The Chemical Abstracts Service Chemical Registry System. II. Augmented Connectivity Molecular Formula†

R. G. FREELAND, S. A. FUNK, L. J. O'KORN,* and G. A. WILSON

Chemical Abstracts Service, Box 3012, Columbus, Ohio 43210

Received November 15, 1978

The Chemical Abstracts Service (CAS) Chemical Registry System grew out of research in the early 1960s. Building upon work supplied by du Pont, CAS perfected an algorithm for generating a unique and unambiguous computer-language representation of a chemical structure. This algorithm became the foundation of the CAS Chemical Registry System which provides unique identification of chemical substances on the basis of their molecular structure. Because of the importance of this algorithm to the CAS Chemical Registry System, techniques have been developed to improve validation of the algorithm and to ensure integrity of the Registry database. These improved validation procedures are based on a simple, highly efficient computer calculation of a compact representation for a structure; this representation is called the Augmented Connectivity Molecular Formula (ACMF). The derivation of the ACMF, its use as a validation tool, and results from its use are described.

INTRODUCTION

Chemists often represent chemical substances by graphs whose vertices represent the atoms making up a molecule and whose edges represent the chemical bonds between the atoms. The vertices are labeled with element symbols, e.g., C = carbon, N = nitrogen, O = oxygen, etc.; and the edges are weighted to indicate the various types of chemical bonds, e.g., single, double, etc. Structural details, e.g., charge, abnormal valence, isotopic mass, etc., can be associated with each vertex. These graphs may in turn be represented by a connection table well suited to computer processing. The connection table is a detailed inventory derived by computer program of the atoms and bonds comprising the basic structure of a substance. O'Korn described various schemes for chemical substance representation and manipulation.[1] The Chemical Abstracts Service (CAS) Chemical Registry System is a computer-based system which uniquely identifies chemical substances on the basis of their molecular structure, and this system utilizes connection tables to provide a unique and unambiguous representation of the chemical substance. Dittmar, Stobaugh, and Watson[2] described features and capabilities of the CAS Chemical Registry System and the details of the connection tables used to represent the chemical substances in the CAS Chemical Registry System. Terminology and concepts presented in that paper are useful background for this paper.

The CAS Chemical Registry System, which is based on a concept originally suggested by G. Malcolm Dyson, grew out of research in the early 1960s. Harry Morgan of the CAS staff, building upon work suppled by du Pont, perfected an algorithm for generating a unique and unambiguous computer representation of a chemical substance, i.e., a unique connection table.[3] This algorithm is the foundation of the CAS Chemical Registry System. Since the CAS Chemical Registry System is based on the unique and unambiguous representation of chemical substances, determination of whether a potentially new substance is already on file is reduced to comparing the unique, unambiguous representation of the candidate substance with those of the substances previously on file. As part of the CAS Chemical Registry System, CAS has developed an additional compact, simply calculated representation for a chemical structure called the Augmented Connectivity Molecular Formula (ACMF). The ACMF was developed for several purposes:

File Validation—Because of the importance of the unique connection table generation algorithm to the CAS Chemical Registry System and the size of the file, a machine-supported validation technique independent of the unique table generation algorithm is desired to ensure that the programmed implementation of the unique table generation algorithm is correct.

Symmetrical Structures—The computer time requirements for the unique table generation algorithm to uniquely label a small number of highly symmetrical structures are prohibitively large; therefore, alternate representations and techniques are needed to support registration of these symmetrical structures.

Direct Access File Addresses—Since the current CAS Chemical Registry System has a direct access database, an easily computed representation is required to provide a precise file address based on structural characteristics of the chemical substances. This is the same type of requirement typically satisfied by some form of "hash code" addressing for any direct access file.

In the following sections, the algorithm for calculation of the ACMF is outlined, an illustrative example is provided, and the results from application of the algorithm to the CAS Registry database are provided.

THE ALGORITHM

To illustrate the calculations of the ACMF, consider the chemical substance shown in Figure 1. (Note that we adopt the convention of omitting hydrogen atoms from the structures in cases where the number of hydrogen atoms attached to an atom would be obvious to a chemist.) One simple way to describe this substance at a less specific level is to compute its molecular formula, in this case $C_3H_6N_2$. For some purposes (e.g., printed indexes sequenced by molecular formula) the molecular formula is a useful description of a chemical substance, but in many situations it is too ambiguous. For example, structures shown in Figure 2 illustrate some of the substances which share the formula $C_3H_6N_2$ but which are, nevertheless, clearly distinct substances.

