The Third BASIC Fragment Search Dictionary*

W. GRAF, H. K. KAINDL, H. KNIESS, and R. WARSZAWSKI

Basel Information Center for Chemistry (Documentation Center of Ciba-Geigy Ltd., F. Hoffmann-La Roche & Co., Ltd., and Sandoz, Ltd.), CH-4002 Basel, Switzerland

Received October 28, 1981

The development, arrangement, and utilization of the third BASIC Fragment Search Dictionary, presently in use with the CAS ONLINE service[1] for substructure searching in the CAS Registry File,[2] are outlined. Changes with respect to the previous version are mentioned, and some special applications are suggested.

INTRODUCTION

The screen dictionary presently in use with the CAS ONLINE service originated in the experimental CAS Registry II Substructure Search System described at the ACS National Meeting in Atlantic City in 1968 and subsequently by Wigington.[3,4] Work fundamental to the concept of augmented atoms was done by Lynch.[5-7] The linear sequence fragment type (from which BASIC derived the atom-, bond- and connectivity-sequence fragments) was studied by CAS in 1968-1969 but never used in the search system. The generation of improved screens was expensive. Since the screenout of this first dictionary was limited, the subsequent atom-by-atom search was time consuming and uneconomical. In fact, little optimism prevailed at that time with respect to the possibility of future routine operation.

Since 1968, CAS has been providing BASIC regularly with their registry structure file containing connection tables[2,8,9] and with their programs for the conversion of this file. Based on the CAS chemical registry system, a retrieval system consisting of the following three files has been created by BASIC: the fragment mask file, which, by input of fragment numbers, enables us to screen out the major part of irrelevant structures; the connection table file, which, by means of the iterative (atom-by-atom) search gives the possibility for the precise elimination of the remaining false drops, if necessary; the REG/CAN file which links the CAS registry numbers retrieved by one of the foregoing two steps to the corresponding citations, i.e., CANs (Chemical Abstracts Numbers).

For screening of the fragment mask file the BASIC Fragment Search Dictionary is used. It contains the numerical designations of those fragments which were selected from a practical point of view among possible fragments of any of the predetermined types. Utilization of the first dictionary, created in 1973 for searching the CAS registry file, demonstrated already that in 28% of all cases the fragment search alone was sufficient for a precise retrieval. After a thorough revision and addition of new fragment types (Linear Sequences, LS; Hydrogen Augmented Atoms, HA) the second Fragment Search Dictionary was created in 1977. As a result of these improvements, in 68% of all retrievals in the CAS registry file the fragment search alone was sufficient, making a subsequent atom-by-atom search dispensable.[10]

In addition to the BASIC substructure search version of the CAS registry file the three Basel chemical firms have their own structure files of internally synthesized substances, recorded and retrievable on the basis of the CAS registry system. In contrast to the CAS registry file, however, which we can search batchwise only, these are searchable online in a dialogue mode by means of the Interactive Fragment Search (IAFS). Since their size (100 000-200 000 records in each) is small in relation to that of the CAS registry, eventual false drops can easily be eliminated manually. Thus, by use of the second Fragment Search Dictionary, in 90% of all cases no subsequent iterative search would be necessary. Moreover, by use of the IAFS in this manner, selectivity could easily be evaluated and missing additional fragments or fragment types defined.

This second Fragment Search Dictionary was considered sufficient for the BASIC batch version of the CAS registry file as well as for our internal files. It was our goal, however, to optimize it further to allow good results even in future on-line retrievals in the entire CAS registry file using the fragment search alone. It was, therefore, recreated completely and tested in our internal files. The corresponding programs were either written by BASIC or modified from existing CAS programs. Along with these, the dictionary was then put at the disposal of CAS for its CAS ONLINE service. The current CAS ONLINE dictionary has been extended beyond the one described in this paper.

THE THIRD BASIC FRAGMENT SEARCH DICTIONARY

The third BASIC Fragment Search Dictionary contains over 5000 fragments distributed among 2047 fragment numbers. Fragments which by themselves occur only rarely but form a single logical family are collected under a single number (designated by a $ sign in the BASIC edition of the dictionary). The concept used for this purpose is being discussed with the atom sequences and is illustrated for several types of screens. Table I allows a comparison between the second and the third dictionary.

