment is found.

The routine that generates the environment descriptors must have access to the file of molecular structures and to the atom-centered fragment library which is constructed by the user. The actual calculation of the environment descriptors proceeds extremely rapidly since both the fragment location and necessary calculations are easily done by a computer.

The concept of the environment is not limited to connectivities, but could take into account electron densities, bond distances, electronegativities, or other physical parameters. This can be done by replacing the values assigned in step 1 by the desired parameters. In this manner, more informative descriptors may be obtained.

Use of the environment descriptors may reveal relations which are not particularly obvious. Note that both structures A and B have the same BED and WED values. These structures, which at first glance appear quite different, do indeed have these parameters in common. However, when one takes into account the type of atoms connected to these bonds the difference becomes apparent. Such relationships may or may not prove significant. Their ultimate utility depends on the type of environment measure, the molecule being coded, and the problem being attacked.

GEOMETRIC DESCRIPTORS

The descriptors discussed so far have all been generated from the molecular connection tables which represent the two-dimensional structural diagram of the molecules. However, since molecules are actually three-dimensional, it would seem only logical to generate descriptors which incorporate the three-dimensionality of the structure.

Unfortunately, obtaining the actual structural shape of any given molecule is not a simple task. The use of x-ray data is not reasonable since the probability is very small that all of the structures in a given data set have been studied. Other empirical methods could be used to obtain structural information, but this would be extremely costly and time consuming for a data set usable in pattern recognition studies. Three-dimensional space models could be constructed by hand, but this would be extremely tedious and the extraction of geometric parameters from the models would be difficult at best.

The research area of molecular mechanics deals entirely with the calculation of molecular geometries and energies using classical mechanical principles. Ideally, these calculations should be done using the appropriate Schrodinger equation for the molecular system under investigation. In practice, this is not done because of the complexities and computational difficulties encountered. Even though simplified quantum mechanical treatments have been reported,[22,23] the desired degree of accuracy for large organic molecules is still to be gained. Nevertheless, this method is clearly the preferred approach, and accurate results for large molecules may be attainable in the future. In the meantime, another technique is being applied to calculate geometric models.

A molecule can be viewed as a collection of particles held together by simple harmonic or elastic forces. These forces can be defined by potential energy functions whose terms are the atom coordinates of the molecule. This function can then be minimized to obtain a strain-free three-dimensional model of the molecule. Geometric parameters can then be extracted. A wealth of information already exists describing the procedures and results of several different molecular mechanics algorithms.[24,25] Therefore, finding and implementing an algorithm to model sets of molecules is a relatively straightforward procedure. A modified version of the molecular mechanics routine described by Wipke et al.[26-28] has been developed and interfaced to the ADAPT system so that geometric descriptors can be derived from the resulting molecular structure.

The molecular mechanics routine, MOLMEC, used in conjunction with the ADAPT system is highly interactive and relies on graphical input and output. The program is executed using a MODCOMP II/25 computer containing 128 K bytes of memory. A Tektronix 4010 storage CRT graphics unit is also supported and is utilized by MOLMEC for the displaying of the molecule being modeled.

Table V. Bond Length Strain Function[a]

$$E_{\text{bond}} = (K/2)(L - L_0)^2$$

Bond type	L_0, Å	K_b
C–C	1.54	312
C=C	1.34	500
C≡C	1.20	500
C⋯C	1.39	312
C–O	1.43	312
C=O	1.22	500
C–N	1.47	312
C=N	1.29	500
C≡N	1.16	500
C⋯N	1.34	312

[a] K_b = bond stretching constant, L_0 = expected bond length, and L = observed bond length.

Table VI. Bond Angle Strain Function[a]

$$E_{\text{angle}} = (K_a/2)(\Theta - \Theta_0)^2$$

Angle type	Θ_0 (57.3)	K_a
sp^3	109.5	80.1
sp^2	120.0	100.0
sp	180.0	150.0
sp^3 [b]	109.5	20.0

[a] K_a = angle strain constant, Θ_0 = expected bond angle, and Θ = observed bond angle. [b] Noncarbon atom centers with tetrahedron shape orbitals.

Table VII. Torsional Angle Strain Function

$$E_{\text{torsion}} = K_t F(\Phi')^2$$

Torsional bond type	K_t	F	Φ' (57.3)[a]
–A–B–	1.0	1.0	60 – Φ, Φ < 60
–A=B–	15.0	1.628	Φ (cis)
			180 – Φ (trans)
=A–B=	0.003	1.628	Φ (cis)
			180 – Φ (trans)
A⋯B	15.0	1.628	Φ
A–B=C=D–E	15.0	1.6	90 – Φ (for angle ABDE)

[a] Φ is the measured dihedral angle in degrees.

