Chemical Structure Codes in Perspective*

M. L. HUBER**
Eastern Laboratory, Explosives Department,
E. I. du Pont de Nemours and Co., Gibbstown, New Jersey
Received April 6, 1964

To chemists, graphic structures are descriptions of reality which not only identify chemical compounds but also convey relationships concerning functions, reactivity, and properties. Chemical formulas and structures also are essential methods of entry through which information is located or correlated in the body of chemical knowledge, but practical methods have not been available for indexing structures, in contrast to indexing representations of them which are further abstractions of chemical reality. Furthermore, bulky structures cannot be printed conveniently with conventional linear type and they have no inherent equivalent for speech. Therefore, chemists and documentalists have devised a variety of descriptive codes or representations to record and communicate chemical structures and the compounds they portray. These methods include nomenclature, structure hierarchy, fragmentation, cipher notation, and topological coding.

CODING, IDENTIFICATION, AND CLASSIFICATION

For effective storage and retrieval, chemists need to identify and index compounds, so that subsequent location is possible, and to classify and organize compounds so that related or generic structures can be grouped for consideration of common attributes. By analogy, these needs are related to the basic recording operations succinctly described by Fairthorne[1] as "parking" and "marking," both of which are used in retrieval systems. Historically, identification of structures has been stressed in connection with the library requirements of indexing and summarizing; in contrast, generic classification has been emphasized along with laboratory experimentation and data correlation.

While a structure code must meet several needs, the two dominant requirements are the ability to provide a unique, specific address for each compound, and the ability to group together compounds with similar structural characteristics.[2] These ability requirements reflect different descriptive levels on the abstraction ladder emanating from chemical reality. In practice, therefore, these functions of a code are often exclusive[3] and the description which performs well for ordering is frequently inadequate for generic classification, and *vice versa*. Nomenclature is a particular example; names are suitable for indexing but are poor for generic organization.

A comparison of coding methods points up the conflicting capabilities and varying success of the available methods in performing the two essential requirements of unique identification and generic classification. Deficiencies in the traditional methods, combined with the need for designations more adaptable to machine methods for handling large numbers of compounds, have stimulated the development of alternate techniques for coding structures. In this context it is both pertinent and significant that major developments and uses for structure coding have evolved within pharmaceutical and medical research organizations where the comparative evaluation of large numbers of complex compounds is a major requirement and where the research strategy is guided *via* relationships of structure and function with properties and activity.

METHODS FOR IDENTIFICATION

The capability for generic organization must be met without abandoning the ability to locate specific compounds in a large collection. Accordingly, a unique identification or address is the most critical requirement. This requirement is fulfilled by three methods of specific identification: a systematic name ("word name"), a line notation ("cipher name"), or an identification number ("number name"); all of the structure codes are associated in some way with one or more of these interchangeable "names," which are illustrated schematically in Figure 1.

Basically, these "names" are descriptions of the two-dimensional structure which symbolizes chemical reality. In the "Ring Index," for example, the identity number and the systematic name (and the cipher, if desired) are alternate identifications of specific structures. If the structure is unknown or indefinite, an identification by name or number can be assigned directly (shown by dotted lines in Figure 1), but systematic nomenclature and cipher notations are primarily abstractions of structure according defined rules. In principle, each of these specific designations can be unique but the goal of an unambiguous identification for each compound is not always attained.

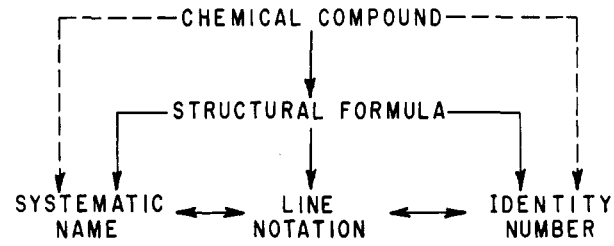

Figure 1.

* Presented before the Division of Chemical Literature, 147th National Meeting of the American Chemical Society, Philadelphia, Pa., April 6, 1964.

