Substructure Retrieval by Means of the BASIC Fragment Search Dictionary Based on the Chemical Abstracts Service Chemical Registry III System

W. GRAF, H. K. KAINDL, H. KNIESS, B. SCHMIDT, and R. WARSZAWSKI*

BASIC, 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 July 24, 1978

The fragment search, used as a primary screen in the topological Substructure Search System developed in Switzerland and based on CAS Registry data and programs for the entire CAS Registry Structure File, has been further refined. The development and further improvements of the BASIC Fragment Search Dictionary are discussed and possible application for on-line retrieval is envisaged.

1. INTRODUCTION

The Chemical Abstracts Service (CAS) Chemical Registry System comprises the structures of all those compounds which have been described in the chemical literature cited in *Chemical Abstracts* since 1965 (at present, over 4 million compounds). The annual increase of about 360000 compounds originates from publications in 14000 journals of 150 countries and from patents of 26 countries. The system, which provides representation of chemical structures in the form of topological tables, is based on the process of Gluck.[1] The unique and unambiguous character of these "connection tables" is achieved by numbering the single atoms according to the Morgan algorithm.[2]

Based on the CAS Chemical Registry System, a retrieval system was created by BASIC via the following steps.

- Since 1968, CAS has provided us semiannually with their Registry Structure File containing connection tables.
- CAS implemented the Registry III system in 1974, but because our programs were written for the previous CAS Registry II format, the Registry Structure File is converted in Basel into a version corresponding largely to the Registry II format. Our version of the Registry Structure File differs from the original mainly in the area of multicomponent compounds (i.e., salts, addition compounds, etc.). Each multicomponent compound is assigned a Registry Number by CAS and, in addition, each of their components is assigned a Registry Number. In our case these compounds are recorded under the Registry Numbers of their components only. The corresponding CAS Registry Number of the entire compound may then be retrieved in a further step following the search. Another difference between the two files is that only completely defined structures are contained in the BASIC version.
- From the BASIC version of the Registry file, various kinds of fragments are recognized and generated by algorithms originally defined by CAS.
- Subsequently, those fragments of predetermined type which are selected from a practical point of view and constitute our Fragment Dictionary are sorted out. In this manner the Fragment Mask File is created.

Thus, for structure searching, two files are at our disposal:

- The Connection Table File for searching by input of partial or entire structures (the so-called topological, atom-by-atom, or iterative search).
- The Fragment Mask File containing on the average 100–500 fragments per record of a medium-sized structure for searching by input of fragment numbers, according to their designation in the Dictionary (the fragment search).

In practice, the fragment screening is the first search, leaving seldom more than 1% of the file to the iterative search. The iterative search is uneconomical with respect to computer time and is used for the precise elimination of false drops in this remaining portion only.

The iterative search yields a printout containing a list of the retrieved CAS Registry Numbers; the corresponding structures may be scrutinized on a microfilm which CAS has put at our disposal. The corresponding citations, i.e., the Chemical Abstracts (CA) Numbers, however, are not obtained. The latter are linked to the pertinent CAS Registry Numbers in a separate Registry Number/CA Number (REG/CAN) file which may be searched automatically in adjunction to each substructure search to obtain a list of abstract numbers, each of them being linked to the pertinent CAS Registry Numbers.

Thus, a retrieval may consist of three steps:

(1) The fragment search (screens the entire file and yields, if required, CAS Registry Numbers of the resulting "candidates")
(2) The iterative search (yields the CAS Registry Numbers of the precise hits)
(3) The REG/CAN search (yields CA citations pertaining to the CAS Registry Numbers of the hits)

This possibility has been implemented for the file in CAS Registry II[3] format and is still available in the case of Registry III. As before in the case of Registry II, CAS has put at the disposal of BASIC the file in Registry III format for substructure retrieval purposes.

