(3) H. E. Dayringer, G. M. Pesyna, R. Venkataroghavan, and F. W. McLafferty, "Computer-Aided Interpretation of Mass Spectra", *Org. Mass Spectrom.*, **11**, 529 (1976).

(4) A. M. Duffield, A. V. Robertson, C. Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg, "Application of Artificial Intelligence for Chemical Inference. II. Interpretation of Low-Resolution Mass Spectra of Ketons", *J. Am. Chem. Soc.*, **91**, 2977 (1969).

(5) B. G. Buchanan, D. H. Smith, W. C. White, R. J. Gritter, E. A. Feigenbaum, J. Lederberg, and C. Djerassi, "Application of Artificial Intelligence for Chemical Inference. 22. Automatic Rule Formation in Mass Spectrometry by Means of the Meta-DENDRAL Program", *J. Am. Chem. Soc.*, **98**, 6168 (1976).

(6) S. Sasaki, H. Abe, Y. Hirota, Y. Ishida, Y. Kudo, S. Ochiai, and T. Yamasaki, "CHEMICS-F: A Computer Program System for Structure Elucidation of Organic Compounds", *J. Chem. Inf. Comput. Sci.*, **18**, 211 (1978).

(7) Frank Harary, "Graph Theory", Addison-Wesley, Reading, MA, 1969.

(8) Takashi Nakayama and Yuzuru Fujiwara, "BCT Representation of Chemical Structures", *J. Chem. Inf. Comput. Sci.*, **20**, 23 (1980).

(9) The method of structure generation based on BCT representation of chemical structures will be dealt in our next paper.

Realistic vs. Systematic Nomenclature

JOHN A. SILK*

Imperial Chemical Industries Ltd., Plant Protection Division, Jealott's Hill Research Station, Bracknell, Berkshire, United Kingdom

The place of systematic nomenclature is appraised by relating its functions to recent developments.

The recent papers by Goodson and others on graph-based chemical nomenclature[1,2] raise afresh in my mind the difficult question of the value of truly systematic nomenclature. We already have systems capable of meeting the great majority of needs, and the fact that some of them are specially designed to meet particular requirements reflects the manifold variety of molecular architecture.

This seems to be an opportune time to reappraise the place of systematic nomenclature by relating the functions it is required to serve to recent developments. For what applications would a completely systematic and comprehensive nomenclature system be useful?

(1) Among chemists the primary means of communication is the structural formula, and the role of nomenclature is secondary in providing linear descriptions of structures in forms which can be both written and spoken. For communication *descriptiveness at an appropriate cognitive level* is required. Current systems of nomenclature reflect this need by employing a suitably rich vocabulary with characteristic names for important ring systems and functional groups. The variety speeds communication, at least among the knowledgeable, by enabling larger entities to be designated by a few syllables. This psychological aspect is important: the names are fit for their purpose.

With a fully systematic nomenclature, by contrast, the vocabulary is deliberately limited to a basic set of terms. While in principle this simplifies the construction and interpretation of names, in practice it has the obvious disadvantage of a higher degree of fragmentation (which lengthens names and slows comprehension) and the less-recognized disadvantages that complex sets of numerals, punctuation, parentheses, and other symbols are required to specify syntactical relations among components and that the names have a greater overall sameness and lack of distinctiveness. Consequently, they become more liable to errors in transcription (direct copying) and translation (to a structural diagram).

Whatever their logical attractions, such names are not generally of a type which chemists would willingly use in preference to established styles. The situation may be likened to that with a synthetic language, such as Esperanto, which has found few supporters as an international means of communication in comparison with two or three living languages. A new nomenclature system is likely to find applications only in areas where it can deal with new situations in a useful manner, for example, cyclophanes, where nodal nomenclature is clearly relevant.

(2) The major use of systematic names is in documenting the literature of chemistry in abstract journals and reference works. Chemical Abstracts Service has taken the lead in this important role by rationalizing current practices and providing valuable accessories, particularly the *Parent Compound Handbook* and the *CA Index Guides*. Despite all this, the role of systematic names is still mainly secondary: the molecular formula index is the primary tool for locating compounds, and the names in it serve merely to distinguish isomers of the same molform.

For this purpose it is not strictly necessary to derive a *unique* name for each compound; an *unambiguous* name suffices. This then accords with the situation of a searcher who is using formula indexes without having expert knowledge of nomenclature systems. He is able to interpret the alternatives which may be presented in a way which is heuristic rather than algorithmic.