Calculation of the molecular formula may be viewed as partitioning the vertices of the graph based on element symbols and counting the number of vertices in each component of the partitioning. Thus, the structure in Figure 1 contains three carbon atoms and two nitrogen atoms. Examination indicates that a finer partitioning is possible, for it is evident that the

† Presented before the Division of Chemical Information, 169th National Meeting of the American Chemical Society, Philadelphia, Pa., April 9, 1975.

0095-2338/79/1619-0094$01.00/0

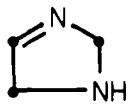


Figure 1. Substance with molecular formula $C_3H_6N_2$.

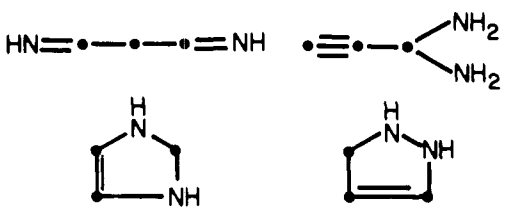


Figure 2. Substances which share molecular formula $C_3H_6N_2$.

three carbon atoms may be differentiated from one another in terms of the kinds of atoms to which they are adjacent and the types of bonds connecting them to the adjacent atoms. Similarly, the nitrogen atoms are in some ways different. To quantify these differences, i.e., to obtain a finer partitioning among the atoms, we calculate the augmented connectivity values[4] for each atom and use these values for calculation of the ACMF.

The algorithm for ACMF computation is as follows:

Step 1. Assign to each vertex of the graph, a "level 1" value based on the element symbol as indicated in Table I. (The values associated with the element symbols are chosen to permit representation of each element in one eight-bit byte, while maintaining the alphabetic sequence and providing for the insertion of one additional element symbol.)

Step 2. Assign to each vertex of the graph, a "level 2" value computed by multiplying the level 1 value for each adjacent vertex times the appropriate bond-type number from Table II and summing these products over all vertices adjacent to the given vertex. (The values associated with the bond types are chosen as primes to cause the calculated products to scatter, thus minimizing the opportunity for overlap among the products.)

Step 3. Given the "level n" values for the vertices of the graph, for each vertex assign the "level $n + 1$" value by summing the "level n" values for all vertices adjacent to the given vertex.

Step 4. If $n < 4$, increment n by 1 and go to step 3.

Step 5. Count the number of distinct "level n" values and the number of distinct "level $n + 1$" values. If there are more "level $n + 1$" values than "level n" values, increment n by 1 and go to step 3. (Note that the number of iterations to terminate the ACMF computation is limited by the number of nonhydrogen atoms, since the maximum number of distinct values that can be obtained is equal to the number of nonhydrogen atoms.)

Step 6. Assign as the augmented connectivity value for each vertex the "level n" value for that vertex.

Step 7. Construct a string consisting of the element value associated with the vertex, the augmented connectivity value of the vertex, and the number of vertices with an equal element value and augmented connectivity value.

Step 8. In ascending sequence, concatenate the strings generated in step 7.

Step 9. Incorporate special structural characteristics as follows:

(a) For each vertex with an abnormal valence,[5] construct a string consisting of the product of the augmented connectivity value for the vertex and its abnormal valence. If there is more than one vertex with an abnormal valence, concatenate these strings in ascending sequence.

(b) For each vertex with an abnormal mass,[5] construct a string consisting of the product of the augmented connectivity value for the vertex and its abnormal mass. If there is more than one vertex with an abnormal mass, concatenate these strings in ascending sequence.

(c) For atoms with isotopes at unknown locations,[5] construct a string consisting of the element value and its mass value. If there is more than one atom with isotopes at unknown locations, concatenate these strings in ascending sequence.

(d) For delocalized charges,[5] multiply the charge times the augmented connectivity value for each of the vertices and concatenate those strings in ascending sequence. If there is more than one delocalized charge, concatenate the constructed strings in ascending sequence.

(e) For alloy composition,[5] construct a string consisting of the percentages of each component listed in ascending sequence.

(f) For each tautomer mobile group,[5] construct a string consisting of the types of charges and the sum of the augmented connectivity values for each vertex within the group. If there is more than one tautomer mobile group, they are listed in ascending sequence.