While fragments of the types AS, AA-specified, HA, RC, AC, EC, and GM were subject to revision, including the elimination of superfluous ones and supplementation with necessary additions, the types CS, AA-generalized, TW, and TR were newly introduced. The type BS was expanded greatly by addition of bond values. The types DC and BC became superfluous and were deleted with the exception of a few fragments which were now assigned the type AA-generalized. Multiple occurrences of linear sequences within a structure are ignored (i.e., the count is always 1); they extend, at the most, to the length of six atoms. The meanings of the symbols used in the following description of fragment types are given in Table II. We refrain at this point from defining those types which have already been described earlier.[10,11]

Atom Sequences (AS). Within the atom sequences as well as within other fragment types sometimes superimposed screens are used. Superimposition, i.e., grouping under a single number of fragments which share some features, not only has been dictated by the necessity to save fragment numbers, the

* Correspondence should be addressed to Dr. H. R. Schenk, BASIC, P.O. Box 4043, CH-4002 Basel, Switzerland.

0095-2338/82/1622-0177$01.25/0

Table I. Basic Fragment Search Dictionary

type of fragment	total of allocated fragment numbers: 3rd dictionary (1979)	2nd dictionary (1976)
linear sequence (LS)		
atom sequence (AS)	512	514
connectivity sequence (CS)	215	
bond sequence (BS)	207	7
augmented atom (AA)		
augmented atom, general (AA)	16	
augmented atom, specific (AA)	757	979
hydrogen augmented atom (HA)	113	115
twin augmented atom (TW)	17	
ring		
ring count (RC)	10	10
type of ring (TR)	51	
other fragment types		
atom count (AC)	19	24
degree of connectivity (DC)		23
bond composition (BC)		119
element composition (EC)	111	131
graph modifier (GM)	19	20
total	2047	1942

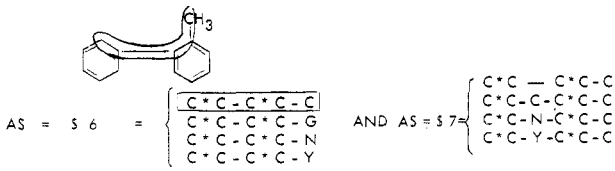


Figure 1.

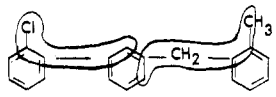


Figure 2.

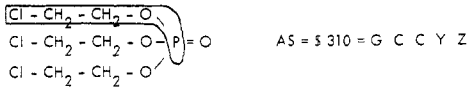


Figure 3.

quantity of which has been limited due to software specifications, but has also allowed us to frame some general questions without resorting to an extensive "OR" logic.

Thus, for example, atom sequences describing ortho substitution of two rings directly joined to each other are represented by the fragment number AS = $6. Under the number AS = $7, ortho substitution of two rings joined to each other directly as well as by mediation of one atom is included. Normally, the use of superimposed screens will lead to satisfactory results. Occasionally, like in the case of using AS = $6 and AS = $7 together for coding the structure in Figure 1, a possible false drop (here containing the sequence C*C-C*C-G as well as C*C-C-C*C-C and shown in Figure 2) may occur. The AS with Z as the endpoint (e.g., AS = GCCCZ) are extended to those having an N or Y adjacent to the Z. Figure 3 shows an example. Further precision can be achieved by using additionally AA = 1759 = 1POOOO together with AS = 326 = G-C-C-Y.

Connectivity Sequences (CS). Connectivity sequences may be used whenever the exact substitution pattern of a substructure region is known. In the connectivity sequences the atoms are not represented by their element symbols but by their connectivity number, i.e., the exact number of nonhydrogen atoms to which a given atom is connected ("non-H count"). This kind of transcription is shown for two isogeometrical structures in Figure 4.

