Searching Chemical Structures[*]

By JULIUS FROME[**]
Defense Documentation Center for Scientific and Technical Information,
Cameron Station, Alexandria, Virginia

Received August 1, 1963

CHEMICAL STRUCTURE SEARCHING

This paper is a brief summary of some of the more important approaches used today for chemical structure searching. An attempt is made (1) to generally describe the problem, and (2) to point out the salient advantages and disadvantages of the major areas of chemical information retrieval research.

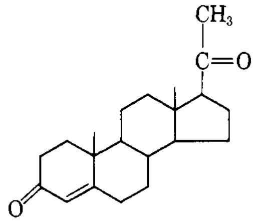
Chemists long have been plagued with the problem of locating chemical information which has been published both in the open literature and in technical reports. Over the years, many attempts have been made to solve this problem. Most of the early work in chemical information control was directed at standardizing chemical nomenclature, but, despite these efforts, chemistry has grown up with diverse and often unrelated systems of naming its products. Some of the nomenclature approaches give clues to the structure being named while others are "trivial" or popular names that are most often unrelated to the underlying chemical configuration. So today, we live with a composite nomenclature system in which a compound like styrene can be named in many different ways, (*e.g.*. styrol, styrolene, cinnamene, cinnamol, phenylethylene, vinylbenzene) and all are correct. Or a simple compound such as acetic acid can be called ethanoic acid, ethylic acid, acetonecarboxylic acid, and several other names.

In an effort to overcome the difficulties of nomenclature and in view of the growth of chemical literature, many chemists have worked on the problem of finding chemical compounds in the literature. The efforts of *Chemical Abstracts* in this field are notable.

The fact that there is no universally employed nomenclature system complicates the present day status of chemical information retrieval. The chemical structural formula remains the only universally accepted means of communication between organic chemists and is therefore the best source of information either for direct coding operations or for "translation" of the structure into a consistent nomenclature.

As we all know, the basic problem today is much broader than finding specifically named compounds. The demand for generic search systems is an ever-growing need of the modern chemical scientist. Many times, he has a need to know all available information about a particular grouping of elements, or a partial configuration, where the remainder of the structure is irrelevant.

The field of chemistry involves literally millions of compounds, some of them complex. It is important that chemists have the ability to find specific compounds or generic concepts within this mass of chemical structures no matter how they are named. Thus, for example, when looking for progesterone, we would also want to find references containing the compound 3,20-diketo-Δ^4-pregnene, since both represent the structure shown below. A searcher looking for information on "unsaturated steroids" would also need to retrieve the same documents.



Progesterone (3,20-diketo-Δ^4-pregnene)

The importance of the chemical structure as a basis for storing and retrieving chemical information takes on an enhanced importance in light of the relationship between structural configuration and biological activity. An enormous amount of research is carried out in this area creating an ever-increasing need for both specific and generic chemical structure searching tools.

In trying to develop solutions for the complex situation of chemical structure searching, three main areas have been studied in at attempt to mechanize the search for chemical information: (1) notation systems, (2) topological systems (element-by-element), and (3) fragmentation schemes.

NOTATION SYSTEMS

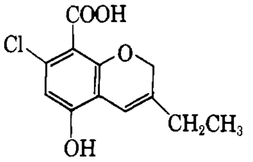
A notation system is an attempt to identify chemical structures uniquely by a consistent and unique linear sequence of symbols. In a sense, then, a notation system is a substitute for nomenclature. The two best known notation schemes are those developed by Dyson[1] and Wisswesser.[2] However, there are a whole host of others such as those published by Silk,[3] Hayward,[4] and Gordon, *et al.*[5]

Modern high speed computers and serial files of notations may prove to be a good method for organizing specific compounds in the chemical literature. The major drawback to broad acceptance of the notation system approach, however, is the complex and unfamiliar system of rules that must be mastered to use these systems and

[*] Presented before the Chemistry Section of the Special Libraries Association. Denver. Colorado. June 12. 1963.

[**] The views expressed herein are those of the author and do not necessarily reflect the views of the U. S. Air Force or the Department of Defense.

to understand the meaning of the linear ciphers. For example, the compound



has the following notations[6]:

Dyson system (International Notation System)-
$B6_2ZQ3C_25CX10Q7Ch9h4$

Wiswesser system-2T66BOCHJD2GQIGJVQ

Semiautomatic file creation and search techniques may help to overcome this handicap by allowing use of the notation for file creation and organization without requiring the searcher to understand fully the complexities of the notation rules. Research continues in this direction, and it was recently announced by *Chemical Abstracts* that its work with the International Notation System had developed to the point at which they could pass from the chemical structure into the computer without the use of the cipher.