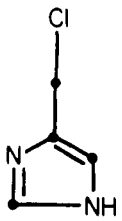
Table I. Element Values for ACMF Calculations

Ac	32	Dy	84	Mg	136	Rn	188
Ag	34	Er	86	Mn	138	Ru	190
Al	36	Es	88	Mo	140	S	192
Am	38	Eu	90	N	142	Sb	194
Ar	40	F	92	Na	144	Sc	196
As	42	Fe	94	Nb	146	Se	198
At	44	Fm	96	Nd	148	Si	200
Au	46	Fr	98	Ne	150	Sm	202
B	48	Ga	100	Ni	152	Sn	204
Ba	50	Gd	102	No	154	Sr	206
Be	52	Ge	104	Np	156	T	208
Bi	54	H	106	O	158	Ta	210
Bk	56	He	108	Os	160	Tb	212
Br	58	Hf	110	P	162	Tc	214
C	60	Hg	112	Pa	164	Te	216
Ca	62	Ho	114	Pb	166	Th	218
Cd	64	I	116	Pd	168	Ti	220
Ce	66	In	118	Pm	170	Tl	222
Cf	68	Ir	120	Po	172	Tm	224
Cl	70	K	122	Pr	174	U	226
Cm	72	Kr	124	Pt	176	V	228
Co	74	La	126	Pu	178	W	230
Cr	76	Li	128	Ra	180	Xe	232
Cs	78	Lu	130	Rb	182	Y	234
Cu	80	Lr	132*	Re	184	Yb	236
D	82	Md	134	Rh	186	Zn	238
						Zr	240

*The element symbol was changed from Lw to Lr to conform to international usage; the element value of 132 remained unchanged.

Table II. Bond Values for ACMF Calculations

Cyclic Single Bond	3	Acyclic Single Bond	19
Cyclic Double Bond	5	Acyclic Double Bond	23
Cyclic Tautomer Bond	7	Acyclic Tautomer Bond	29
Cyclic Delocalized Bond	11	Acyclic Delocalized Bond	31
Cyclic Alternating Bond	13	Acyclic Triple Bond	37
Cyclic Triple Bond	17		



Registry Number: 23785-22-0
CA Index Name: 1*H*-Imidazole, 4-(chloromethyl)-

Figure 3. Illustrative example of augmented connectivity molecular formula calculation.

(g) For each vertex with an abnormal charge,[5] construct a string consisting of the product of the augmented connectivity value for the vertex and its abnormal charge. If there is more than one vertex with an abnormal charge, concatenate these strings in ascending sequence.

(h) For hydrogen isotopes, construct a string consisting of the product of the connectivity value of the vertex, the mass of the hydrogen, and the number of hydrogens of that mass at that vertex. If there is more than one hydrogen isotope, concatenate these strings in ascending sequence.

Step 10. Concatenate the string constructed in step 8, the strings constructed in steps 9(a–h), and the stereochemical descriptor.

Step 11. If the string constructed in step 10 is not a multiple of eight bytes, pad the string constructed in step 10 with binary zeroes to make it a multiple of eight bytes.

Step 12. Treat each eight-byte segment of the string constructed in step 11 as a binary number, add the first eight bytes to the second eight bytes, and rotate the sum left one bit. (Rotate left means shift left one bit and place the original high order bit in the low order position of the new string.)

Step 13. Add the third eight bytes to the total and rotate the sum left one bit.

Step 14. Continue until all eight-byte segments have been reflected in the sum. The result, an eight-byte binary string, is the ACMF.

ILLUSTRATIVE EXAMPLE

Calculation of the ACMF is illustrated using the structure provided in Figure 3.

As shown in Figure 4A, the level 1 value associated with each nonhydrogen atom of the given structure is the value associated with each element symbol listed in Table I. For this example, C = 60, Cl = 70, and N = 142.

For each atom, the level 2 augmented connectivity values are calculated by summing the products of the attached atom's element value times the bond value for each of the attached atoms. The bond values from Table II for this example are: acyclic single bond = 19, cyclic single bond = 3, and cyclic double bond = 5.

As shown in Figure 4B, the level 2 value for the Cl atom is the product of 60 (the value for carbon) times 19 (the value for the attaching acyclic single bond). The level 2 augmented connectivity value for the C joining the ring and the acyclic string is the sum of the product 60 (the value for the attached carbon) times 19 (the value for the acyclic single bond) plus 60 (the value for the attached carbon) times 5 (the value for a double cyclic bond) plus 142 (the value for the attached nitrogen) times 3 (the value for a single cyclic bond), i.e., (19 × 60) + (5 × 60) + (3 × 142) = 1866. These calculations are made for every atom. Through these calculations we have reflected the element symbols and the bonds into our augmented connectivity value calculations.