** Secretary's Department, E. I. du Pont de Nemours and Co., Wilmington, Del.

In their representation of chemical compounds and structure, these identifications differ in their characteristics and in the information they contain. Thus, verbal names can be communicated by speech as well as print, and the word description can be derived from or transposed into structure by those who know the formal procedures. Ciphers do not have a convenient speech equivalent, but for those who have learned the special symbols and rules, the structure and its linear code which has internal meaning are interchangeable. Identity numbers usually are meaningless in themselves and require auxiliary codes or references (analogous to symbols and rules for names and ciphers) to convey structure, but numbers are manipulated easily by machines and, in the judgement of many people, serve as the least ambiguous common denominator in human communication.

Both the structure of a compound and the primary identifications are associated with auxiliary descriptors in practice. For example, the molecular formula is used as a general representation for a group of isomeric structures; trivial names, trade names, and cryptograms supplement the formal nomenclature; the line-formula shorthand is a linear description analogous to the cipher; and secondary descriptors such as local numbers, names, or the colloquial references used in everyday communication are associated with primary identity numbers.

In performing their essential ordering and indexing function, the three identifications are utilized in different ways. That is, names are ordered by alphabet or by class relationships to place them in a specific location, and auxiliary association is provided with cross references; ciphers are listed or permuted by the cipher symbols or letters, and names or numbers serve as auxiliary references; identity numbers are arranged by numerical sequence, and names, ciphers, or topological structure codes provide corollary designations. Some of these relationships are illustrated further in Figure 2.

METHODS FOR GENERIC CLASSIFICATION

The traditional basis for generic or substructure classification has been a structure hierarchy, expressed in terms of either a formal nomenclature or a parent-class order; alternatively, by defining fragments of structure, a more flexible arrangement can be obtained. This approach through structure dissection is refined further in the cipher notation which not only shows segments but encodes the complete structure. Further discrimination can be developed *via* a topological code, which records the complete structure through its atom-bond combinations. Pertinent points for comparing the various coding practices are summarized under these categories.

Nomenclature.—The nomenclature of any science, as Dyson has emphasized,[4] must consider not only a convenient general language for every day use (simple, trivial, or trade names) but also a legal language where terms are strictly defined (unique, exact, formal names). These latter, systematic names are convenient tools for identification and indexing, but they provide a poor basis for meaningful classification or functional organization.[5] Although formal nomenclature, which is a hierarchy of defined units of functional segments, has some capability for generic organization through its syllables and roots, this utility is limited by the scattering of alphabetical listings.[6] Nevertheless, a partial classification by nomenclature is widely used in abstract journals and in information systems; to combine this feature with indexing,

