Chemical Substance Retrieval System for Searching Generic Representations. 1. A Prototype System for the Gazetted List of Existing Chemical Substances of Japan

YOSHIHIRO KUDO*

Japan Association for International Chemical Information, Gakkai Center Building, 2-4-16 Yayoi, Bunkyo-ku, Tokyo 113, Japan

HIDEAKI CHIHARA

Department of Chemistry, Faculty of Science, Osaka University, Toyonaka 560, Japan

Received July 9, 1982

A prototype information retrieval system has been developed to search for either a specific substance or a family of substances of which the query compound is a member. The query itself can be generic. The data base of the system consists of a name file and a notation file, the latter being searched with specially designed representations as the keys. Three different representations of varying levels of generality were designed to permit a generic search corresponding to a specific compound the searcher inputs. A small-scale data base was built from the gazetted list of Existing Chemical Substances (Japanese legislation). Examples of searches are given.

INTRODUCTION

In chemical documents, particularly in chemical patents and inventory lists under laws, chemical substances are described in wide variety of ways.[1] They range from very generic Markush formulas to specific names. To search such documents for a particular substance or a group of substances is very important but tedious work which makes information specialists in industries feel uncomfortable because ready systems of generic representations searchable by computers are not always easily available. CAS ONLINE[2] of Chemical Abstracts Service is the only practical system in the world for searching nearly 6 million of chemical structures stored in the Chemical Abstracts Service Resistry file, but it is a file of specific substances. The GREMAS[1] screens are well designed for searching generic formulas, but its coding is not easy except to specially trained searchers. GENSAL by Lynch et al.[3] is a kind of comprehensive compiler which would be convenient both for a searcher and a data base builder because it provides easy communication between them and the data base system. Their articles illustrate their general design of the system, but, so far, how to search a data base of generic representations has not been fully disclosed yet.

This paper will describe a pilot system specially designed for searching generic or specific names in the Handbook of Existing Chemical Substances gazetted by the Japanese Government. The system also has a potential capability of being applicable to similar data bases of generic chemical names.

The data base used to illustrate the system is a small one, but it is large enough to show how the searching works.

LIST AND HANDBOOK OF EXISTING CHEMICAL SUBSTANCES

The List of the Existing Chemical Substances is the attachment of "The Law of Examination and Regulation of Manufacture, etc. of Chemical Substances" legislated originally in 1973 and has been updated several times. The List is published in the form of a handbook, the "Handbook of Existing Chemical Substances", both in Japanese and in English. The 1981 English edition[4] consists of two parts: (1) About 20000 names of Existing Chemical Substances and (2) about 600 names of New Chemical Substances added afterward. Each name is assigned a unique number (referred to as the G number or GN in this paper) which is called the class-reference number in the Handbook. Also, there is another numbering system which corresponds to structure classes (referred to as the S number or SN in this paper). In the Handbook the names are classified according to the chemical and structural features. For instance, squalane (GN 9-762) is located in the subsection "Hydrocarbons/Saturated hydrocarbons" of Section 2 which lists Low Molecular Chain-Like Organic compounds, as shown in Table I, and is given as SN 2-11 which means the 11th name of Section 2. Because the List is not a scientific but a legal document, no single system of nomenclature is adopted, but a wide variety of names may be found in the List and Handbook, e.g., octene, alkene (C 10-50), and poly (3-5) chloropropane, which look almost arbitrary.

Because the number of names in the Handbook is not large, it may seem not to be extremely difficult for a searcher to judge if a particular substance is contained in the List as a specific substance or as a member of a generic name. However, for the purpose of such judgment, the chemical and structural feature is the only key for a systematic search which leads only to a section or subsection which may possibly contain the answer. A particular name can only be used afterward for browsing through the section or subsection. Therefore, one can hardly be certain about the search result, particularly when he gets no hits, unless he uses a sufficient length of time and care to allow for one-by-one comparison.[4]

ANALYSIS OF NAMES IN THE HANDBOOK

Various types of names are used indiscriminately as examples in Table I illustrate, and these have been taken from the 1981 English version. There is no systematic nomenclature employed in naming the chemical substances in the Handbook. The same substance may be and indeed is located at more than one place by different names. It will be useful to examine Table I in order to understand the features of the Handbook that the present paper will treat.