For level 3 calculations and beyond, the augmented connectivity value for each vertex is calculated by summing the

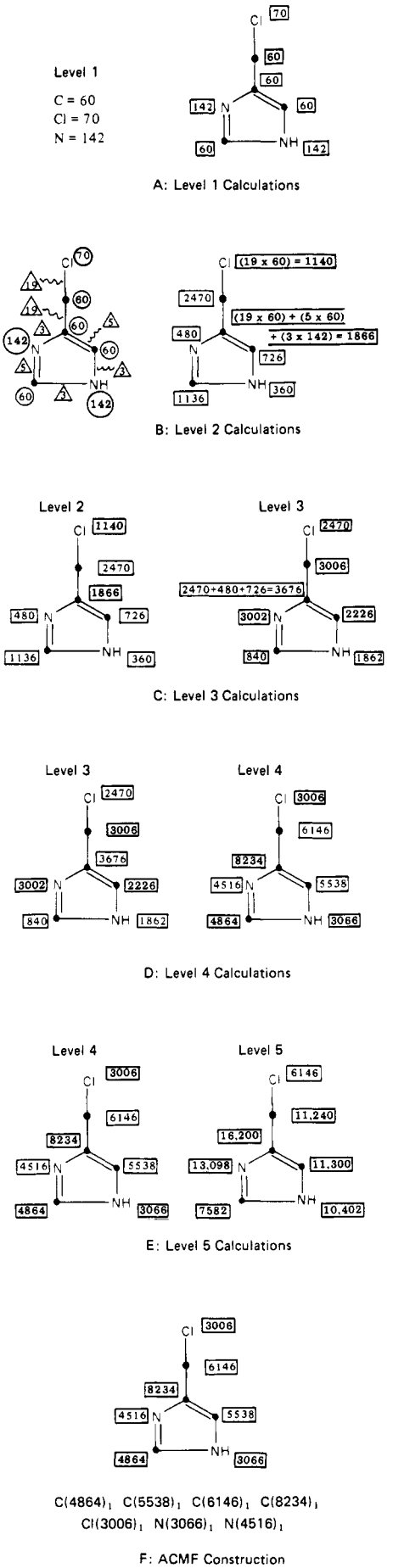
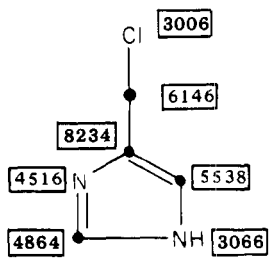
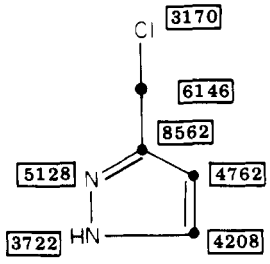


Figure 4. Illustrative example of the augmented connectivity molecular formula calculations.



Registry Number: 23785-22-0
CA Index Name: 1*H*-Imidazole, 4-(chloromethyl)-
Molecular Formula: $C_4H_5ClN_2$
ACMF: $C(4864)_1$ $C(5538)_1$ $C(6146)_1$ $C(8234)_1$ $Cl(3006)_1$ $N(3066)_1$ $N(4516)_1$



Registry Number: 23784-89-6
CA Index Name: 1*H*-Pyrazole, 3-(chloromethyl)-
Molecular Formula: $C_4H_5ClN_2$
ACMF: $C(4208)_1$ $C(4762)_1$ $C(6146)_1$ $C(8562)_1$ $Cl(3170)_1$ $N(3722)_1$ $N(5128)_1$

Figure 5. ACMF distinguishing between positional isomers.

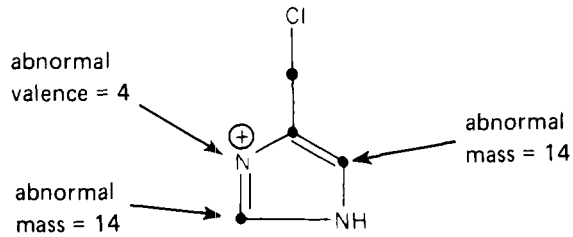


Figure 6. ACMF calculations with special structural characteristics.

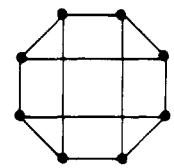
augmented connectivity values of the previous level for each of the attached atoms. As shown in Figure 4C, the level 3 augmented connectivity value for the C joining the ring and the acyclic string is the sum of 2470 + 480 + 726 = 3676, i.e., the sum of level 2 augmented connectivity values for attached atoms. This calculation is made for every atom.