2. THE IDEA OF A NEW BASIC FRAGMENT SEARCH DICTIONARY

We prepared the first Fragment Search Dictionary in 1973 for searching in Registry II, having no practical retrieval experience over the whole file. This may have been of minor importance in view of the fact that our objective was limited to the implementation of an iterative search at a reasonable cost. The dictionary in question was based on a CAS Dictionary containing about 1200 fragments. In view of the high generation costs involved, we even refrained from using additional, more selective fragment types. Although this first dictionary was conceived under unfavorable circumstances, surprisingly, it has already produced a screenout of 99.6% on the average (leaving about 10000 structures for an iterative search in a file containing 2.4 million compounds at that time). Moreover, in 28% of all cases no iterative search was necessary, since the resultant number of answers was so small that it could be reviewed manually and consisted in most of the cases of precise hits.

With the introduction of the CAS Registry III System,[6] topological recording of chemical structures has been refined

* Correspondence should be addressed to Mr. D. Ligtenberg, BASIC, Secretary, P.O. Box 273, CH-4002 Basel, Switzerland.

0095-2338/79/1619-0051$01.00/0

and unified; in addition, some principal differences with respect to Registry II have been established. In view of this, and in order to eliminate shortcomings recognized in the old Dictionary, the need to create a new one became evident. Our prerequisites have been incomparably better this time. Aside from the experience gained with a wide range of fragment types, we had a working knowledge of the frequency and selectivity of each fragment and knew the gaps in the old Dictionary. Moreover, the generation of special fragment types, too costly before, became attainable owing to the availability of improved computer programs.

Thus, a new objective was envisaged to attempt to make the iterative search increasingly dispensable through stepwise improvements of the Fragment Dictionary, since

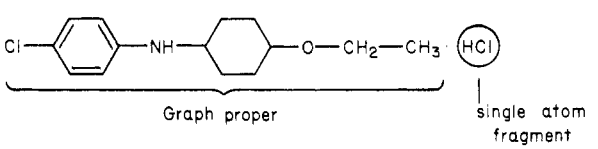
- a mere fragment search system is easier to learn and may be made accessible to a wide range of users
- a fragment file may be inverted, i.e., sorted according to fragment numbers instead of Registry Numbers, and, as such, is predestined for on-line retrieval (this step has already been successfully implemented in an internal file of about 100 000 structures)
- such a fragment file could easily be linked to other, nonstructural data, for example, entries in the CA Subject Index Alert (CASIA) or *CA Condensates* (CACon)
- in those few cases where a fragment search alone would not be sufficient, an improved Fragment Dictionary would simplify coding for the subsequent iterative search by limiting it to such parts of the substructure where fragment coding is not exhaustive or precise enough.

In contrast to older fragment codes like that of DERWENT, designed for specialists and, in most cases, adjusted to the punched card format, other requirements should be met by a code generated by computer. To this end, in analogy with the previous Dictionary, the following selection criteria were applied.

- All fragments must be inclusive, i.e., those with a higher degree of specification must be implicit in those with lower specification.
- Fragments occurring too rarely are to be replaced by others of a more general nature or grouped under single collective bit numbers.
- The Dictionary should cover as large as possible an area of chemistry in as uniform a manner as possible. The selection, therefore, should not be biased in favor of such areas of interest which are specific to the chemical companies in Basel.
- The fragments should be normalized in a general schematic form. Instead of individual definitions for each fragment, fragment types must be created which may be generated by as few simple programs as possible. In view of the large number of fragments, the arrangement of a dictionary must be lucid; of crucial importance are separate groupings and a unified alphanumeric sequence for each fragment type. This allows an immediate utilization of the Dictionary without prior detailed study by an information scientist. Fragments with special definitions, though technically not excluded, are expensive and would have to be learned individually.

3. ARRANGEMENT AND UTILIZATION OF THE NEW BASIC FRAGMENT SEARCH DICTIONARY

Until now, only the seven types of fragments illustrated in the following example were available:



Graph proper is that portion of the computer structural record which identifies the atom number, connection, bond, H-count, and ring closure pairs for each Registry Number. Elements are represented by their atomic symbols, or, in some cases, summarized as G (=halogen), Y (=O, S), or Z (=P and other nonmetals). A ring bond is represented by an asterisk (*) and a chain bond by a hyphen (-) (bond operators). The bond value is indicated by the numeral 1, 2, 3, or 4 for the single, double, triple, or "nonlocalized" bond (an aromatic bond or one with a migrating H or a charge), respectively.

