A Simplified Chemical Structure Fragmentation System*

J. T. MAYNARD
Elastomer Chemicals Department, E. I. du Pont de Nemours & Co., Inc., Experimental Station, Wilmington, Del. 19898

Received April 1, 1970

A fragmentation scheme for the indexing of chemical structure information is proposed. The system is based on that described by Edge, Fisher, and Bannister in 1957, the major change being that functional groups are described as existing, if one of 13 common groups, and otherwise as derived from the six most common groups, i.e., $-NH_2$, $>NH$, $>C=O$, —COOH, —OH, and —SH, by the elimination of the elements of water or hydrogen. This information, combined with an indication of the elements present in each group, provides a simple and effective description of functional groups and eliminates the need for an open-ended thesaurus of named groups. Additional valuable information is provided by using in the formula atom count nonskeletal carbon and nonskeletal hydrogen. The advantage of this system is that the entire set of rules for indexing functional groups is presented in one page, thus ensuring easy understanding and consistent application by indexers. Discrimination between related compounds is good. It is found that most organic compounds can be described by an average of eight to 12 indexing terms.

In the evolution and proliferation of chemical notation systems over the past two decades, the greatest attention in recent years has gone to linear notations and topological coding systems.[1-3] The goal of the proponents of these systems has been to provide a means of describing the structure of a chemical compound in a machine-manipulateable language of such a nature that the compound's structure can be completely regenerated from the stored information. It seems to be taken as an article of faith that the capability of regenerating a structure is essential to the usefulness of a chemical coding system. However, this may not be so, and the purpose of this paper is to examine the usefulness of a simple fragmentation code that stores enough descriptive terms to ensure retrieval of compounds that meet search criteria but that is not necessarily able to provide all the information needed to recreate the structure of each encoded compound.

In spite of the extensive attention given to linear notations and topological coding methods, it is quite apparent that many individual chemical information systems, perhaps a majority of the smaller systems, still depend on some type of a fragmentation code. There seem to be several reasons that fragmentation systems remain popular.

They employ the "natural language" of chemically-trained scientists.

They are readily tailored to the information needs of relatively narrow fields of chemical interest.

The coded information can be manipulated with relatively simple equipment such as notched or punched cards.

They are especially useful in handling generic information, which is often encountered in the patent literature.

A generally applicable chemical structure fragmentation system was devised in the early 1950's by Edge, Fisher, and Bannister.[4] It has served the rather wide-ranging needs of the Du Pont Central Research Department well and has been adapted to the computer-based system of the Du Pont Central Patent Index.[5] Its major weakness is that every functional group must be individually named and assigned a code number or term. Experience has shown that such an open-ended collection of terms becomes unwieldy—more than 4000 such terms are now in the system. Many are encountered only rarely, and the chances for coding error are great.

It is proposed here that a much simpler approach to the handling of functional groups, combined with certain minor changes in the system of Edge *et al.*, can give a generally useful fragmentation method that will meet the needs of many small-to-medium-sized chemical information systems.

GENERAL CONSIDERATIONS

Let us consider the objectives of a coding system. In the case of specific compounds, we want to describe the compound in the fewest possible terms that will distinguish it from all the several million other known compounds. Our zeal in this aim may need to be tempered by the cost of storing and searching a large file of terms. We may have to accept a certain amount of false drop. What we are after is a series of screens that will eliminate from consideration all compounds except the one of interest, possibly with a few close relatives.

It is immediately apparent to a practicing chemist that the most powerful screen is a molecular formula index. Inspection of the formula indexes of *Chemical Abstracts* shows that the number of compounds having the same molecular formula reaches a peak at compounds having about 12 carbon atoms, where a maximum of about 200 different compounds will be found that have the same

* Presented before the Chemical Documentation Section of the Fifth Middle Atlantic Regional Meeting, ACS, University of Delaware, Newark, Delaware, April 1, 1970.

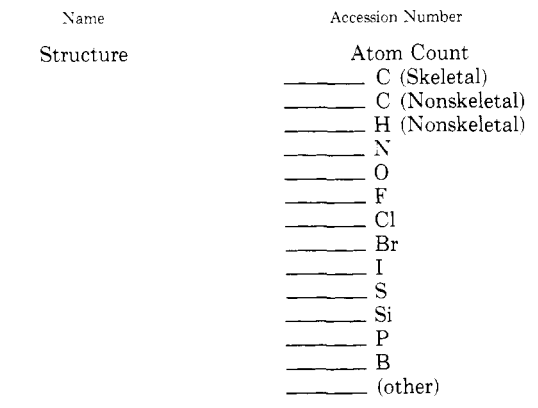


Figure 1

formula. Thus, molecular formula alone is the most powerful tool we have and should be the first screen in any index of specific compounds. Clearly, it should take only a few further screening terms to provide a near-unique description of any given compound, and the design of an indexing system should be based on this viewpoint.