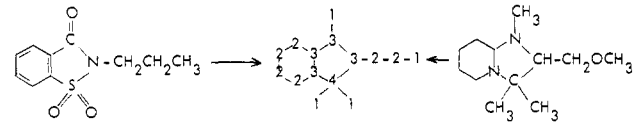


Figure 4.

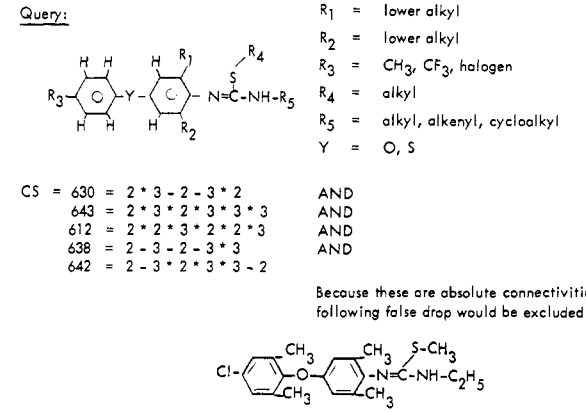


Figure 5.

Figure 6.

From the large number of theoretically possible combinations of three to six atoms we have selected only 196, intending to eliminate some of the shortcomings of the old set of fragment types. Most connectivity sequences consist of connectivity numbers and bond types; in some cases the latter are omitted.

While in atom and bond sequences any higher degree of substitution of the atoms is allowed, the meaning of connectivity sequences is limited to their actual definition. Thus, a connectivity number 1 always means an endpoint of a chain (as a halogen, methyl, hydroxy, amino group, etc.), connectivity number 2 indicates an unbranched node (–CH_2–, –O–, –N=, etc.), connectivity 3 means a branching point connected to three atoms, and connectivity 4 is a quaternary branching point as in the trifluoromethyl, sulfonic acid or sulfonyl, and the phosphoric ester group as well as in spiro rings and so on. Higher connectivity numbers are not included in this fragment type. Using connectivity sequences it is now possible, for example, to select a monosubstituted six-membered ring only or two single-atom substituents in the ortho position of a ring.

Two nitro groups in the 1,4-positions can now be distinguished from amino groups by adding the sequence 3-3*2*2*3-3. Ethyl esters of carboxylic acids may now be distinguished from higher alkyl esters by the sequence 1-2-2-3-1.

It is to be considered that, in contrast to some other screen types, connectivity sequences represent the *exact* sequence at the atomic substitution pattern. While inadvertent use may lead to a loss of what might be relevant answers, a discriminating application of these screens may facilitate or even enable a search by virtue of an additional reduction of the number of candidates prior to iteration. [In the CAS ONLINE service Substructure search (SSS) the CS and TW screens are not being generated automatically from the structural query input and have to be added individually.]

A typical use of CS fragments for a class of structures is shown in Figure 5.

Bond Sequences (BS). These are sequences of three to five bonds between nonhydrogen atoms. The bond types are given

Table II. Definition of the Symbols

A = any atom except H
D = non fusion node (cf. Type of Ring section)
G = common symbol for F, Cl, Br, I
M = common symbol for metals
T = fusion node (cf. Type of Ring section)
Y = common symbol for O, S
Z = common symbol for the nonmetals B, Si, P, As, Se, Te
$ = symbol indicating that several fragments have the same fragment number
* = ring bond
- = chain bond
1, 2, 3, 4 = bond value

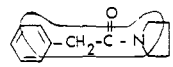


Figure 7.

with or without bond values; the elements connected, however, are never specified. Sequences which are superfluous, too specific or otherwise of little use are not included. Thus, for the structure in Figure 6 the bond sequence A*4A-1A*1A*1A-1A is not present in the dictionary but is represented by the more general form BS = A*A-A*A*A-A*. In a search, the fully defined BS can be approximated, for example by using the combination of BS = 803 = A*A-A*A*A-A AND 798 = A*1A-1A*4A AND 873 = A-1A*1A*1A-1A.

Bond sequences which occur very frequently as well as all those which must be used separately are given separate fragment numbers. Thus, most of the shorter BS fragments have unique and most of the longer ones collective fragment numbers.