ATOM COUNT = AC. Number of atoms present (excluding hydrogen) in the graph proper, e.g., AC = 17.

RING COUNT = RC. Number of rings present, e.g., RC = 2.

DEGREE OF CONNECTIVITY = DC. Count of the branching atoms, with an indication of the minimum number of nonhydrogen atoms to which a branching atom is bound (only ≤ 3 and ≥ 6 atoms are considered); e.g., 4 times 1 atom bound to 3 atoms: DC = 4 × 3.

BOND COMPOSITION = BC. Number of bonds present (excluding those to hydrogen atoms); consists of bond operator and bond value.

*	ring bond (value not defined)
-	chain bond (value not defined)
*1	single ring bond
-1	single chain bond
*2	double ring bond
-2	double chain bond
*3	triple ring bond
-3	triple chain bond
*4	aromatic or alternating (also equalized or delocalized) ring bond
-4	equalized or delocalized chain bond

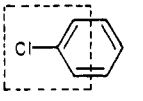
e.g., BC = 12*
= 6*1
= 6*4
= 6-1

ELEMENT COMPOSITION = EC. Number of elements present. The most common elements are directly searchable by their specific fragment number; the others, by combining the vertical and horizontal line in the periodic table, e.g., EC = 1 C1.

GRAPH MODIFIER = GM. Fragment type used to specify structural characteristics identified in the abnormal mass, abnormal valence, charge, and single atom fragment (SAF) fields, e.g., GM = 1 C1.

AUGMENTED ATOM = AA. Atom surrounded by other directly bound atoms. Surrounding atoms which are not mentioned are not excluded. In this notation the central atom is cited first, e.g.,

AA = 6 C*4C*4C
AA = 1 C*4C*4C-1Cl



For some of the AA-Fragment types, the bond operator (-, *) and, eventually, bond values (1, 2, 3, 4) are indicated. The count always specifies the number of such fragments present.

We were able to utilize our retrieval experience for an improved selection of fragments. In doing so, about 360 positions of the old Dictionary could be vacated and saved for other purposes. Using the space gained in this manner, two new fragment types were introduced. In many cases not all ligands of the central atom were defined in the AA type, and such undefined ligands could imply hydrogen or any other

atom. In the Hydrogen Augmented Atom (HA) type, both are fully defined. Thus, for example, for the definition of the aldehyde (C—CH=O) group the fragment

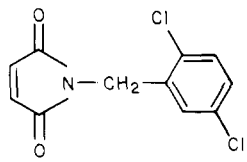
AA: C-1C-2O

was available only, implying merely that a central atom (written on the left in the AA convention) is singly bound to another carbon and doubly bound to an oxygen atom. Such a fragment, however, is present in ketones, carboxylic esters, and others as well. The newly introduced fragment, on the other hand,

HA: CH1-1C-2O

defines the aldehyde group exclusively. Thus, HA fragments provide the possibility of searching essential groups like hydroxyl, primary and secondary amines, etc., in a selective manner.

Linear Sequences (LS) define sequences of four, five, or six atoms in the same order as in the structure, contributing substantially to the description of the topological context in which the Augmented Atoms appear. The structure



contains among others the following Linear Sequences:

Y-C*C*C*C-Y
Y-C*N*C-Y
Y-C*N-C
C*N-C-C*C
N-C-C*C-G
N-C-C*C*C-G
G-C*C*C*C-G

As in Augmented Atoms, generalized Linear Sequence fragments in which the bond operator is not defined, are also present, such as

Y C N C Y

In addition to these, there are sequences present which describe the bond operator only

* - - *

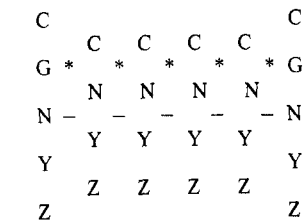
which implies that two rings are bound to each other through an intermediate atom (in the above example, the carbon of the CH_2– group).

We have refrained from specifying the bond values in the Linear Sequences in view of the large number of theoretically possible variations. In addition, the symbol Y is used collectively for oxygen and sulfur, G for all halogens, and Z for phosphorus and other nonmetals. If a further distinction of atoms or bonds is required, other fragment types (primarily Augmented Atoms) must be used.