Information scientists who have tried to use names for generic search, particularly in online mode, have come to realize the inherent limitations of even *9 CA* names as a basis for uniformly predictable descriptions of molecular structure. These arise from three general rules, which are designed to lead to unique names. They are, firstly, the priority rankings among functional groups, secondly, the precedence which is always given to the longest carbon chain, and thirdly, the alphabetical sequencing of substituents. The special methods used for naming symmetrical structures also cause substantial variations in styles of names (see Figure 1). Consequently, relatively small changes in structure can often lead to major changes in the forms of names for related compounds.

While CAS can be its own arbiter, the IUPAC rules for organic nomenclature[3] illustrate another facet of the problem. For many classes of structure two or even three alternative styles of name are permitted. Moreover, it has been my experience that the guidance provided is inadequate for many compounds encountered in practice and that even experts in IUPAC nomenclature differ over details of the name for a

* Address correspondence to 5 Albert Road, Wokingham, Berks RG11 2AL, United Kingdom

0095-2338/81/1621-0146$01.25/0

$C_2H_5OCOC(C_2H_5)=CH-$	2-ethoxycarbonyl-1-butenyl
$C_3H_7OCOC(C_2H_5)=CH-$	2-propoxycarbonyl-1-butenyl
$C_2H_5OCOCH=CH-$	3-ethoxy-3-oxo-1-propenyl
$C_2H_5OCOCF=CH-$	3-ethoxy-2-fluoro-3-oxo-1-propenyl
$C_3H_7OCOCH=CH-$	3-oxo-3-propoxy-1-propenyl

4(3*H*)-quinazolinone, 3-(methylamino)-
acetamide, *N*-methyl-*N*-(4-oxo-3(4*H*)-quinazolinyl-

Figure 1. Variations in styles of *CA* names for closely related structures.

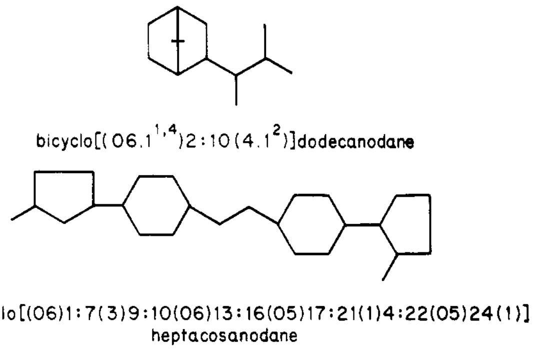
bicyclo[(06.1^1,4)2:10(4.1^2)]dodecanodane
tetracyclo[(06)1:7(3)9:10(06)13:16(05)17:21(1)4:22(05)24(1)]
heptacosanodane

Figure 2. Nodal nomenclature.

given structure. Is this not a further indication of the true magnitude of the problems and at least a hint that the effort may be misplaced?

While it may be argued that a truly systematic nomenclature would help resolve these problems, the reality is that online substructure search is about to displace traditional means of searching the literature to a large extent and make the role of systematic names of any sort more secondary than it is now.

(3) The requirement for linear representation of structures extends beyond the scientific community to meeting the needs of commerce and law for accurate descriptions of chemical substances. Would these users' needs be better met by a more systematic nomenclature?

Such users are largely indifferent to the kind of description. They are mostly not expert in any system, and in the absence of a recognized generic name of the type adopted for drugs and pesticides, the CA Registry Number would generally provide a more convenient description. It has already been adopted by several organizations, such as the EPA and ECDIN. Its eight or so digits can easily be copied and checked. Through an online terminal, names and synonyms can quickly be obtained, although easier and surer cross-reference than now between numbers for different salts and stereoisomers is a growing need. If a precise, linear description is required, the Wiswesser line notation[4] can provide concise ciphers, and it demands much less learning effort than nomenclature.

(4) In amplification of the above statement concerning disadvantages of fully systematic names, namely, complex punctuation and inherent monotony, one may instance examples in the paper on general principles of nodal nomenclature[1] (Figure 2), where widely different skeletons have the same root name simply because they have the same number of nodes, and Dyson's unpublished monograph, "Some new concepts in organic chemical nomenclature".[5] The latter is particularly valuable because it exemplifies names for many typical compounds, not just basic chains and ring systems (Figure 3). At this level of analysis, ciphers seem more appropriate than names, the two being virtual alternatives.