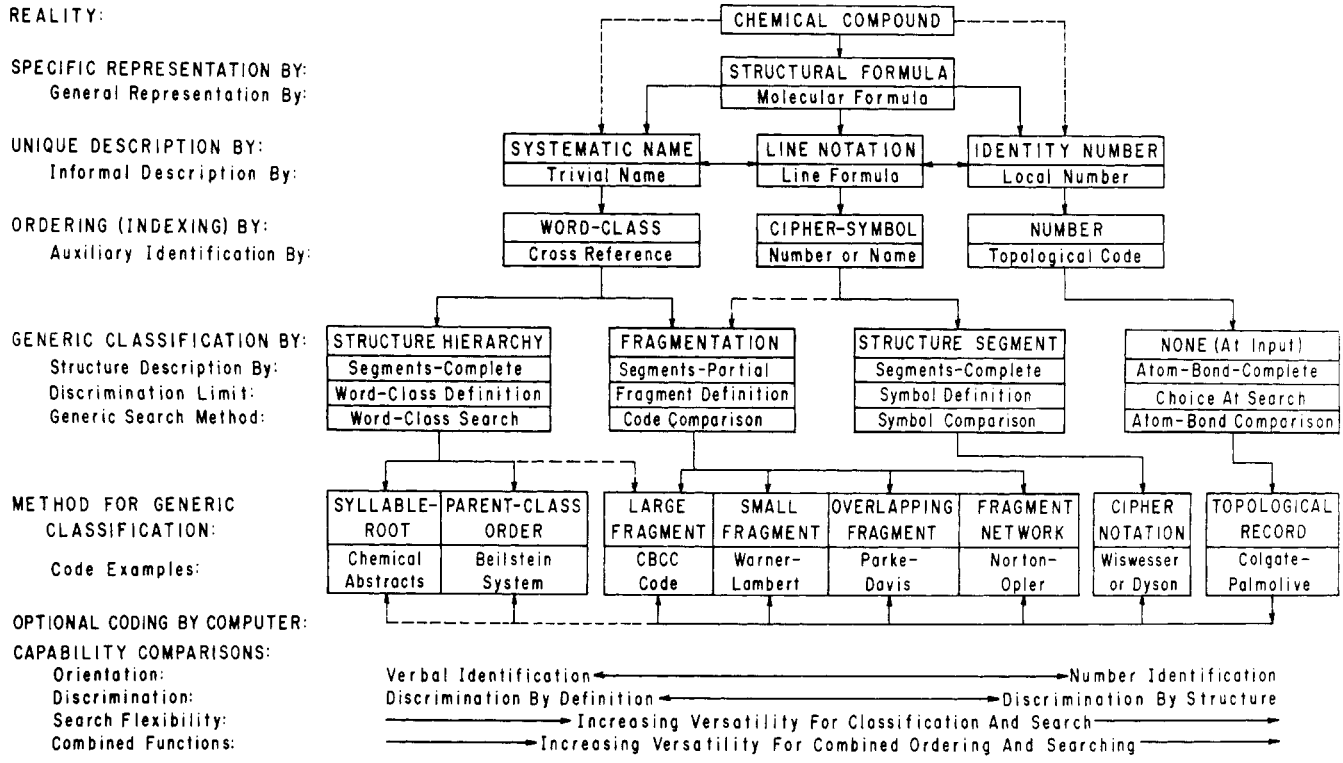

Figure 2.—Description and coding of chemical compounds.

supplementary generic headings and generous "see also" cross references are added.

Throughout the past 70 years, international groups have made outstanding contributions and improvements in developing a formal nomenclature. However, the goal of defining single, systematic names is becoming increasingly difficult because many of the formal names and nomenclature rules are too complicated for general use, and systematic names cannot be developed for many complex structures. In addition, the confusion of a variety of names and alternate designations is a major problem for chemists and information specialists, as noted by several authors.[6-8] Because of these deficiencies, other techniques have been developed to supplement nomenclature descriptions. Also, in recent years, proposals have been made for the use of structural formulas and molecular formulas (rather than words or names) to index information about compounds[6]; such techniques are being evaluated currently in journals and in proprietary information collections.

Parent-Class Hierarchy.—Beilstein's "Handbuch" exemplifies the classical hierarchy arrangement of structure wherein each compound is located in one address according to arbitrary rules relating to structural features. The system includes main divisions based on rings, classes based on functional groups, subclasses containing homologous series, and rubrics or parent compounds to which derivatives are related. Historically the Handbook has been a very valuable reference and in some information collections this classification method is still used as a compromise technique for locating compounds and for associating groups with some common attributes. However, this complex, inflexible system is frequently illogical and inadequate for generic combinations[2] and to a large extent the hierarchy of systematic nomenclature has superceded the Beilstein method for indexing, and fragmentation codes have been developed for more discriminating generic associations.

Fragmentation.—By this method, a compound is represented as a composite of its predominant structural features. Following the pioneering applications of the Wiselogle Code,[29] the Frear Code,[10] and the CBCC Code,[11] more than thirty modifications of fragment coding have been developed in the past two decades. This evolution has occurred in response to a need for information association and retrieval on the basis of generic traits, and the various codes emphasize substructure classification primarily. Typically, functional groups or defined segments and their interrelationships are assigned numerical descriptors which can be recorded on punched cards or magnetic tapes for subsequent search to retrieve structures having specified generic characteristics.

With reference to structure, one of the significant differences in the fragment codes is in the general size of the defined substructures. For example, some codes utilize relatively large segments,[11] others list small combinations,[9] several permit adjacent atoms to be included in more than one fragment,[12] and a few interrelate structural parts by a numerical network.[13] Each individual code includes refinements related to proprietary needs or to limitations of machine equipment, but all of the fragment codes function by representing compounds through a collection of pieces, as in a jigsaw puzzle.