While the compilation of HA fragments did not pose particular problems, the selection of Linear Sequences proved to be much more difficult. Considering sequences of four, five, or six atoms, the C, N, G, Y, and Z symbols, bonds with no value, and taking each sequence once only, the number of possible variations alone grows into thousands. In addition, all other fragment types have to be taken into account to avoid, if possible, the concomitance of such LS and AA fragments which could serve the same purpose equally well. Thus, for a sequence of six atoms containing element symbols as well as bond operators and taking into account that G must be at an end position and bonded by a chain bond only, 82 232 variations are possible.

Table I

fragment type	no. of fragments
AC (ATOM COUNT)	24
RC (RING COUNT)	10
DC (DEGREE OF CONNECTIVITY)	23
BC (BOND COUNT)	119
EC (ELEMENT COMPOSITION)	131
GM (GRAPH MODIFIER)	20
AA (AUGMENTED ATOM)	979
HA (HYDROGEN AUGMENTED ATOM)	115
LS (LINEAR SEQUENCE)	521
Total	1942



If bond operators or element symbols are omitted, 3240 or, respectively, 20 possibilities remain.

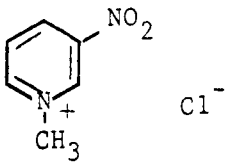
The following procedure was chosen: by chemically significant combinations of the symbols C, G, N, Y, and Z, sequences of four, five, and six atoms were created and listed without bonds initially. Following a test of their utility, those sequences which proved too special were erased. The same kind of selection was repeated if bond operators were added. The final choice was based on a comparison with AA fragments already present in the old Dictionary. Independently, all 1666 fragments of the old Dictionary were examined, missing ones were added, and others with low selectivity eliminated. In Table I, the number of fragments of each type and their total number in the new Fragment Search Dictionary are listed.

In spite of the program limitation to about 2000 fragments, a structure record in the Fragment Mask File is very redundant in comparison to what it would be in a classic fragment code. This is one of the consequences of the inclusivity principle which states that fragments with a high degree of specification are implicit in fragments with a low degree of specification. Thus, for example, the count of each atom always represents the *minimum* condition only. In a record of a molecule with ten atoms, the most specific "Atom Count 10" fragment is implicit in each of the entire Atom Count series (at least 1, at least 2, at least 3, ..., up to 10) and all these fragments must be searchable. Thus, a relatively small structure like that of 1-methyl-3-nitropyridinium chloride contains 101 fragments (see Figure 1). Especially redundant are small counts and nonspecific fragments already included in higher counts and more specific fragments and, therefore, used in queries of a very general character only. If fully defined structures (or, with caution, Markush formulas) are searched, it is advantageous to utilize the Boolean NOT logic.

For an unambiguous search of 1-methyl-3-nitropyridinium chloride, a combination of the following six fragments is fully sufficient (others, however, are equally possible):

AC: not 11
GM: C1 single
AA: N-1C-2O-2O N*4C*4C-1C
HA: CH3-1N
LS: C-N*C*C*C-N N*C*C*C*C*C

Optimal coding is especially easy in on-line retrieval.



```
AC (10 fragments):  minimum count 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
BC ( 1 fragment ):  minimum count 1
DC ( 3 fragments):  minimum count 1, 2, 3 branched atoms
BC (20 fragments):  minimum count 1 *, 2 *, 3 *, 4 *, 5 *, 6 * (ring bonds general)
                    1 *4, 2 *4, 3 *4, 4*4, 5*4,6 *4 (aromatic type ring bonds)
                    1 -, 2 -, 3 -, 4 -, 1 -1, 2 -1, 1 -2, 2 -2 (chain bonds)
EC (10 fragments):  minimum count 1 C, 2 C, 3 C, 4 C, 5 C, 6 C, 1 N, 2 N, 1 O, 2 O
GM ( 3 fragments):  1 Cl single atom, 1 non del.charge, 1 non table valence
AA (42 fragments):
```