Additionally, the experience of those concerned with coining generic names for drugs and pesticides should be noted, although the objective here is somewhat different. Although, at first, names were constructed almost entirely from relevant chemical syllables, it came to be recognized that this led to too many similar-sounding, multisyllabic names and that the introduction of entirely different syllables was desirable, particularly where such a syllable could designate a significant moiety common to members of a class.

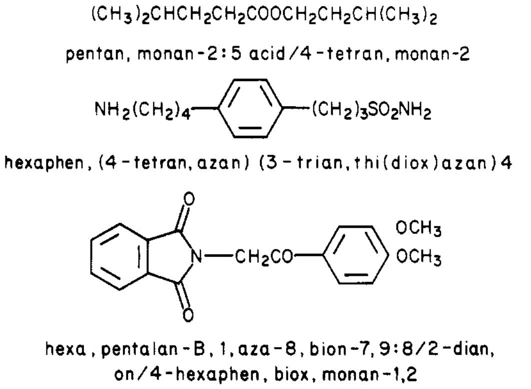
(CH3)2CHCH2CH2COOCH2CH2CH(CH3)2
pentan, monan-2:5 acid/4-tetran, monan-2
NH2(CH2)4
(CH2)3SO2NH2
hexaphen, (4-tetran, azan) (3-trian, thi(diox)azan)4
O
N
CH2CO
OCH3
OCH3
O
hexa, pentalan-B, 1, aza-8, bion-7, 9:8/2-dian,
on/4-hexaphen, biox, monan-1,2

Figure 3. Dyson's nomenclature.

These considerations emphasize the need to consider descriptiveness at an appropriate cognitive level. Names are for human communication, while ciphers and connection tables provide the descriptions required for unambiguous identification and computer processing.

(5) The approach shown by nodal nomenclature provides a unique numerical locant for every node of a graph, i.e., an overall numbering of an entire skeleton (rings plus chains), and for ring systems per se it selects the largest encompassing ring rather than the smallest set of smallest rings. While this may be advantageous from a systematic, taxonomic viewpoint, it is quite at variance with current practice, except for bridged rings. Since nodal nomenclature initially disregards the chemical nature of a node, be it atom or ring system, perhaps this nomenclature should be looked upon as a more generic descriptive system, which operates at a lower level than conventional nomenclature, and therefore supplements rather than supplants it.

If it is accepted that for ring systems a new approach, such as Goodson's development[1] of Taylor's proposals, is desirable, it would appear that some additional concepts should be introduced to take account of geometric fundamentals. The present description of ring fusions in terms of locants works smoothly only for simple edge-fused ring systems. With peri-fused systems it becomes quite irregular, as shown particularly with highly symmetrical structures like coronene, while for cage parents the complexity of the locant sets is surely a sign that a nonoptimal algorithm has been devised to describe a regular enumeration pattern. (Moreover, how many chemists would in practice translate such descriptions confidently back to structural diagrams?) In the case of peri-fused systems, I showed many years ago[6] that by recognizing the characteristic Y atom which occurs in these structures, concise, elegant descriptions and regular enumeration patterns could be obtained. For bridged and cage structures it might help to recognize the Platonic solids (the five regular polyhedra, namely, tetrahedron, octahedron, cube, icosahedron, and dodecahedron in hierarchical order[7]) as fundamental structures in their own right, and assign them parent names upon which others could be built. Adamantane, for example, can be seen as a tetrahedron with CH occupying each vertex and with CH_2 inserted into each edge, and this approach could be extended to deal with a number of multicyclic bridged systems. Many carboboranes have the icosahedral structure.

In summary, I am arguing that the time for directing effort to devising a new, comprehensive system of organic nomenclature is past, because the role of systematic names is declining in importance with the development of the CA Registry Number and its associated connection table as a unique descriptor and of online searching systems for substance identification. Further developments should be restricted to areas where existing systems are inadequate to meet practical needs.

REFERENCES AND NOTES

(1) Goodson, A. L. "Graph-Based Chemical Nomenclature. 2. Incorporation of Graph-Theoretical Principles into Taylor's Nomenclature Proposal". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 172-176.

(2) Lozac'h, N.; Goodson, A. L.; Powell, W. H. "Nodal Nomenclature—General Principles". *Angew. Chem., Int. Ed. Engl.* **1979**, *18*, 887-889.

(3) IUPAC, "Nomenclature of Organic Chemistry: Sections A, B, C, D, E, F, and H"; Pergamon Press: Oxford, 1979.