Fragment codes have proved to be adequate and very practical for many classification and correlation purposes. However, such descriptions are only a partial representation of structure, based on an aggregate of predetermined, fixed segments. Thus, related compounds may have the same fragments and descriptors, and for some structures there are no suitable segments to use. With the discrimination limited to the arbitrary structural units and their definitions, fragmentation coding is quite satisfactory for many compounds for which the composite approaches a complete description but is deficient in delineating structures which are closely related or cannot be divided into suitable segments.

Cipher Notation.—Cipher notation employs letters and symbols in linear sequence to describe the structure. Generally, the characteristic segments and atom relationships in the molecule are represented by assigned letters, numbers, and symbols which can be printed and manipulated by machine methods. Through careful and intricate symbol definitions, a unique code is assembled to depict the complete structure. In one sense the cipher is a "name," the use of which is feasible in writing but not in concise speech; in another sense it is a convertible code with internal meaning from which the structure can be regenerated.

Substructure information in notations is carried in the cipher symbols and their combinations; generic retrieval is accomplished through symbol comparison in a manner analogous to searching *via* fragment descriptors. Because of the precision of the cipher codes, the selectivity in searching can be more definitive than in fragment matching, but the discrimination is limited to the symbols and definitions used at input. The dependence of this method on hierarchial rules and on an intricate symbol language analogous to the systematic nomenclature is a handicap; however, this deficiency is neutralized to some extent by the possibility of generating the cipher *via* a computer program and a topological code.[8]

The predominant cipher codes (Dyson,[14] Wiswesser,[15] Silk,[16] and Hayward[17]) differ primarily in the prominence and precedence given to particular structural features, but the systems do combine the ability to identify structures for indexing and to delineate structure components for classifications.[3] Unfortunately, the controversy and argument among the proponents of these methods frequently obscure the desirable elements and strengths of each. Superimposed on the dispute, however, are the basic questions of whether a cipher is needed for a particular application and whether the need can be fulfilled practically and economically in an alternate way.

Topological Coding.—A topological code is a structure description showing the atom-to-atom connections in a compound with the atoms as nodes and the connections as branches of a network. In a sense, the code is a mathematical snapshot of the complete structure, and the corresponding matrix representation is a numerical analog and a structural model which can be handled *via* computer programs. The theories and techniques of network analysis which have evolved in recent years provide the basis for several proposals and several experiments using matrices to encode structures.[8, 18-22, 25]

In the few practical systems which have been developed, generic information is stored within the topological

structure code and retrieval is determined by the substructure selected for the search. In contrast to other methods, the degree of discrimination is optional and is determined by the choice at the time of search and not by a prior structure dissection at the time of indexing. Thus, in concept, the topological description approaches a maximum in coding technique: a unique, complete description of structure and a flexible, unlimited substructure classification. That is, the code can provide a versatile method both for identification and for generic classification.

The unique, matrix-derived code of a compound is a type of numerical identification, but in practice it is more convenient to assign a corresponding identity number or registry number. This numerical "name" may be supplemented with auxiliary names or descriptors but the primary functions of identification and structure description are performed by the identity number and the associated topological code.

STRUCTURE CODES IN PERSPECTIVE

Figure 2 is a graphical summary of the various structure coding methods and their interrelationships, both for indexing and for classification. The vertical steps from the top illustrate abstractions from chemical reality, with verbal orientation emphasized at the left and numerical orientation at the right. As noted, the three dominant identifications are interchangeable, and one or more of them is used along with each of the generic classification methods.

The vertical sequence on the left side depicts the traditional technique of verbal identification combined with generic description through defined hierarchies by name form or by structure class; for retrieval, verbal terms or ordered classes are consulted. Where the primary need is to describe and retrieve specific compounds and where broad generic definitions are sufficient, this method is adequate and is widely used.