min.count	item	min.count	item	min.count	item
2	C *4C	1	C *4N	1	N C O
1	C C C	1	C - N	1	N - C- O
1	C * C *C	1	C -1N	1	N -1C-2 O
1, 2, 3	C *4C*4C	1	N C C	1	N C O O
1	C C C N	1	N * C * C	1	N - C- O- O
1	C * C* C- N	1	N * C- C	1	N -1C-20-20
1, 2, 3	C C N	1	N *4C*4C	1	N - O
1, 2	C * C *N	1	N *4C-1C	1	N -20
1, 2	C *4C*4N	1	N C C C	1	N O O
1, 2, 3, 4	C N	1	N * C* C- C	1,2	O N
1	C * N	1	N *4C*4C-1C		

```
HA ( 1 fragment ):  1 CH3-1N
LS (11 fragments):
```

- * * -	C*C*C*C*N	C*C-N-Y	N C C N Y
C*C*C*C*C	C*C*N*C*C	C-N*C*C-N	N*C*C-N-Y
C*C*C*C*C*N	N C C C C N	N*C*C-N	

Figure 1.

Table II

	no. of searches	screenout	cand./hit	applied fragments/ question
CAS Registry II Test	94	99.807	88.60	13
CAS Registry III Test	94	99.971	12.50	20

Table III

	no. of searches	screenout	cand./hit	applied fragments/ question
CAS Registry II 1975-1976	1014	99.615	25.47	7
CAS Registry III	912	99.932	8.64	11

4. RESULTS ACHIEVED WITH THE NEW BASIC FRAGMENT SEARCH DICTIONARY

To test the efficiency of the new Fragment Search Dictionary, retrievals for 94 real questions already processed in the Registry II file were repeated utilizing the CAS Registry III system. The results are summarized in Table II where the candidates are the structures resulting from the fragment search. As indicated above, the relation of candidates to hits shifted greatly in favor of the fragment search: instead of 88, only 12 false drops accompanied each hit. In 1977, 912 searches in the CAS Registry III file were commissioned for the three chemical companies in Basel. The results of a comparison with earlier such retrievals in the CAS Registry II file are shown in Table III. The results are similar, the candidates/hits ratio being 8:1 (it was 25:1 before). Since, previously, in 28% of all retrievals the fragment search alone was sufficient, it was of great interest to know whether the introduction of the new Fragment Search Dictionary increased this proportion. To facilitate this, we have continued using the iterative search, refraining from using it in those cases only where a low number of candidates was expected. It turned out that in 68% of the 912 searches the fragment search alone was sufficient. The need for a subsequent iterative search, however, is not determined by the candidates/hits ratio alone, but also by the absolute number of the resulting candidates. If, for example, 500 candidates result, an iterative search is indicated in spite of the favorable ratio of 2:1, but would not be necessary if the resulting number were smaller. Some examples are given in Table IV. For practical purposes, the coding effort must be weighed against the size of the expected

Table IV

cand./question	hits/question	iterative-search necessary?
500	250	yes
500	480	no
30	1	no

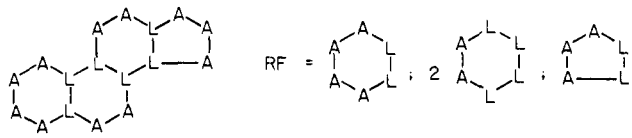
output. Less sophistication may be justified if few candidates result, and hits may be selected by viewing the corresponding structures. The optimal balance between coding and evaluating, however, will be found only if the possibility of on-line retrievals is given.

5. POSSIBLE SUPPLEMENTATION OF THE NEW BASIC FRAGMENT SEARCH DICTIONARY

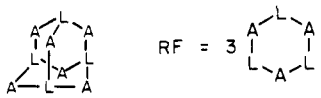
According to the first test results, in 32% of all cases an iterative search was still necessary. An analysis of those retrievals, where coding by the existing fragment types was impossible or insufficient, revealed the missing structural characteristics.

A. Ring Information. The fragment types available until now do not supply the most essential information like ring size or degree and points of fusion. Thus, for example, such characteristic features as a cyclopropane ring or an adamantane skeleton can be described in special cases only. To improve the selectivity in this respect, the following suggestions are made.

