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.

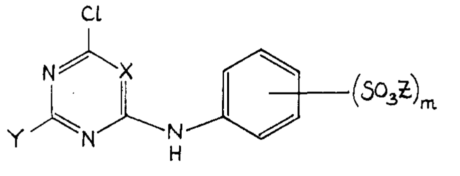
Previous work on means of representing generic chemical structures includes that of Sneed, Turnipseed, and Turpin[4], who described an adaptation of the Hayward notation for the description of Markush structures, limiting it to what they called determinate structures, i.e., those with variable radicals, 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;



WHERE X IS N OR >C-CL ; WHEN X IS N,
Y IS 1-6 C ALKOXY, AND WHEN X IS >C-CL,
Y IS CL ; Z IS ALKALI METAL OR AMMONIUM;
M IS 1 OR 2.

Figure 1. A generic chemical expression typical of these found in patents.

(c) determination of whether two or more generic expressions have specific molecules in common.

These requirements demand a representation in which potentially all of the class members described in the expression may be generated, at least to the extent to which the expressions are well-defined (a formal definition of the term "well-defined" is not practicable at this stage). The potential for generation of individual molecules does not necessarily imply that all such members need be generated explicitly; indeed, given the great variety of structures often possible, this would be computationally unfeasible. Analysis of generic expressions based on a wide variety of patent documents suggests that the classes are identified principally by means of expressions such as alkyl, cycloalkyl, heterocycle, aryl, or aralkyl. The members of certain of these, including alkyl and cycloalkyl, may be generated over finite ranges, although this process is inherently time consuming, even when quite low limits to the compass of the groups are set, as shown by Masinter et al.[11] in relation to the exhaustive generation of isomeric groups. What is less clear at present is the extent to which expressions such as aryl may support potentially exhaustive generation, although it appears likely that the recognition of specific groups, by assignment to classes, may be more readily achieved. Other property-defined expressions such as "electron-withdrawing group", "easily hydrolyzable group", etc., also occur, but are not amenable to generation. It may be possible to deal with these via appropriate dictionaries. Functional groups, on the whole, are separately defined, as in chloroalkyl, alkoxycarbonyl, alkoxy, etc., and only rarely are other formulations used which imply even simple functional groups such as ethers or amines by expressions such as "polymethylene group interrupted by oxygen or nitrogen atoms".

The generic formulas included in patents, as exemplified in Figure 1, are interpretable by the expert in chemical nomenclature, yet are not sufficiently well formalized to serve as a representation for input and manipulation within a computer-based information system. Thus a primary requirement appears to be to provide a sufficiently exact and formalized statement of the structural and logical relations in these expressions to permit the algorithmic generation of any member of the class denoted by a well-defined expression, although this capability may be called into action only in rare cases.

A second requirement, bearing in mind the need in operational systems for substructure search of files of many millions of individual chemical substances, is the ability to create descriptions of localized areas of the molecules, in the form of both specific and generalized screens. These provide the indexing of structures to permit, in suitable combinations, effective file reductions to be carried out prior to atom-by-atom searches, or, indeed, to obviate them entirely in many circumstances. Thus it is necessary to provide a means of characterizing both the invariant and the variable components of generic chemical structures and to describe their interrelations, so that structural searches of the kinds described above can be performed.

In more specific terms, the requirements call for a full and unambiguous description of generic structures, the creation of screen descriptions from these, and for both approximate and exact matching algorithms. In addition to recognizing that limitations on these capabilities may be posed by the degree of definiteness or indefiniteness of the expressions in the original descriptions, it must be noted that further difficulties are posed by specifications such as "optionally substituted", as opposed to "optionally substituted by", to which is appended a list of specific or generic expressions.

GENERAL STRATEGY

These considerations have led us to develop a formal language which appears well suited to the representation of the great majority of generic structure descriptions published in patents and capable in many instances of translation into an extended form of connection table. Further, we have considered methods, on the basis of formal grammar theory, which would appear to permit the generation of individual members of certain classes of radicals included in generic formulas and, through this, the automatic assignment of screens which fully describe, through limitation to local environments, both the invariant components and the potential variety of well-defined variable components.

Reexamination of the choice of structural representation and the increasing opportunities for use of chemical structure graphics have lead us to select the connection table as the primary medium for storage and manipulation. Close examination of the complexities which result from use of linear notations, for instance, the problems of specifying substituent positions on ring systems where the locants of substitution may be dependent on a ring of variable size which is cited earlier, has convinced us of the advantages of this choice.