Molecular formulas are obviously of limited utility as a screen in describing generic groups of compounds, but here again it is apparent that it should take only a few screening terms to describe compounds that meet given structural specifications, if it is not taken as a criterion that it must be possible to reconstruct the structure from its coded description. Our purpose should be to locate the one or more compounds or generic structures that have features corresponding to the search question, which may be highly specific or quite general. This will often call into play only a very few indexing terms.

The coding of ring systems has preoccupied the designers of many fragmentation systems—for example, the Ring Doc code used in the pharmaceutical field. This seems unnecessary in view of the existence of the Ring Index.[6] The present system simply uses the appropriate Ring Index number, with the additional terms needed to put the ring in context, thus taking advantage of a complete, systematic compilation that is universally available.

The final and perhaps most important consideration is that an indexing system should have the fewest possible rules, so that they are easily learned and uniformly applied. In the system proposed here, all the rules for describing functional groups can be presented in one page, and the remainder of the rules are self-evident from the terms provided on the indexing forms (Figures 2 and 3).

THE SYSTEM

The information that is indexed includes for specific compounds a modified atom count (molecular formula), and for both specific and generic structures a description of functionality, of ring systems present, and a set of descriptive configurational terms.

Indexing is recorded on a one-page form combining Figures 1, 3, 4, and 5. Each compound is assigned an accession number. Atom count is recorded as in Figure 1. A feature that adds considerable screening effectiveness to the usual simple molecular formula is the distinguishing of skeletal and nonskeletal carbon atoms. Thus, any carbon that is part of a hydrocarbon chain or ring is skeletal, while those that are part of functional groups such as —COOH or —NCO or —$NHCONH_2$ are nonskeletal. In keeping with the present trend to not storing redundant information, skeletal hydrogen atoms are not indexed. However, nonskeletal hydrogens are counted, because they provide significant screening information. The actual count of all other atoms is recorded.

Functional groups are indexed according to the rules shown in Figure 2, and are recorded on the indexing sheet as shown in Figure 3. This method is the major difference from the system of Edge *et al.* Only the most common, simple functional groups are indexed directly. The very large number of more complex functional groups are described by the overlapping fragments from which they can be regarded as derived by the elimination of the elements of H_2 or H_2O. Derived groups and those that cannot be described as above are given an additional screening term, the "element group" of which they are composed. This combination of terms, while not sufficient to reconstruct the structure of the subject compound without some degree of intellectual effort, provides adequate screening power to distinguish different compounds and to search for compounds having the structural characteristics of search questions.

Ring structures are indexed as in the system of Edge *et al.*, and recorded by selecting and/or entering terms as in Figure 4. For specific compounds, only the Ring Index number, degree of unsaturation, and number of units needs to be entered. Generic groups may need an indication of ring type and the presence of heteroatoms.

Configuration terms are entered on the form shown

Figure 2. Rules for fragment indexing of functional groups

Organic compounds are described as derivatives of hydrocarbon moieties altered by unsaturation or substitution. The number of double and triple bonds present is entered. Heteroatom substituent groups are described in terms of six Basic Functional Groups:

Primary Amine (—NH_2)
Secondary Amine (> NH)
Carbonyl
Carboxyl
Hydroxyl
Thiol

These Basic Groups, if present, as well as tertiary amine, aldehyde, halogen, ether, sulfide, nitro, and nitrile are entered as EXISTING. All other hetero groups are entered as DERIVED according to the Basic Functional Groups from which they can be formed by elimination of water or hydrogen without any changes in valence state.

Element Groups are also entered showing for each derived heteroatom grouping the elements present, in alphabetical order, without regard to the number of each element. Element Groups are also entered for any heteroatom group that cannot be described as above.

The following rules apply:

Carbon is included as part of the heteroatom grouping if it is multiply bonded to a heteroatom. Hydrogen attached to such a carbon is also included.

Carbon both doubly and singly bonded to heteroatoms, not including hydrogen, is considered derived from carboxyl.

Except for C, B, N, P, O, S, Si, and the halogens, all elements are coded M in the Element Groups.

Unsaturation:

_____ C=C
_____ C=C

Substituents:

Existing	Derived
1° Amine	_____
2° Amine	_____
Carbonyl	_____
Carboxyl	_____
Hydroxyl	_____
Thiol	_____
Aldehyde	
3° Amine	
Ether	
Halogen	
Nitrile	
Nitro	
Sulfide	

Element Groups

_____
_____

Figure 3. Functionality

Type	Ring Index No.
No Ring	Benzene 292
Carbocyclic	
Heterocyclic	
Fused	
Spiro	

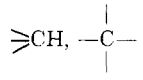
Unsaturation
Maximum
Partial
None

Units

Hetero Atoms	
N in Ring	P in Ring
O in Ring	B in Ring
S in Ring	Si in Ring

Figure 4. Rings

in Figure 5. These are the same as in the parent system, with the addition of an indication of the presence of tertiary and quaternary carbon atoms,



to discriminate branching in hydrocarbon moieties.