In the structure

$$RCH_2CH{=}CHC{\equiv}CCH_2R$$

the longest BS fragment would be

BS = A-1A-2A-1A-3A-1A

Being too specific, however, it is grouped together with other fragments of the type

A-1A-*n*A-1A-*n*A-1A (*n* = 2, 3, or 4)

resulting in the collective number

BS = $932 = A-1A-2A-1A-2A-1A
A-1A-2A-1A-4A-1A
A-1A-2A-1A-3A-1A
A-1A-4A-1A-4A-1A
A-1A-4A-1A-4A-1A
A-1A-3A-1A-3A-1A

To describe the abovementioned structure in a search, this should be combined with the number for the corresponding shorter fragment:

BS = $932 = A-1A-2A-1A-3A-1A AND 910 = A-2A-1A-3A

Sequences of five bonds describing two rings separated by three chain bonds are located in two collective fragment number sets. In the first the ring bonds are varied, and in the second the chain bonds are varied.

To search e.g. the structure of Figure 7 one would use the following two sets in combination:

set 1: BS = 834 (single bond chain between any rings)

A *1A -1A -1A -1A *1A
A *1A -1A -1A -1A *2A
A *1A -1A -1A -1A *4A
A *2A -1A -1A -1A *2A
A *2A -1A -1A -1A *4A
A *4A -1A -1A -1A *4A

AND

set 2: BS = 837 (any chain bonds between specified rings)

A *1A -1A -1A -1A *4A
A *1A -1A -2A -1A *4A
A *1A -1A -4A -1A *4A
A *1A -1A -4A -4A *4A
A *1A -2A -1A -1A *4A
A *1A -2A -1A -4A *4A
A *1A -4A -1A -1A *4A
A *1A -4A -1A -4A *4A
A *1A -4A -4A -1A *4A
A *1A -4A -4A -4A *4A

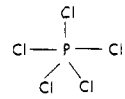
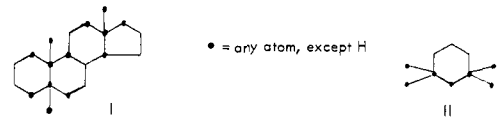


Figure 8.



AA = 949 = 3 A * A * A * A - A (structure I)
AA = 942 = 2 A * A - A - A (structure II)

Figure 9.

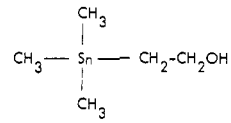


Figure 10.

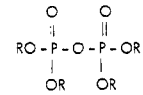


Figure 11.

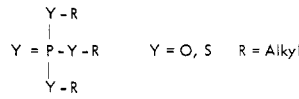


Figure 12.

Augmented Atoms (AA-generalized). In this type of AA fragment the nonhydrogen atoms are indicated by means of the letter A only; their nature is not specified. The attachment of additional (unspecified) atoms is not excluded. The bond type is always indicated; in some cases the bond value is also given.

Some fragments of this type can be of special value when used in the negative mode; i.e., AA = NOT 1A*A*A*A = NOT 940 eliminates all structures having a ring fusion. With the exception of searching for specific structures, negative logic is, in general, dangerous since its inadvertent use may lead to a loss of relevant answers. It may be useful, however, and is indispensable for special applications like those described later (cf. Possible Further Applications of the Third Basic Fragment Dictionary).

The collective AA fragment 950 = 1AAAAAA in the structure of Figure 8 indicates any unspecified nonhydrogen atom surrounded by at least five such atoms (connectivity number 5 or more).

Geminal or angular substitution can be searched in a general manner, i.e., for the structures of Figure 9.