(4) Smith, E. G.; "The Wiswesser Line-Formula Chemical Notation"; Chemical Information Management Inc.: Cherry Hill, NJ, 1976.

(5) Private communication, Chemical Notation Association (UK Chapter).

(6) Silk, J. A. "An Improved System for the Enumeration and Description of Ring Systems" *J. Chem. Doc.* **1961**, *1*, 58-62.

(7) Critchlow, K. "Order in Space"; Thames and Hudson: London, 1969.

Computer Storage and Retrieval of Generic Chemical Structures in Patents. 1. Introduction and General Strategy

MICHAEL F. LYNCH,* JOHN M. BARNARD, and STEPHEN M. WELFORD

Postgraduate School of Librarianship and Information Science, University of Sheffield, Western Bank,
Sheffield S10 2TN, United Kingdom

Received March 25, 1981

The strategy of an approach to representing and searching the generic chemical formulas (Markush formulas) typical of chemical patents is outlined. The methods under development involve the following stages: (a) the description of generic chemical expressions by means of a formal language, GENSAL; (b) an approach to the generation and recognition of substituents or radicals defined by generic nomenclatural expressions, via formal grammars; (c) methods for automatic generation of screen characteristics, individually and within the relational and logical frameworks defined by generic formulas; (d) search techniques for identification of specific structures and substructures within generic formulas based on these methods.

INTRODUCTION

Chemists make extensive use of generic chemical nomenclature; arguably, they use it more than specific chemical nomenclature except when they are searching an index or data base for specific chemical molecules. Expressions such as phenoxyalkanolamines, nitro-substituted anthraquinones, etc., abound in the literature and in chemists' conversation. The use of generic chemical nomenclature attains particular importance in chemical patents, where classes of molecules are described which may be either finite or potentially infinite in number, depending on the constraints placed on the possible position and variety of substituents or other variable characteristics. The economic importance of adequate information systems in this area needs little emphasis.

While computer-based chemical information systems are now widely used for the retrieval of specific structures and groups of compounds related by their having substructures in common, the methods available for representing classes of molecules lag far behind. The problems posed are substantially more complex, so that, for instance, the World Patents Index of Derwent Publications Ltd., and, in particular, the Farmdoc, Agdoc, and Chemdoc services, depends on manually assigned chemical fragmentation codes which suffer from substantial defects which have recently been documented by Kaback.[1] Facilities for generic structure searching are also included in the system of International Documentation in Chemistry, using the GREMAS code,[2] and the IFI/Plenum data base.[3]

The example of a generic chemical structure shown in Figure 1 illustrates the problems of representing and retrieving such information whether in a search for a single structure which may or may not be a member of the class described or in a search for a substructure which may lie partly in an invariant part of the structure and partly in a variant part. Determining automatically whether two classes of molecules described by generic expressions of this type have members in common poses yet more difficult problems.

The example of Figure 1, a relatively simple case, illustrates the presence of a constant skeleton, variable components X and Y with a logical exclusion relating them, and substituents on the phenyl group, the number, nature, and position of which are variable (in many instances, the invariant part of the molecule may be vestigial). This generic expression is estimated to represent at least 10 000 different molecules, but if the limit on the size of the alkyl group were removed, the number would become infinite. Simpler examples can readily be found, where the positions of substituents are not variables, and the substituents are given as a list of discrete members; these are also familiar from publications reporting the preparation of series of analogous compounds for testing.

Previous work on means of representing generic chemical structures includes that of Sneed, Turnipseed, and Turpin[4], who described an adaptation of the Hayward notation for the description of Markush structures, limiting it to what they called determinate structures, i.e., those with variable radicals, the members and positions of which are specified, and those with a diradical (such as methylene) which occurs a variable number of times. These constraints limit the applicability of the method to a small subset of generic formulas. Geivandov subsequently outlined methods which relaxed these rigid constraints.[5] More recently, Silk has reviewed the services available and has made suggestions for the facilities needed for the improvement of these services.[6] Krishnamurthy and Lynch have described a preliminary study of the problems and have suggested a possible methodology for their solution,[7,8] within the context of application of the ALWIN notation.[9,10]

REQUIREMENTS OF A GENERIC STRUCTURE SEARCH SYSTEM

The types of searches which a fully developed generic structure search system should support include the following:

(a) searches for specific molecules within a given generic expression;

(b) searches for substructures within generic expressions, regardless of whether the substructure lies wholly or only partly within the invariant part of the structure;

0095-2338/81/1621-0148$01.25/0 © 1981 American Chemical Society