Configuration terms describe the relationship between two functional groups (FGs) and the type of carbon atom to which a hetero atom FG is attached. The terms have the following meanings:

0,0 - Used to indicate the presence of two or more hetero atom functional groups joined together by a single bond between carbons that are integral constituents of the FGs.

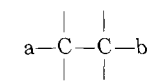
1,1 - Used to denote the configuration

where a and b are any hetero atom FG, alike or different.

0,0
1,1
1,2
1,3
1,4
Aromatic
FG on CH_2
FG on CH
FG on C
Vinyl
α,β
Allyl
Conjugated
CH
C

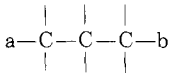
Figure 5. Configuration

1,2 - Used to denote the configuration



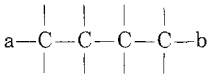
where a and b are hetero atom FGs, alike or different.

1,3 - Used to denote the configuration



where a and b are hetero atom FGs, alike or different.

1,4 - Used to denote the configuration

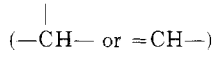


where a and b are any hetero atom FGs, alike or different.

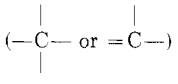
Aromatic - Used to denote that a hetero atom FG is attached to an individual ring having maximum double bond unsaturation.

FG on CH_2 - Used to denote that a hetero atom FG is attached to a secondary carbon atom (CH_2—).

FG on CH - Used to denote that a hetero atom FG is attached to a tertiary carbon atom



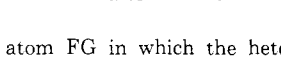
FG on C - Used to denote that the hetero atom FG is attached to a quaternary carbon atom



provided in the latter case

that the carbon atom is not in an individual ring of maximum double bond unsaturation.

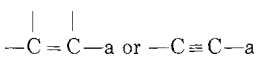
Vinyl - Used to denote the configuration



where a is any hetero atom FG in which the hetero atom

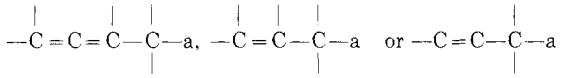
attached to the unsaturated carbon is saturated. The double bond may be in a ring having partial double bond unsaturation but not in a ring having maximum double bond unsaturation.

α,β - Used to denote the configuration



where a is any hetero atom containing FG in which the atom joined to the unsaturated carbon is unsaturated. The double bond may be in a ring having partial double bond unsaturation, but not in a ring having maximum double bond unsaturation.

Allyl - Used to denote the configuration



where a is any FG containing a hetero atom. The double bond may be in either a ring of partial or maximum double bond unsaturation. A double bond in a ring of maximum unsaturation is not recorded as a C=C FG but is recognized as part of Allyl configuration. The carbon atom to which "a" is attached may not be a

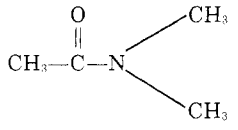
carbon of a carbocyclic ring of maximum double bond unsaturation.

Conjugated - Used to denote conjugated carbon-to-carbon unsaturation which may be olefinic or acetylenic. Double bonds are identified as part of the conjugated unsaturation only if they are coded as FGs, i.e., if they are in an open chain or a ring of partial double bond unsaturation, but not if they are in a ring of maximum double bond unsaturation.

CH - Used to denote the presence in a carbon chain of a tertiary carbon atom.

C - Used to denote the presence in a carbon chain of a quaternary carbon atom.

EXAMPLES OF INDEXING

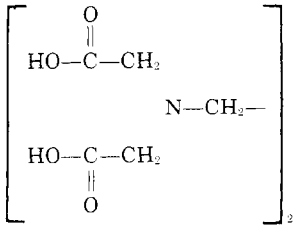


Atom Count

3 C (skeletal)
1 C (nonskeletal)
1 N
1 O

Functionality

1 Derived 2° amine
1 Derived carboxyl
1 CNO element group



Atom Count

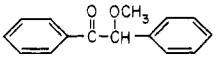
6 C (skeletal)
4 C (nonskeletal)
4 H (nonskeletal)
2 N
8 O