Although notation systems can identify and retrieve specific compounds, in some instances, generic searches may be awkward. Most of the systems are now making an attempt to include degrees of genericity in their file organizations. One method involves creating permuted search files where the entire notation is in each case entered under every indidvidual portion of the notation. This requires a large amount of storage per compound, but developments in modern computer technology may help to minimize the practical effects of increased storage and allow widespread use of this approach for generic retrieval.

TOPOLOGICAL SYSTEMS (Element-by-Element)

Since there are only about a hundred elements used as basic building blocks for compounds, many researchers have felt that it would be desirable to code and search chemical structure in an element-by-element approach, or topologically, the structures being described by the sequential connection of the various elements. This seemed impossible by the sheer magnitude of the file creation and storage problem before the general availability of high speed search hardware, but recent years have seen a growth in the number of researchers developing this approach. Some of them, such as Waldo,[7] Ray,[8] Ballard and Neeland,[9] and Norton and Opler,[10] have made considerable progress in element-by-element search systems. Theoretically, they all appear to be workable, but none has proven to be practical enough to gain widespread support. One of the great difficulties is the tremendous volume of coding associated with topological systems. Although intellectually it may be a simple task to visualize the dissection of a given structure, the application of a logical element-by-element coding program is invariably a long, involved, and tedious process. This approach does have the advantage that rules for coding are simple, but the amount of coding and the number of codes per structure runs high. Furthermore, the search program is complicated, costly, and takes much computer time and storage. Most researchers in this field believe that any successful computer element-by-element search would search by a chemical scheme using preconstructed combinations of elements rather than merely by the individual elements.

In an attempt to alleviate the topological coding problem, Jacobus and his associates[11] have developed a chemical typewriter that has the ability to type chemical structural formulas. A punched paper tape is a by-product which may be used as computer input eliminating the need for human coding and therefore substituting a typist for the coder. However, to date no practical search program for economical use of the equipment has been developed, but good work is going on in this area.

FRAGMENTATION

Some of the most promising methods for chemical coding, storage, and searching seem to be the systems based upon structure fragmentation. Chemists think of chemical structures as combinations of fragments. These are the usual building blocks known to chemists, *e.g.*, nitro, amino, carboxy, sulfonic, mercapto, etc. The fragments are well known to react generally as a unit, and their properties and reactions constitute a major portion of the chemist's basic training. This factor was recognized early by Frear[12] in his work on fragmentation. In this particular instance, an index was created assigning each fragment a code, and compounds were cataloged by a code number composed of combinations of the fragment codes. This was a forerunner of the Chemical-Biological Coordination Center system[13] in which fragmentation was carried to a high degree of sophistication. The latter system was mechanized to retrieve CBCC codes for over 50,000 compounds. One of the difficulties in this approach is the fact that no relationships between fragments are coded producing a large amount of noise, or false drops, in searching in the system. Recognizing this problem, the U. S. Patent Office carried out research in fragmentation schemes which stressed fragment relationships. The first of these systems was developed in the area of steroids[14] where the fragments attached to steroid nuclei are coded, along with the steroidal location involved. This approach is effective and is used by most major pharmaceutical houses and many Patent Offices throughout the world. Following the steroid approach, the Patent Office created a system for organophosphorus compounds[15] where the relationships between fragments are coded on a series of matrices. Both of these systems proved to be operationally practical in the U. S. Patent Office and were successful with document files of up to 10,000 entries.

In recent years, the Patent Office extended these principles to be applicable to all organic chemistry. The PACIR[16] system is a logical extension of the earlier work and has rules for fragmenting organic compounds along with roles and interfixes for highly selective retrieval. A semiautomatic coding scheme is employed allowing more efficient processing of input and a variable user-oriented search descriptor list. Many of the PACIR concepts are also employed in a related search system with a code sheet indexing approach using a coding format with a limited number of descriptors and "open-ended" capabilities for new fragments.[17]

CONCLUSIONS

Each of these systems has its particular advantages and it is likely that all of the approaches will receive sustained emphasis and use because of the variety of needs in the chemical information retrieval field. One thing is certain—advances in computer technology will definitely play a major role in determining the relative importance and general acceptability of the three directions.