Where better generic discrimination is desired, fragmentation is employed, frequently in combination with nomenclature; for retrieval, structure fragments and their interrelationships are compared. This technique has been a favorite for organizing structures *via* functional segments or defined substructures.[10,23]

The central part of the figure illustrates the linear notation in which the complete structure is coded by ciphers; for retrieval, cipher symbols are selected and matched. Particularly where the file includes numerous compounds with closely related structures, ciphers are used as a convenient method of organization both for identification and for collecting compounds with specific features. Also, the cipher symbols provide a basis for file subdivisions or for preliminary screening of a large collection.[3,24]

The sequence on the right side diagrams the topological method which is inclined toward numerical identification, with an associated record of the complete structure by a matrix code; for retrieval, a flexible selection is possible, based on the atom combination chosen at the time of search. Where there is a large collection of compounds requiring the capacity and speed of computers for centralized operation, and where there is a need for optional searching from multiple points of view, the topological technique promises to be the method of choice.

In summary, Figure 2 illustrates the spectrum of structure coding methods and their evolution from a generic discrimination by definition toward a discrimination by structure and optional choice. The techniques toward the right display an increasing versatility for generic classification and search, and an increasing versatility in performing the combined functions of indexing and of classification.

It is apparent that each of the techniques has strengths which enhance its use, and the choice of a preferential method is dependent on the needs and the economics in specific situations. This conclusion is reflected strongly also in a recent comprehensive survey of chemical notation systems and their use[26] and in another review.[27]

IMPLICATIONS FOR THE FUTURE

If the recent evolution emphasizing topological coding continues, several implications and trends may be forecast.

(1) The development and acceptance of suitable topological codes for compounds should permit the adoption and use of a single registry (perhaps national or international), with identity numbers for primary ordering and indexing, and with the associated codes for structure searching from many points of view.

(2) With the inherent potential to generate other necessary generic descriptions from the topological matrix *via* computer search programs, cipher codes and fragment codes (and probably names, eventually) can be provided, as the need arises, by machine programs rather than by human analysis at input.

(3) As the use of identity numbers (and associated codes) supercedes the use of names for primary reference, the urgency of selecting unique, systematic names for structures will diminish since many of the needs for verbal names will become secondary. As a corollary, the traditional concern and emphasis on the selection of formal names will tend to shift instead to a different level of abstraction of chemical reality, *i.e.*, to a selection of the singular, unique structure which best describes each compound.

REFERENCES

(1) R. A. Fairthorne, "Towards Information Retrieval," Butterworths, London, 1961, pp. 85, 95.
(2) E. L. Buhle, F. Y. Wiselogle, *et al.*, *J. Chem. Educ.*, **23**, 375 (1946).
(3) H. T. Bonnett, *J. Chem. Doc.*, **3**, 235 (1963).
(4) G. M. Dyson, "Chemical Nomenclature," Advances in Chemistry Series, No. 8, 1953, p. 104.
(5) J. W. Perry, *ibid.*, p. 107.
(6) B. Loev, *J. Chem. Doc.*, **1** (2), 27 (1961).
(7) L. Schmerling, *ibid.*, **1** (1), 46 (1961).
(8) G. M. Dyson, *et al.*, *Inform. Storage Retrieval.*, **1**, 69 (1963).
(9) F. H. Arendell, *J. Chem. Doc.*, **1** (3), 47 (1961).
(10) D. E. H. Frear, "Punched Cards—Their Applications to Science and Industry," 2nd Ed., R. S. Casey, *et al.*, Ed., Reinhold Publishing Corp., New York, N. Y., 1958, Chapter 22.