Functionality

4 Existing carboxyl
2 3° Amine

Configuration

1,1
1,2
FG on CH_2



Atom Count

14 C (skeletal)
1 C (nonskeletal)
2 O

Functionality

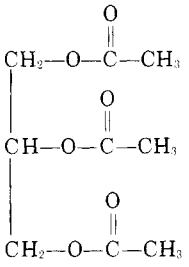
1 Existing carbonyl
1 Ether

Rings

R. I. 292
2 Units

Configuration

1,1
Aromatic
FG on CH
Allyl



Atom Count

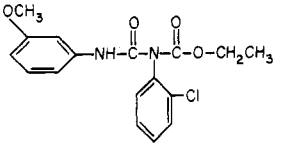
6 C (skeletal)
3 C (nonskeletal
6 O

Functionality

3 Derived carboxyl
3 Derived hydroxyl
3 CO element groups

Configuration

1,2
1,3
FG on CH_2
FG on CH



Atom Count

15 C (skeletal)
2 C (nonskeletal)
1 H (nonskeletal)
2 N
4 O
1 Cl

Functionality

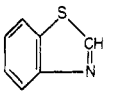
2 Derived 1° amine
2 Derived carboxyl
1 Derived hydroxyl
1 Ether
1 X
1 CHNO element group

Rings

R. I. 292
2 Units

Configuration

1,2
1,3
Aromatic
FG on CH_2



Atom Count

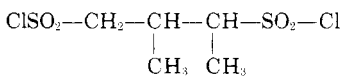
6 C (skeletal)
1 C (nonskeletal)
1 H (nonskeletal)
1 N
1 S

Functionality

1 Derived 1° amine
1 Derived carboxyl
1 Derived thiol
1 CHNS element group

Rings

Max. unsaturation
R. I. 1152
1 Unit



Atom Count

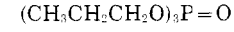
5 C (skeletal)
4 O
2 Cl
2 S

Functionality

2 OSX element groups

Configuration

1,3
FG on CH_2
FG on CH
CH



Atom Count

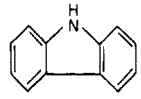
9 C (skeletal)
4 O
1 P

Functionality

3 Derived hydroxyl
1 OP element group

Configuration

FG on CH_2



Atom Count

12 C (skeletal)
1 H (nonskeletal)
1 N

Functionality

1 Existing 2° amine

Rings

Max. unsaturation
R. I. 2927
1 Unit

USE OF THE SYSTEM

Indexing of organic compounds in this way yields an average of eight to 12 terms per compound. These can be stored and manipulated in any of the well-established ways that have been described in recent years. Perhaps the most useful approach is to employ the indexing terms in an inverted file as the second level of a chemical information system.

In such a system individual compounds are posted at the first level, e.g., as registry numbers, to reference sources such as abstracts, journal articles, or patent citations. At the second level, fragment indexing terms are posted to compound names or registry numbers. A small system might use a peek-a-boo file in which each fragmentation term has a card into which is punched the accession number of each compound or generic structure indexed by that term. No more than 200 terms will ever be used in the coding of the great majority of compounds. If peek-a-boo cards that have a 100 × 100 matrix are used, all the screening information for 10,000 compounds can thus be stored and searched in a deck of fewer than 200 cards.

DISCUSSION

This proposed fragmentation system represents a point of view rather than a working procedure. It has not been put into practice, although sufficient test indexing and searching has been done to satisfy us that the approach is effective. There is more redundancy in the indexing as described than may be desirable for maximum efficiency, and individual information systems might find it best to use even fewer terms than are suggested. It may be necessary to add terms for particular fields. The nature of the particular collection of chemical information being manipulated will dictate which terms should be emphasized.

On the other hand, a fragmentation system with only a few simple rules, as is proposed here, has the great advantage that indexing is sure to be consistent and uniformly applied by all users. This would assure compatibility between different collections if it becomes desirable to merge two or more chemical information groups.

This system places emphasis on the more common structural features of organic compounds, which account for the vast majority of known compounds. The few per cent of functional groups that cannot be described by the restricted list of allowable terms are indexed merely by the appropriate "element group," and this information in conjunction with atom count and configurational terms can be expected to provide entirely adequate screening power.

We have found that most compounds can be indexed in one to two minutes by this procedure, once the structural formula has been drawn. Very little mental effort is needed for an indexer with normal training in organic chemistry.

ACKNOWLEDGMENT

The assistance and advice of Patricia A. Dorler in test indexing and in designing the indexing forms is gratefully acknowledged.

LITERATURE CITED

(1) "Survey of Chemical Notation Systems," NAS-NRC Publication **1150**, Washington, 1964.
(2) "Chemical Structure Information Handling—A Review of the Literature, 1962-1968," NAS-NRC Publication **1733**, 1969.
(3) "Survey of European Non-Conventional Chemical Notation Systems," NAS-NRC Publication **1275**, Washington, 1965.
(4) Edge, E. B., N. G. Fisher, and L. A. Bannister, *Am. Doc.* **VIII**, 275 (1957).
(5) Rasmussen, L. E., and J. G. Van Oot, J. CHEM. DOC. **9**, 201 (1969).
(6) "The Ring Index," 2nd ed., ACS, Washington, D. C., 1960; and supplements I, II, & III.