BIBLIOGRAPHY

(1) G. M. Dyson, "A New Notation and Enumeration System for Organic Compounds," Longmans, Green and Co., London and New York, 1947, 63 pp.

(2) W. J. Wisswesser, "A Line-Formula Chemical Notation," Thomas Y. Crowell Co., New York, N. Y., 1954, pp. 125–126.

(3) J. A. Silk, *J. Chem. Doc.*, **3**, 189 (1963).

(4) H. W. Hayward, "A New Sequential Enumeration and Line Formula Notation System for Organic Compounds," Office of Research and Development, Patent Office, U. S. Department of Commerce, Washington, D. C., November, 1961 (Patent Office Research and Development Reports No. 21).

(5) M. Gordon, C. E. Kendall, and W. H. T. Davison, *Proc. Intern. Congr. Pure Appl. Chem., 11th London, 1947.* **II,** 115 (1950).

(6) H. T. Bonnett, *J. Chem. Doc.*, **3**, 235 (1963).

(7) W. H. Waldo and M. DeBacker, *Proc. Intern. Conf. Sci. Inform., Washington, 1958*,**1**, 711 (1958).

(8) L. C. Ray and R. A. Kirsch, *Science*, **126**, 3278 (1957).

(9) D. L. Ballard and F. Neeland, *J. Chem. Doc.*, **3**, 196 (1963).

(10) A. Opler and T. R. Norton, "A Manual for Programming Computers for Use with a Mechanized System for Searching Organic Compounds," Research Dept., Western Division, Dow Chemical Co., Pittsburg, Calif., April 25, 1956.

(11) A. Feldman, D. B. Holland, and D. P. Jacobus, *J. Chem. Doc.*, **3**, 187 (1963).

(12) D. E. H. Frear, "A Catalogue of Insecticides and Fungicides," Chronica Botanica Co., Waltham, Mass., Vol. 1, 1947; Vol. 2, 1948.

(13) Chemical-Biological Coordination Center, "A Method of Coding Chemicals for Correlation and Classification," National Research Council, Washington, D. C., 1950.

(14) "Revised Steroid Search System Coding Manual," Patent Office Research and Development Report No. 19, U. S. Department of Commerce, Washington, D. C.

(15) J. Frome, P. T. O'Day, F. S. Sikora, and M. S. Gannon, "Manual for a Punched Card Retrieval System for Organic Phosphorus Compounds," Patent Office Research and Development Report No. 22, U. S. Department of Commerce, Washington, D. C.

(16) J. Frome and P. T. O'Day, *J. Chem. Doc.*, **2**, 249 (1962).

(17) J. Frome and P. T. O'Day, *ibid.*, **4**, 33 (1964).

The Singularity Sub-Link—A New Tool for Use in the Storage and Retrieval of Information*

By J. FREDERIC WALKER
Research Division, Electrochemicals Department, E. I. du Pont de Nemours and Co.,
Wilmington 98, Delaware

Received August 21, 1963

The singularity sub-link gives additional scope and utility to concept-coordination systems for the storage and retrieval of information. It simplifies the indexing process, reduces false retrieval, and makes it possible to locate documents in which various materials, chemicals, special agents, processes, and independent variables are compared. It does this by indicating that an indexing term or key-word is a member of a series of alternates that have been studied separately for comparison in a given context. Previous methods of doing this have required an excessive number of links since for accuracy each alternative term must be placed in a separate link in which all the nonalternative terms are repeated. Furthermore, if this proliferation of links is avoided by placing the alternative terms in one link, false retrieval of information relative to combination effects resulting from mixtures not described in the documents is an inevitable sequel. The term "COMPARISON" has been previously avoided in our concept-coordination systems of indexing because of its indefinite nature as a search term. The singularity sub-link indicates specific comparisons so that this concept becomes available by implication without ambiguity. Every reference marked by the sub-link indicates a document reporting the comparative evaluation or study of a term with other terms employed in the same context.

The functions and relations of terms are lost when they are isolated from each other in a simple inverted file. Links coordinate the terms in the respective intellectual subdivisions of the report or document. Roles preserve relationships by indicating functions, such as raw material,

* Presented before the Division of Chemical Literature, 145th ACS National Meeting, New York, N. Y., September 11, 1963.