Table VIII. Nonbonded Strain Function[a]

$$E_{\text{nonbonded}} = (K_{nb}/2)(D - D_0)^M$$

Interaction type	D_0, Å
C–C–C (1,3 nonbonded)	2.52
C–C–heteroatom (1,3 nonbonded)	2.25
All other nonbonded interactions	3.50 (poor models)
	3.0 (good models)

[a] K_{nb} = 28.76, D_0 = expected interatomic distance, D = observed interatomic distance, M = 2 for poor models, and M = 6 for good models.

The structure input section of MOLMEC has been designed to allow the user to either read the molecule's connection table from ADAPT's disc files or else accept the structure from the CRT via UDRAW. Thus, MOLMEC can be used independently of the ADAPT system. Once the molecule has been entered, control branches to the interactive section where the user can direct the different phases of modeling as well as monitor the results.

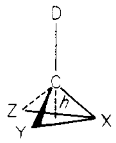
In the strain minimization section, the atom coordinates are systematically altered until a minimum is found in the strain or potential energy function. The actual strain function used in MOLMEC is:

$$E_{\text{strain}} = E_{\text{bond}} + E_{\text{angle}} + E_{\text{torsion}} + E_{\text{nonbonded}} + E_{\text{stereo}}$$

The first four terms of the function are commonly found in all molecular mechanics strain functions and are defined more explicitly in Tables V through VIII. The last term of the

Table IX. Stereochemistry Strain Function

$$E_{\text{stereo}} = 0.5 + 100(h - 0.2)^2$$



function has been added to assure the proper stereochemistry about an asymmetric atom. Its form is given in Table IX.

The actual minimization of the function is best accomplished by some type of nonlinear programming method (e.g., steepest descent). In MOLMEC, an adaptive pattern search routine[29] is used because it does not require analytical derivatives. The amount of time necessary to obtain good molecular models depends upon the number of atoms in the molecule, the initial strain of the molecule, and the degrees of freedom in the structure. If a small molecule is being modeled, only one pass through the minimization section may be sufficient to obtain a good structure. However, this is seldom the case. Usually, the molecules are rather large and require several passes. The actual amount of time per pass is limited by a cutoff parameter so that the user may analyze the progress of the modeling at different intervals.

The graphics interaction section of MOLMEC contains routines capable of rotating and aligning the molecule into any desired position. Since the graphics unit is only a two-dimensional screen, rotation is essential to obtain a good view of the structure. Furthermore, these routines are useful in locating atoms trapped in local minima. If such an atom is found, the user can move the trapped atom to a new position by a MOVE routine found in the graphics section. Naturally, if the structure is altered, the molecule should be passed through the minimization routine at least once more.

When the molecule is finally in a low strain energy conformation, the molecular parameters can be either listed on an output device, or else the structure's coordinate matrix can be stored on a disc file for further processing.

An automatic version of MOLMEC has also been developed so that large molecular data sets can be modeled without continuous supervision. The program consists of an input section, which reads the molecule's connection table and present coordinate matrix from the ADAPT files, a minimization section with all output suppressed, and a section which stores the final coordinate matrix. Good models can easily be obtained in this manner. However, before the coordinate matrices can be used for calculating descriptors, the structures are reviewed to make sure that the molecules are in acceptable conformations. Once modeling is complete, geometric descriptors can be derived.

Presently, two basic types of geometric descriptors are calculated from the molecular structures. The three principal axes of the molecule form the basis for the first type of geometric descriptor. These axes are calculated by a radius of gyration program the steps of which are:

Step 1. Calculate the center of mass, $\bar{u}$, for the molecule where the element u_j is:

$$u_j = \frac{1}{M_T} \sum_{i=1}^{N} m_i x_{ij} \text{ for } j = 1, 2, 3$$

where M_T is the total mass of the molecule, m_i is the atomic mass of atom i, N is the total of atoms in the structure, and x_{ij} is the jth coordinate for the ith atom.

Step 2. Calculate the tensor of gyration matrix **R** where the element r_{jk} is:

$$r_{jk} = \frac{1}{M_T} \sum_{i=1}^{N} m_i(x_{ij} - u_j)(x_{ik} - u_k)$$

for $j = 1, 2, 3$; $k = 1, 2, 3$.

Step 3. Diagonalize the tensor of gyration matrix to obtain the eigenvalues.