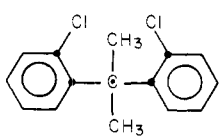
- A combination of the available Dictionary with such types of ring fragments which are already well known and, in part, generated:
  - Nomenclature (thering Parents of CAS, for example).
  - Schematic ring information, for example, of the kind used already in the NIH/EPA (National Institutes of Health/Environmental Protection Agency) or NLM (National Library of Medicine)–CHEMLINE system.
- Deduction of new ring bits, like Ring Fusion Points (RF) from the connection tables:
  - The RF type would define the ring size (3-, 4-, 5-, 6-, and 7-rings) as well as fused or bridged systems. To be generally appreciable, it would not contain any element symbols. RF fragments would allow, for example, an improved description of the following structures:



A = isolated ring atom;
L = ring fusion point



B. Connectivity Sequences. Another type of structure is difficult to define because the branching points in the molecule cannot be located.

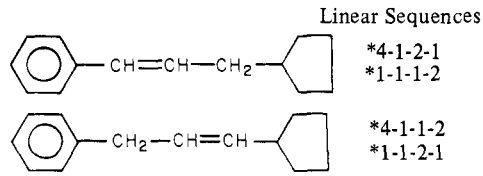


This structure may be described using the Augmented Atom C–C–C–C–C and the Linear Sequence C*C–C–C*C, but not as a quaternary C between two rings. This could be remedied by generation of selective "Connectivity Sequence". Using the Connectivity Sequence, the above structure could be defined as:

3*3-4-3*3

The numerals represent the number of non-H-atoms bound to each atom.

C. Bond Type Sequences. Another selective and easily comprehensible fragment type could be Linear Sequences based on bond types only. Such Linear Sequences would allow the distinction between the following structures:



All these new fragment types would increase selectivity, while at the same time being generally applicable and easily comprehensible, an objective of our Dictionary. Introduction of these additional fragment types should be especially worthwhile in view of a future on-line availability of the CAS Registry file, which is under consideration.

On-line operation seems advantageous to us, since, in a dialogue, the efficiency of fragments could be fully exploited. In addition, a link to other CAS files, like CASIA, would be possible. Introduction of the additional fragment types is not a problem since they can be generated from the existing connection table file at any time. The space available in our Fragment Search Dictionary is sufficient, since, without changing the present programs, about 100 bit numbers are not yet occupied. Should these reserved positions be insufficient, new ones could be made vacant by additional elimination of rarely used fragments or combining these. The new types would possibly make some old fragments, e.g., the Degree of Connectivity series and, partially, Bond Composition, superfluous.

ACKNOWLEDGMENT

BASIC would like to thank CAS from making available the Registry and REG/CAN Files, along with the software, and especially for the fragment generation programs from which the main portion of our programs was derived. Further BASIC acknowledges the continuous support as well as technical advice given by CAS staff which made it possible to handle the data in the new Registry III format using the slightly modified Registry II software. Specialists from the data processing departments of the three BASIC companies have made important contributions to the subject of this publication.

REFERENCES AND NOTES

(1) J. D. Gluck, "A Chemical Structure Storage and Search System Developed at Du Pont", *J. Chem. Doc.*, **5**, 43 (1965).
(2) H. L. Morgan, "The Generation of a Unique Machine Description for Chemical Structures. A Technique Developed at Chemical Abstracts Service", *J. Chem. Doc.*, **5**, 107 (1965).
(3) H. R. Schenk and F. Wegmüller, "Substructure Search by Means of the Chemical Abstracts Service Chemical Registry II System", *J. Chem. Inf. Comput. Sci.*, **16**, 153 (1976).
(4) G. G. Vander Stouw, C. Gustafson, J. D. Rule, and C. E. Watson, "The Chemical Abstracts Service Chemical Registry System. IV. Use of the Registry System to Support the Preparation of Index Nomenclature", *J. Chem. Inf. Comput. Sci.*, **16**, 213 (1976).
(5) "Substructure Search", 2nd ed, Background Information and Question Coding Instructions, American Chemical Society, Washington, D. C., 1970.
(6) P. G. Dittmar, R. E. Stobaugh, and C. E. Watson, "The Chemical Abstracts Service Chemical Registry System. I. General Design", *J. Chem. Inf. Comput. Sci.*, **16**, 111 (1976).