As shown in Figure 4D and 4E we continue this process until at least the fifth level and stop when the number of distinct values at a given stage is less than or equal to the number of distinct values at the previous stage. At this point, we use the values of the previous stage. For this example, the number of distinct values at level 5 is equal to the number of distinct values at level 4, and the values from level 4 are used in constructing the ACMF.

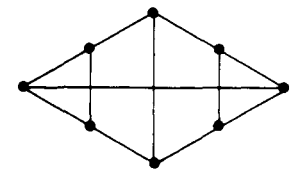
As shown in Figure 4F, the ACMF is constructed by concatenating the element symbol, the augmented connectivity value, and the number of vertices with the same element and augmented connectivity value. The substance used as an example does not contain a stereochemical descriptor or any special structural characteristics. We reduce the string shown in Figure 4F to a fixed length field of eight bytes by hash coding the string as described in algorithm steps 11 through 14. The resulting eight-byte field is the ACMF. The actual eight-byte ACMF can vary based on choices made during implementation, e.g., storage mode, field length, field formats, etc.

By incorporation of the element symbols and bond values in the ACMF calculations, the ACMF is able to achieve a finer partitioning of the atoms of a substance. This finer partitioning is illustrated in Figure 5 where the ACMF is able to distinguish between the two positional isomers.

The example illustrated in Figure 4 does not include any special structural characteristics. Let us consider these characteristics and the way in which they are incorporated into the ACMF as illustrated in the example shown in Figure 6. Figure 6 contains the same substance illustrated in Figure 4 but the Figure 6 substance has the following special characteristics: an abnormal valence of 4 on the nitrogen atom, a positive charge on the nitrogen atom, and an abnormal mass of 14 on two of the carbon atoms in the ring.



Registry Number: 277-10-1
CA Index Name: Pentacyclo[4.2.0.$0^{2,5}$.$0^{3,8}$.$0^{4,7}$]octane
Molecular Formula: C_8H_8



Registry Number: 20656-23-9
CA Index Name: Pentacyclo[3.3.0.$0^{2,4}$.$0^{3,7}$.$0^{6,8}$]octane
Molecular Formula: C_8H_8

Figure 7. Distinct substances with the same ACMF.

In deriving the ACMF, steps 1 through 8 are calculated in the same manner and are not affected by the special structural characteristics. Step 8 yields the following results:

$$C(4864)_1\ C(5538)_1\ C(6146)_1\ C(8234)_1\ Cl(3006)_1\ N(3066)_1\ N(4516)_1$$

To reflect the abnormal valence on the nitrogen atom in the ACMF, we build the following string containing the product of the augmented connectivity value for the nitrogen atom and its abnormal valence:

$$4516 \times 4 = 18064$$

To reflect the carbon atoms with abnormal mass in the ACMF, we build the string containing the product of the augmented connectivity value and the abnormal mass for the atoms with an abnormal mass as follows:

$$5538 \times 14 = 77532 \text{ and } 4864 \times 14 = 68096$$

These fields are sorted in ascending sequence, i.e.,

$$68096 \quad 77532$$

To reflect the charge on the nitrogen atom in the ACMF, we build the string containing the augmented connectivity value for the nitrogen atom and its charge as follows:

$$4516 \times 1 = 4516$$

The string from step 8 is concatenated with the three strings corresponding to abnormal valence (step 9a), abnormal mass (step 9b), and charge (step 9g), i.e.,

$$C(4864)_1\ C(5538)_1\ C(6146)_1\ C(8234)_1\ Cl(3006)_1\ N(3066)_1\ N(4516)_1 \parallel 18{,}064 \parallel 68{,}096\ 77{,}532 \parallel 4516$$

If the substance had other special structural characteristics or stereochemistry, the representation for these properties would be incorporated as described in step 9 of the algorithm. We produce the ACMF by reducing this string to a fixed length field of eight bytes by applying algorithm steps 11 through 14.

RESULTS

The ACMF is not always successful in distinguishing substances sharing the same molecular formula; i.e., it is not absolutely unambiguous. The substances shown in Figure 7 are examples of distinct chemical substances which share the same molecular formula and ACMF. The occurrence of such pairs is infrequent in files of chemical substances. The database for the CAS Chemical Registry System consists of over four million distinct substances with a growth rate of 360,000 substances per year; and within this database there are fewer than 1500 examples of distinct substances having identical ACMF's.