The actual diagonalization of the tensor of gyration matrix is done by the Jacobi method since the matrix is symmetrical. The eigenvalues obtained correspond to the three principal radii of the molecule. Since the orientation of the original molecule in space is essentially random, the radii must be sorted in some manner. This is done by arbitrarily assigning X to the longest radius, Y to the second longest radius, and Z to the shortest radius. Once sorted, the three ratios, X/Y, X/Z, and Y/Z, are also calculated. Because of their small values, all of the radii are multiplied by some constant scaling factor to prevent loss of information during truncation. These six geometric parameters are then used as new descriptors and constitute the first set of geometric descriptors.

The van der Waals volume of a molecule is the other type of geometric descriptor generated in the ADAPT system. Before this calculation can be done, the bond distances and the van der Waals radii of the atoms must be known. The bond distances are easily obtained from the molecular modeling results. For the van der Waals radii, an article published by Bondi[30] was consulted. The volume occupied by an atom is taken as that of a sphere with radius equal to the van der Waals radius of the atom minus the volume of overlap with adjacent atoms. The overlap volumes are calculated from standard spherical geometry formulas. The actual volume is not found for two reasons: the assumption of sphere and spherical segments is not totally correct, and the radii used were selected as being the "best" values from a large collection of data using an empirical selection method. Table X contains the van der Waals radii actually used in the calculation. The total molecular volume for the molecule is taken as the sum of the contributions for each atom found as described above. The volume contributions of attached hydrogens are also included in the calculation of the total volume.

Table X. van der Waals Radii Used in the Molecular Volume Calculation

Atom type	Radius, Å	X-H, Å^3/H atom
C—	1.70	1.83
C=	1.70	1.83
C≡	1.78	1.36
C⋯	1.77	0.50
O—	1.52	2.29
O=	1.50	
N—	1.55	2.38
N=	1.55	2.38
N≡	1.60	2.23
N⋯	1.60	2.23
S—	1.80	5.55
S=	1.75	
F—	1.50	
Cl—	1.75	
Br—	1.85	
I—	1.97	
P—	1.80	2.86
H—	1.20	
H⋯	1.00	

In order to make the routine more versatile, the option of either using standard bond distances or modeled bond distances is included. Since MOLMEC uses the standard bond distances to determine a low strain geometry, it is not surprising that for a well-modeled data set the molecular volumes calculated using the two different bond distances are very similar. However, discrepancies can arise when the molecule contains rings of five or fewer atoms which cause a large amount of bond strain. The volumes are initially calculated in units of cubic angstroms per atom but are then converted to units of cubic centimeters per mole. The molecular volume can then be used as another geometric descriptor. In Table XI, the

Table XI. Geometric Descriptors

Descriptor	Value[a]	Descriptor	Value[a]
X	719	*X/Y*	9
Y	82	*X/Z*	360
Z	2	*Y/Z*	41
		Volume	94

[a] All values of the radii were scaled by a factor of 100 and then rounded off.

values for the seven geometric descriptors derived from the example molecule are given.

Each geometric descriptor contains some information about the molecule. The radii and ratios describe the general shape of the molecule which may be very important in systems where receptor sites are involved. However, this is only a relative shape since the model obtained is for the molecule in a vacuum: in some environments, the molecule's shape will change, especially if long chains are present. On the other hand, the molecular volume is essentially constant regardless of how the molecule is bent. However, like any other descriptor, the actual value of any geometric descriptor depends upon the specific application in which it is used.

DISCUSSION

The descriptor routines discussed in this paper represent four different approaches which are being taken to extract information from a molecular structure. Each routine requires a different level of computational effort and yields different information about the structure being encoded.

Fragment descriptors are easy to compute from the molecular connection table and reduce molecules to their simplest units. Although structural information is lost, information about the chemical nature of the entire molecule is retained and may be useful in obtaining the desired results.

The environmental and substructural descriptors are similar in that they both contain information about the molecule's structural make-up which was lost in the fragmentation process. However, they differ considerably in the amount of computational effort necessary to generate a descriptor. For the environmental descriptors, an atom-by-atom search for a single atom fragment followed by a straightforward calculation is all that is necessary to calculate the descriptor. To generate substructure descriptors, a more complex searching algorithm, which can handle multiple atom fragments, is required. Although generated in different manners, both of these descriptors carry information about the chemical and structural nature of the molecule.

Finally, there are the geometric descriptors which mainly contain information about the overall molecular shape and very little about the chemical nature of the molecule. Unfortunately, in order to obtain geometric descriptors, a molecular mechanics program must be implemented and the data set must be modeled. Both steps require a considerable amount of time. However, once the molecules are modeled and the coordinate matrix calculated, the generation of the descriptors is quickly done.