In this work, we have had in mind the appropriateness of applying our earlier work on the generation of algorithmically defined screens, i.e., those in which the statistics of occurrence of both atom and bond types, and their combinations, are considered in connection with files of specific molecules.[12] In that work, a methodology of some generality was developed, in which the probability of occurrence of members of particular fragment types in a sample of a data base was used as a guide to the degree of generality or specificity at which screens derived from each fragment type were to be expressed. The results of this earlier work are reflected in the screen set used in the substructure search system developed by the BASIC group at Basel,[13] which has more recently been adopted by the Chemical Abstracts Service as the basis for CAS ONLINE, inaugurated in 1980 as a trial service.[14]

Theoretical considerations aside, there are substantial practical reasons for the adoption of an existing and well understood approach to screening, not least the probable familiarity of the method to information scientists. While it must be noted that the detailed statistical basis for the selection of a screen set is rather less well founded in this instance, where the reality of the domain of known molecules has less relevance to the domain of all those potentially included in patent claims, the general principle still appears to be valid.

Little has been published on the effectiveness of screening systems such as these in operational substructure search systems dealing with specific molecules. It can readily be assumed that screening sets of a similar nature will perform much less effectively when used to represent generic formulas. Hence it would appear essential to provide multiple levels of screening at which, successively, further relational information about the substitutional and positional options and limitations are brought into play. As yet, only preliminary attention has

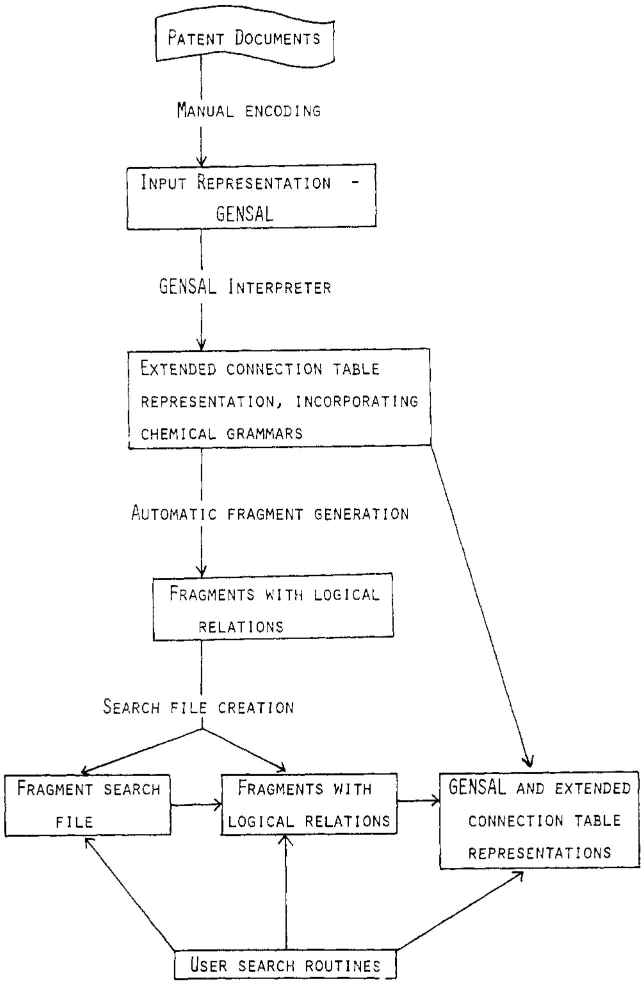

PATENT DOCUMENTS
MANUAL ENCODING
INPUT REPRESENTATION - GENSAL
GENSAL INTERPRETER
EXTENDED CONNECTION TABLE REPRESENTATION, INCORPORATING CHEMICAL GRAMMARS
AUTOMATIC FRAGMENT GENERATION
FRAGMENTS WITH LOGICAL RELATIONS
SEARCH FILE CREATION
FRAGMENT SEARCH FILE
FRAGMENTS WITH LOGICAL RELATIONS
GENSAL AND EXTENDED CONNECTION TABLE REPRESENTATIONS
USER SEARCH ROUTINES

Figure 2. Outline scheme of intended overall process.

been devoted to this aspect of the work, but it appears that several options exist for reflecting these needs.

The considerations have led us to the scheme outlined below:

(a) the formalized description of generic formulas

(b) the possibility of potentially exhaustive generation and recognition of members of classes of molecules described by generic chemical formulas, via formal grammars,

(c) the generation of screens, individually and within the relational and logical frameworks defined by the generic expressions, which describe the members of the classes.

This approach can be summarized in Figure 2 in which the elements and stages of the procedures are delineated.

Much theoretical knowledge of the basis on which artificial languages such as programming languages may be rationally constructed is now available. Thus the theory of grammars, as described in Aho and Ullman[15] and Cleaveland and Uzgalis,[16] greatly facilitates the design of a formal language so that complete and efficient recognizers or translators may be implemented.