Augmented Atoms (AA-specified). Some important modification and additions have been made. Fragments with symbols like G, M, Y, and Z are less selective than those with the exact symbol. They are, however, more flexible in their applicability and therefore of more general interest. Thus, for the structure presented in Figure 10 the AA = 1 C-SN is approximated by the new general fragments AA = 1390 = 1CM (=carbon bound to metal) and AA = $1392 = 1C-M (=any nonmetal chain bound to metal) and the corresponding fragment numbers of the element composition (EC): EC = $1935 = 1SN (horizontal element group in the periodic table) and EC = $1983 = 1SN (vertical element group in the periodic table). The new fragment AA = $1718 = 1OZZ can be used for searching (Figure 11) , and the fragment AA = 1PYYYY = 1769 can be used for the very frequently occurring search of general formulas, like as in Figure 12. All

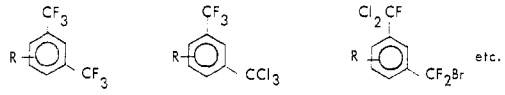


Figure 13.

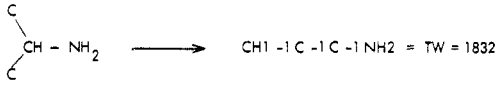


Figure 14.

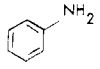


Figure 15.

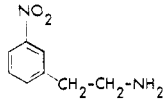


Figure 16.

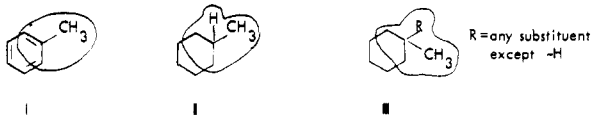


Figure 17.

possible combinations of O and S can be formed by using one fragment number only.

The same search facility is given for polyhalogenated compounds. Thus, for example, the structures presented in Figure 13 can be found by using the general fragment numbers: AA = 1362 = 2CGG and AA = 1364 = 1CGGG.

Twin Augmented Atoms (TW). In some cases it may be useful to specify completely a CH_3, NH_2, or OH/SH group adjacent to an augmented atom. In the resulting "twins" all surrounding atoms as well as the types and values of the bonds connecting these to the central atom are indicated. As in the case of other augmented atom fragments, the notation is linear, the central atom being cited first (Figure 14). By introduction of the TW fragments a specific search of the very frequent NH_2, YH, or CH_3 groups as substituents on a chain or ring carbon became possible. Until now an NH_2 group at a benzene ring (Figure 15) could be approximated only by AA = 1C*4C*4C-1N and AA = 1NH2-1C, fragments which too frequently caused false drops (e.g., Figure 16). The TW fragments in the structures shown in Figure 17 are presented in the Fragment Dictionary in the following form

I: TW = 1C*4C*4C-1CH3

II: TW = 1CH1*1C*1C-1CH3

III: TW = 1CH0*1C*1C-1CH3

Structure II as well as III contains a CH_3 substituent at a ring carbon; the latter structure, however, also contains an additional geminal substitution at this atom. This can be distinguished by means of the specified presence or absence of H substitution at the central C atom (CH1 or CH0, respectively).

Type of Ring (TR). These are closed sequences of ring atoms specified as fusion points (T) or nonfusion points (D). Since these sequences never exceed a primary ring in a ring system, they also define the ring size by their length for three- to seven-membered rings. For larger ring sizes there is one indicator only. Envelope rings are not considered, but the smallest set of smallest rings. Thus, according to this definition the structure in Figure 18 contains no six-membered ring.

The TR fragments do not define the elements, bond values, or the position of substituents at the ring. This must be done

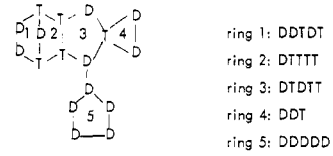


Figure 18.

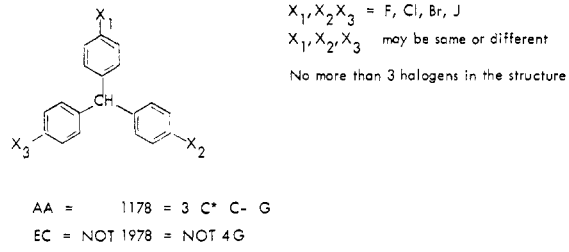


Figure 19.