Although studies done entirely with the ADAPT system have not been published, the descriptors discussed in this paper have been used in two published reports. For the generation of the low-resolution mass spectrum of a given molecule,[9] fragment, substructure, environment, and geometric descriptors were employed to obtain the desired results. In the classification of psychotropic drugs as sedatives or tranquilizers,[14] fragment, environment, and substructure descriptors were used. These studied demonstrate how the descriptors are implemented in pattern recognition studies, and they illustrate the versatility of substructure and environment descriptors for describing different data sets.

As can be seen, the generation of representative descriptors from molecular structures is no easy task. However, the ability to predict different chemical or biological properties of a molecule from its chemical structure is a worthwhile goal. Presently, barbiturate, triazine, and olfaction molecular data sets are being studied using the ADAPT system; results will be published in the future. Along with these studies, ideas are being formulated for generating more informative geometric descriptors: algorithms for calculating quantum mechanical parameters and molecular lipophilicities are also being planned. Thus, this paper represents merely the beginning of the area of descriptor development with the future holding the key for new and improved methods.

ACKNOWLEDGMENT

The authors acknowledge the partial financial support by the National Science Foundation of the MODCOMP computer used in this project.

REFERENCES AND NOTES

(1) P. C. Jurs, B. R. Kowalski, and T. L. Isenhour, *Anal. Chem.*, **41**, 21 (1969).
(2) P. C. Jurs, B. R. Kowalski, T. L. Isenhour, and C. N. Reilley, *Anal. Chem.*, **42**, 1387 (1970).
(3) P. C. Jurs, *Anal. Chem.*, **43**, 1812 (1971).
(4) B. R. Kowalski, P. C. Jurs, T. L. Isenhour, and C. N. Reilley, *Anal. Chem.*, **41**, 1945 (1969).
(5) R. W. Liddell and P. C. Jurs, *Appl. Spectrosc.*, **27**, 371 (1973).
(6) B. R. Kowalski and C. A. Reilly, *J. Phys. Chem.*, **75**, 1402 (1972).
(7) B. R. Kowalski and C. F. Bender, *Anal. Chem.*, **44**, 1405 (1972).
(8) L. B. Sybrandt and S. P. Perone, *Anal. Chem.*, **43**, 382 (1971).
(9) G. S. Zander and P. C. Jurs, *Anal. Chem.*, **47**, 1562 (1975).
(10) C. Hansch, S. H. Unger, and A. B. Forsythe, *J. Med. Chem.*, **16**, 1217 (1973).
(11) S. A. Hiller et al., *Comput. Biomed. Res.*, **6**, 411 (1973).
(12) G. W. Adamson and J. A. Bush, *Nature (London)*, **248**, 406 (1974).
(13) K. C. Chu, *Anal. Chem.*, **46**, 1181 (1974).
(14) A. J. Stuper and P. C. Jurs, *J. Am. Chem. Soc.*, **97**, 182 (1975).
(15) B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.*, **96**, 916 (1974).
(16) K. C. Chu et al., *J. Med. Chem.*, **18**, 539 (1975).
(17) M. F. Lynch, J. M. Harrison, W. G. Town, and J. E. Ash, "Computer Handling of Chemical Structure Information", Macdonald, London, 1971.
(18) W. E. Brugger and P. C. Jurs, *Anal. Chem.*, **47**, 781 (1975).
(19) E. H. Sussenguth, Jr., *J. Chem. Doc.*, **5**, 36 (1965).
(20) T.-K. Ming and S. J. Tauber, *J. Chem. Doc.*, **11**, 47 (1971).
(21) J. Figeras, *J. Chem. Doc.*, **12**, 237 (1972).
(22) R. Hofffmann, *J. Chem. Phys.*, **39**, 1397 (1963).
(23) R. Hoffmann, *Tetrahedron*, **22**, 521 (1966).
(24) E. M. Engler, J. D. Andose, and P. v. R. Schleyer, *J. Am. Chem. Soc.*, **95**, 8005 (1973).
(25) J. E. Williams, P. J. Strang, and P. v. R. Schleyer, *Annu. Rev. Phys. Chem.*, **19**, 531 (1968).
(26) W. T. Wipke, T. M. Dyott, and J. G. Werbalis, Abstracts, 161st National Meeting of the American Chemical Society, Los Angeles, Calif., March 1971.
(27) W. T. Wipke, P. Gund, J. G. Verbalis, and T. M. Dyott, Abstracts, 162nd National Meeting of the American Chemical Society, Washington, D.C., Sept 1971.
(28) W. T. Wipke, P. Gund, T. M. Dyott, and J. G. Verbalis, unpublished manuscript.
(29) E. S. Buffa and W. H. Taubert, "Production-Inventory Systems, Planning and Control", Rev. Ed., R. D. Irwin, Inc., Homewood, Ill., 1972.
(30) A. Bondi, *J. Phys. Chem.*, **68**, 441 (1964).