(11) Anon., "A Method of Coding Chemicals for Correlation and Classification," National Academy of Sciences–National Research Council, Chemical-Biological Coordination Center, Washington, D. C.
(12) H. A. Geer, *et al.*, *J. Chem. Doc.*, **2**, 110 (1962).
(13) A. Opler and T. R. Norton, *Chem. Eng. News*, **34**, 2812 (1956).
(14) "Rules for I.U.P.A.C. Notation for Organic Chemistry," Longmans, Green and Co., London, 1961.
(15) W. J. Wiswesser, "A Line-Formula Chemical Notation," Thomas Y. Crowell Co., New York, N. Y., 1952, revised 1962.
(16) J. A. Silk, *J. Chem. Doc.*, **3**, 189 (1963).
(17) H. W. Hayward, U. S. Patent Office Research and Development Report, No. 21, 1961.
(18) E. Meyer and K. Wenke, *Nachr. Document*, **13**, 13 (1962).
(19) L. Spialter, *J. Am. Chem. Soc.*, **85**, 2012 (1963).
(20) D. Gould, *et al.*, *J. Chem. Doc.*, **5**, 24 (1965); see also *Chem. Eng. News*, **41**, No. 46, 5 (1963).
(21) D. J. Gluck, *J. Chem. Doc.*, **5**, 43 (1965); see also *Chem. Eng. News*, **41**, No. 49, 35 (1963).
(22) D. L. Ballard and F. Neeland, *J. Chem. Doc.*, **3**, 196 (1963).
(23) H. A. Geer and C. C. Howard, *ibid.*, **2**, 51 (1962).
(24) H. T. Bonnett and D. W. Calhoun, *ibid.*, **2**, 2 (1962).
(25) E. H. Sussenguth Jr., *ibid.*, **5**, 36 (1965).
(26) "Survey of Chemical Notation Systems," Publication 1150, National Academy of Sciences–National Research Council, Washington, D. C., 1964.
(27) J. Frome, *J. Chem. Doc.*, **4**, 43 (1964).

Syntactic Scanning of Chemical Information

PAUL L. FEHDER* and M. P. BARNETT**
Cooperative Computing Laboratory, Massachusetts Institute of Technology, Cambridge 39, Massachusetts
Received October 1, 1964

1. INTRODUCTION

Digital computers were designed originally to perform numerical computations, but they have been used increasingly during the last few years to deal also with nonnumerical processes in many fields of study. Nonnumerical work now accounts for a substantial portion of computer usage in many organizations, and it is likely to increase considerably in the future, with an attendant trend toward the design of computers to facilitate such work.

Several of the techniques of nonnumeric information processing are being used in various laboratories to deal with chemical problems. This present paper describes a particular technique, namely mechanized syntactic analysis, by reference to a chemical topic of an illustrative nature. This is done because the authors believe that future developments may make the technique of practical value in nonnumerical chemical information processing. It should be stressed, however, that the specific example of mechanized syntactic analysis which is described in this paper was chosen for explanatory purposes. It is not suggested as an application of immediate practical value.

In discussions of mechanized information processing, the term "syntax" is applied to linear notational systems, systems of nomenclature, stylized subsets of natural language, and other forms of alphabetized information, with a meaning that is covered by a slight extension of the following dictionary definition of the general usage of the term

> "The arrangement of words by which their connexion and relation in a sentence is shown"[1]

This definition may be modified, for use in discussions of information processing, to

> "The arrangement of substrings by which their connexion and relation in the string which they form is shown"

The term "string" is used for a sequence of characters, taken from a finite set of characters, such as a conventional alphabet, or the set of characters that can be typed on a given keyboard machine. The term "substring" is used for any string that forms part of a longer string. Notations that use subscripts and superscripts can be mapped into strings by the use of several simple conventions, and more elaborate two-dimensional notations can be linearized in rather less convenient ways.

Syntax description and syntactic analysis have been used extensively in connection with nonnumeric applications of computers. Early work of this type has been described.[2-5] Several methods of representing syntaxes are now used in mechanized information processing.[6] We use a method[4] that allows "generic names" to be given to the types of string to which frequent reference is made in an application that is under consideration, and allows the name for each of these types of string to be defined by reference to the generic names of the constituent substrings and to the ways in which these substrings can be arranged and varied. A syntax is defined, in this approach, by a set of sentences of a certain simple form that is described in section 2. The set of sentences is called a verbal definition table. The words that form the sentences are abbreviated, in accordance with certain conventions, for input to our computer programs. These

* Chemistry Department, California Institute of Technology, Pasadena, Calif.
** University of London, Institute of Computer Science, 44 Gordon Square, London, WC1, England.