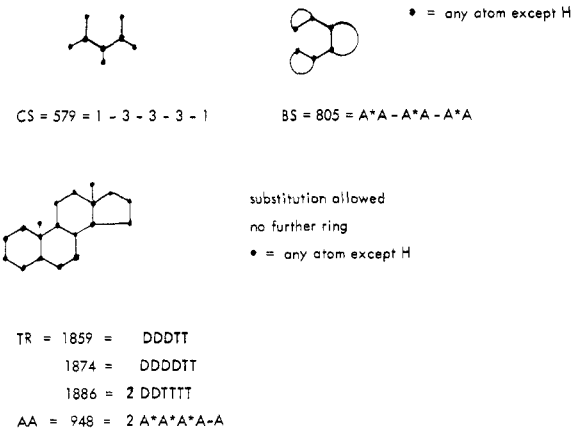
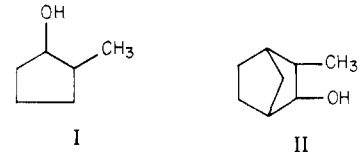


Figure 20.

using additional fragment types. Like the connectivity sequences, the TW and TR screen types are for absolute specificity, and must be used with caution, e.g., if I is a potential answer to the query for II, then the TR screen DDDDD must not be used.



I II

Element Composition (EC). It is now possible to search for halogen general, count >1, by using a single fragment number. Thus, for definition all halogen options for the structure in Figure 19, two fragment numbers are sufficient, obviating the necessity to search all 20 combinations of three halogens and to negate all 35 EC combinations of four halogens by using the OR logic in both cases. All these have already been considered during fragment generation.

POSSIBLE FURTHER APPLICATIONS OF THE THIRD BASIC FRAGMENT DICTIONARY

In addition to the usual kind of substructure searching for literature retrieval, the new dictionary offers the possibility of some special applications. Generally speaking, the fragments can be subdivided into "chemically oriented" and "graphically oriented" ones; the latter are of the CS, RC, TR, AC, and with some limitations also the BS and AA-generalized types. Some examples of these are given in Figure 20.

Thus, for example, by use of these in conjunction with the Boolean AND, OR, and NOT operators (which are not always available in other retrieval systems), various combinations are

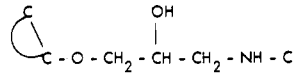


Figure 21.

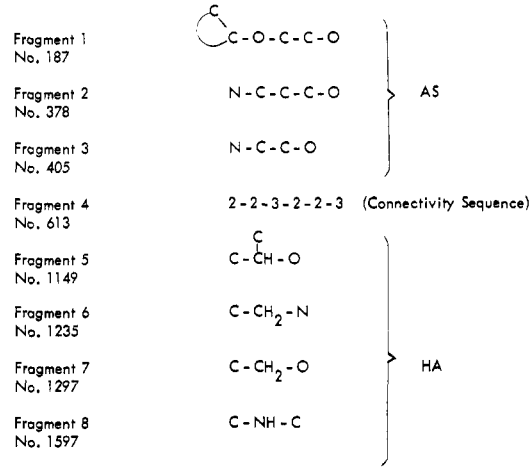


Figure 22.

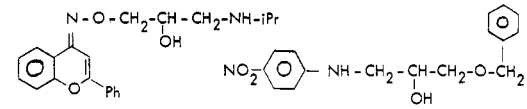
possible to find classes of compounds which may share or exclude certain structural characteristics not necessarily related to the nature of the atoms involved or located in the same manner.

Of possible interest to the industrial chemist could be retrieval with systematic variation of structural characteristics, as illustrated below. As a prerequisite for this procedure, connection tables from which the fragments are derived are structural representations of single compounds (not Markush formulas or substance classes). Thus, the Boolean NOT may be used for the selective elimination of definite features within a single structure.

As a rule, a research chemist tries to modify existing (for example, biologically active) compounds in order to find new ones with improved properties. These share in most cases some structural characteristics in common with the lead compound. In the process of modification it is left to the intuition of the chemist to find areas not already covered by his competitors. This may be facilitated, as shown in the following example.