The formalized input language is GENSAL. Its structure is described in part 2 of this series.[17] Here it is necessary only to say that its design, coupled with the generative and recognitive grammars, initial work on which is described in part 3 of the series,[18] is intended to support the generation of extended generic structure connection tables, in which relational symmetric list structures permit the description of the substituents and substitution patterns and other features evident in generic chemical formulas, together with the logical relations prescribed therein.

ACKNOWLEDGMENT

We thank Dr. J. Silk of ICI Plant Protection Division, Ltd., P. Norton and staff of Derwent Publications Ltd., members of the AIOPI Patents Study Group, and Dr. P. Willett for helpful discussions and advice. We acknowledge the support of the Department of Education and Science in providing Information Science Research Studentships to J.M.B. and S.M.W.

REFERENCES AND NOTES

(1) S. M. Kaback, "Chemical Structure Searching in Derwent's World Patent Index", *J. Chem. Inf. Comput. Sci.*, **20**, 1–6 (1980).

(2) E. Meyer, "The IDC System for Chemical Documentation", *J. Chem. Doc.*, **9**, 109–113 (1969).

(3) M. Z. Balant and J. M. Emberger, "A Unique Fragmentation System for Indexing Patent Literature", *J. Chem. Inf. Comput. Sci.*, **15**, 100–104 (1975).

(4) H. M. S. Sneed, J. H. Turnipseed, and R. A. Turpin, "Line-Formula Notation System for Markush Structures", *J. Chem. Doc.*, **8**, 173–178 (1968).

(5) E. A. Geivandov, "A Language for Notation of Generalised Structures of Organic Compounds Containing Alternative Delocated Fragments (Markush Structures)", *Nauchno-Tekh. Inf., Ser. 2*, (10) 21–24, 46 (1972).

(6) J. A. Silk, "Present and Future Prospects for Structural Searching in the Journal and Patent Literature", *J. Chem. Inf. Comput. Sci.*, **19**, 195–198 (1979).

(7) E. V. Krishnamurthy and M. F. Lynch, "Coding and Analysis of Generic Chemical Formulae in Chemical Patents", *J. Inf. Sci.*, **3**, 75–79 (1981).

(8) E. V. Krishnamurthy and M. F. Lynch, "Formal description, Coding and Computer Handling of Generic Formulae in Chemical Patents", British Library Research and Development Report No. 5490 (1979).

(9) E. V. Krishnamurthy, P. V. Sankar, and S. Krishnan, "ALWIN-Algorithmic Wiswesser Notation System for Organic Compounds", *J. Chem. Doc.*, **14**, 130–149 (1974).

(10) S. Krishnan and E. V. Krishnamurthy, "Compact Grammar for Algorithmic Wiswesser Notation Using Morgan Name", *Inf. Process. Manage.*, **12**, 19–34 (1976).

(11) L. Masinter, N. S. Sridharan, J. Lederberg, and D. H. Smith, "Applications of Artificial Intelligence for Chemical Inference. XII. Exhaustive Generation of Cyclic and Acyclic Isomers", *J. Am. Chem. Soc.*, **96**, 7702–7714 (1974).

(12) M. F. Lynch, "Screening Large Chemical Files in Chemical Information Systems", J. Ash and E. Hyde, Eds., Ellis Horwood Ltd., Chichester, 1974, pp 177–194.

(13) W. Graf, H. K. Kaindl, H. Kneiss, B. Schmidt, and R. Warszawski, "Substructure Retrieval by Means of the BASIC Fragment Search Dictionary Based on the Chemical Abstracts Service Chemical Registry III System", *J. Chem. Inf. Comput. Sci.*, **19**, 51–55 (1979).

(14) N. A. Farmer and M. P. O'Hara, "CAS ONLINE. A New Source of Substance Information from Chemical Abstracts Service", ***Database***, 10–25, Dec, 1980.

(15) A. V. Aho and J. D. Ullman, "The Theory of Parsing, Translation, and Compiling", Prentice-Hall, Englewood Cliffs, NJ, 1972.

(16) J. C. Cleaveland and R. C. Uzgalis, "Grammars for Programming Languages" (Programming Languages Series, 4), Elsevier, New York, 1977.

(17) J. M. Barnard, M. F. Lynch, and S. M. Welford, "Computer Storage and Retrieval of Generic Structures in Patents. 2. GENSAL, A Formal Language for the Description of Generic Chemical Structures", *J. Chem. Inf. Comput. Sci.*, following paper in this issue.

(18) S. M. Welford, M. F. Lynch, and J. M. Barnard, "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 3. Chemical Grammars and their Role in the Manipulation of Chemical Structures, *J. Chem. Inf. Comput. Sci.*, paper 3 in this series.