While the ACMF is not adequate to support complete substance identification as required by the CAS Chemical Registry System, it serves several useful functions: support

Registry Number: 102-54-5
Molecular Formula: $C_{10}H_{10}Fe$

Figure 8. Ferrocene.

of database integrity, providing for symmetrical substance registration, and maintaining direct access file addresses.

Within the CAS Chemical Registry System, the ACMF is used to support database integrity by retrieving for chemists' review those substances new to the CAS Chemical Registry database which have an ACMF equal to substances already registered in the CAS database. This approach is valid since the ACMF calculation is independent of the unique table generation algorithm. The use of a manual review is practical since there are so few ambiguous ACMF's, i.e., distinct substances with identical ACMF's.

The computer time requirements for the unique table generation algorithm to uniquely label this small number of highly symmetrical substances is prohibitively large. Ferrocene, shown in Figure 8, is an example of a structure which would require 10! or 3,628,800 iterations in order to be uniquely labeled. For substances of this type, which cannot be uniquely labeled within a reasonable amount of computer time, the ACMF is utilized to determine if the substance is new to the CAS Chemical Registry database. The ACMF is utilized to identify file substances which are potentially identical with the candidate substance. For these highly symmetrical substances, final determination of whether the candidate substance is new is made via computer-based atom-by-atom structure comparison of the candidate substance with those file substances identified as potentially identical by the ACMF.

Finally, since the CAS Chemical Registry System utilizes a direct access database, it is necessary to have a precise file address based on structural properties. This is provided by the ACMF.

As noted earlier, the ACMF is not always successful in distinguishing chemical substances sharing the same molecular formula. Since the initial installation of the CAS Chemical Registry System, minor improvements have been made to the ACMF algorithm, but the problems illustrated with the substances in Figure 7 are more deeply rooted and a solution to these problems will require a substantially more complex algorithm. However, since the ACMF fails in so few cases, it provides a practical tool for use with files of chemical substances. In addition, it is a failsafe mechanism because substances are added to the CAS Registry database only after applying the unique table generation algorithm.

ACKNOWLEDGMENT

The development of the CAS Chemical Registry System was substantially supported by the National Science Foundation (Contract C656). Chemical Abstracts Service, a division of the American Chemical Society, gratefully acknowledges this support.

REFERENCES AND NOTES

(1) O'Korn, L. J. "Algorithms in the Computer Handling of Chemical Information", ACS Symposium Series No. 46, "Algorithms for Chemical Computations"; American Chemical Society: Washington, D.C., 1977; p 122.
(2) Dittmar, P. G.; Stobaugh, R. E.; Watson, C. E. "The Chemical Abstracts Service Registry System. I. General Design". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 111-124.
(3) Morgan, H. L. "The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service". *J. Chem. Doc.* **1965**, *5*, 107-113.
(4) Connectivity values are numbers associated with each atom of a substance which are derived as part of the unique table generation algorithm. They have been used in the CAS Chemical Registry System since 1965. These connectivity values reflect the atom-by-atom interconnections but ignore the element value and bond type. The adjective "augmented" indicates the inclusion of the element value and bond-type information in the computation of the connectivity values.
(5) Dittmar, Stobaugh, and Watson have provided descriptions for these special structural characteristics.

Mass Spectral Library Searches Using Ion Series Data Compression

GREGORY T. RASMUSSEN and T. L. ISENHOUR*

Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina 27514

JOHN C. MARSHALL

Department of Chemistry, Saint Olaf College, Northfield, Minnesota 55057

Received October 27, 1978

A series of library searches using a mass spectral data compression method based on fractional ion currents of specific ion series is described. The method offers efficient data compression, reasonable search performance, and a capability for use with file partitioning to reduce search times.

Since its initial development, the computerized search of mass spectral data has gained widespread acceptance as a useful tool for the identification of diverse compounds. The variety of mass spectral search systems reported in the literature and the automatic inclusion of at least one library search program in computerized data systems now commercially available for mass spectrometers attest to the importance of the technique. Computerized searches require that mass spectral data for every compound in a reference library be retained on some storage device peripheral to the computer, such as a magnetic tape or a disk pack. The growth of reference collections of mass spectral data to the extent that such collections routinely contain data for tens of thousands of compounds implies that a large amount of computer-ac-

0095-2338/79/1619-0098$01.00/0