To find substances with β-sympatholytic activity the following group, characteristic for this type of compounds, is subjected to systematic alteration (Figure 21). From this substructure eight typical fragments have been chosen (Figure 22). To find similar compounds, the search was then limited to structures containing in each case only seven of these fragments. Thus, in each of the searches described below another one of the eight fragments was omitted.

Using the CAS ONLINE service [containing at that time about 1.9 million substances (January 1981)], we carried out nine searches. In the first of these, structures containing all eight fragments were searched, yielding 3474 substances with the substructure of Figure 21. In the second, all fragments with the exception of the first (fragment 1) were used, yielding 3760 substances. From these, by use of the NOT logic, those of the first search were subtracted, yielding 286 structures missing the feature of fragment 1. These were tested visually. Some typical representatives are shown in Figure 23. This big number of externally known substances indicated that the



Reg. No. 66928-88-9 Reg. No. 71106-77-9

Figure 23.

modification in question has already been subject to much attention.

The analogous procedure was used for searches in which the fragments 2–8 were singly excluded, yielding, respectively, 22, 67, 7, 12, 0, 7, and 95 hits.

To find substances suitable for biological screening, we carried out an analogous series of retrievals in our internal file which has been constructed on the basis of the same fragments as the CAS ONLINE service. Structures with characteristics corresponding to those in the CAS registry file were found in six of those searches. In this manner the novelty and frequency of citations for given molecular modifications may be examined; at the same time, screening of internally available substances may help to validate or disprove a theory.

ACKNOWLEDGMENT

We are greatly indebted to CAS for placing at our disposal files, information, and software which constitute the foundation of our work and for the most valuable stimulus of numerous personal contacts and discussions. We thank the management of BASIC for its encouragement and provision of resources, Dr. J. Haeuser of the data processing department of Ciba-Geigy Ltd. for development and creation of the software which made generation of the third fragment dictionary possible, and Dr. R. J. Rowlett of CAS for reviewing the draft of this paper and for many helpful suggestions.

REFERENCES AND NOTES

(1) Farmer, N. A.; O'Hara, M. P. "CAS ONLINE, a New Source of Substance Information from Chemical Abstracts Service". *Database* **1980**, *3*, 10.
(2) Dittmar, P. G.; Stobaugh, R. E.; Watson, C. E. "The Chemical Abstracts Service Chemical Registry System. I. General Design". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 111.
(3) Wigington, R. L. "Machine Methods for Accessing Chemical Abstracts Service Information". Proceedings of the IBM Symposium on Computers and Chemistry, IBM Data Processing Division, White Plains, New York, 1969.
(4) "System Documentation for the CAS Registry System"; Chemical Abstracts Service: Columbus, OH, 1968.
(5) Lynch, M. F.; Orton, J.; Town, W. G. "Organisation of Large Collections of Chemical Structures for Computer Searching". *J. Chem. Soc. C* **1969**, 1732.
(6) Crowe, J. E.; Lynch, M. F.; Town, W. G. "Analysis of Structural Characteristics of Compounds in a Large Computer-Based File. I. Noncyclic Fragments". *J. Chem. Soc. C* **1970**, 990.
(7) Adamson, G. W.; Lynch, M. F.; Town, W. G. "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. II. Atom-Centered Fragments". *J. Chem. Soc. C* **1971**, 3702.
(8) Gluck, J. D. "A Chemical Structure Storage and Search System Developed at Du Pont". *J. Chem. Doc.* **1965**, *5*, 43.
(9) H. L. Morgan, "The Generation of a Unique Machine Description for Chemical Structures. A Technique Developed at Chemical Abstracts Service". *J. Chem. Doc.* **1965**, *5*, 107.
(10) Graf, W.; Kaindl, H. K.; Kniess, H.; Schmidt, B.; Warszawski, R. "Substructure Retrieval by Means of the BASIC Fragment Search Dictionary Based on the Chemical Abstracts Service Chemical Registry III System". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 51.
(11) Schenk, H. R.; Wegmüller, F.; "Substructure Search by Means of the Chemical Abstracts Service Chemical Registry II System". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 153.