Methane (SN 2-1) through propane (SN 2-3) are specific names. Butane (SN 2-4) through nonane (SN 2-9) are groups of hydrocarbons each having the same molecular formula. Alkane (C10-29) (SN 2-10) is more generic and can be represented with a single generic molecular formula, C_nH_{2n+2} (n = 10-29). The name for GN 2-176 (SN 2-160 and 2-201) contains a clause to describe the restriction of the range of generality and is located at the subsections of "saturated amine" (SN 2-160) and "unsaturated amine" (SN 2-201). The Handbook provides multiple names. Thus,

0095-2338/83/1623-0109$01.50/0

Table I. Examples of Names in the Handbook

```
   Section 2.   Low Molecular Chain-like Organic compounds

      B 1  Hydrocarbons
      B 11 .Saturated hydrocarbons

(SN)     (GN)

                    CnH2n+2
2-1        2-1         Methane
2-2        2-2         Ethane
2-3        2-3         Propane
2-4        2-4         Butane
2-8        2-8         Octane
2-9        2-9         Nonane
2-10       2-10        Alkane (C 10-29)
2-11       9-762       Squalane
2-12       9-1317      2,6,10,15,19,23-hexamethyltetracosane

      B 12 .Unsaturated hydrocarbons
           ..with one double bond

                    CnH2n
2-13       2-12        Ethylene
2-14       2-13        Propylene
2-15       2-16        Butene
2-16       2-19        Pentene
2-17       2-22        1-Hexene
2-18       2-2359      1-Heptene
2-19       2-24        Octene
2-20       2-25        Nonene
2-21       2-27        Alkene(C 10-50)
2-22       2-29        1,1-dialkyl (C4-10) vinylidene
             .
             .
             .
B 2 Halogenated hydrocarbons

B21   .Saturated halogenated hydrocarbons
      ..with halogen other than fluorine
      ... derivins from propane
2-63       2-80        Monochloropropane
2-64       9-576       1,3-Dichloropropane
2-65       2-81        1,2-Dichloropropane
2-66       2-83        Poly(3-5)chloropropane
2-67       2-73        1-Bromopropane
2-68       2-76        Isopropyl bromide
2-69       2-84        1,2,3-Tribromopropane
2-70       9-370       1-Chloro-3-bromopropane
2-71       9-1247      1-Bromo-3-chloropropane
2-72       9-2007      1-Bromo-3-chloropropane
2-73       2-82        1,2-Dibromo-3-chloropropane
             .
             .
2-160      2-176       N,N,N-Trialkyl(or alkenyl, at least one
                         of the alkyl or alkenyl group is C
                         8-24, others are H or C 1-5) amine
                           (See also SN 2-201)
             .
             .
2-201      2-176       (See SN 2-160)
             .


Section 3   Low molecular Carbo-monocyclic Organic Compounds.

3-1        3-1         Benzene
3-2        3-60        Mono-(or di-) methyl (ethyl,bromoallyl,
                         bromopropyloxycarbonyl or chloro-
                         propyloxycarbonyl)benzene
                         (See also SN 3-61,3-1399,3-1480)
           3-2         Toluene
           3-28        Ethylbenzene
           3-3         Xylene
           3-13        Diethylbenzene
3-3        3-21        n-Alkylbenzene(C 3-36)
           3-11        Butylbenzene
3-4        3-22        Branched alkylbenzene(C 3-36)
3-5        3-15        Alkyl (C 2-4)toluene
           3-12        Cymene
3-6        3-27        Dodecyl toluene
3-7        3-25        Dialkylbenzene(C 3-36)
3-8        9-1782      Dialkyl benzene (C 10-13)
3-9        3-3427      Trialkyl (C 1-4) benzene
3-10       3-7         Tri- or tetramethylbenzene
3-17       3-31        Monochlorobenzene
3-18       3-41        Dichlorobenzene
3-23       9-1869      Dichloro-m-xylene
3-31       3-52        p-Dibromobenzene
3-61       3-60        (See the first name at SN 3-2)
3-1399     3-60        (See the first name at SN 3-2)
3-1480     3-60        (See the first name at SN 3-2)
```

squalane (SN 2–11) and 2,6,10,15,19,23-hexamethyltetracosane (SN 2–12) are the same molecule. 1-Chloro-3-bromopropane (SN 2–70) and 1-bromo-3-chloropropane (SN 2–71, 2–72) are separate entities. Among the five entries of SN 3–2, the first name, "mono- (or di-) methyl (ethyl, bromoallyl, bromopropyloxycarbonyl, chloropropyloxycarbonyl) benzene" (GN 3–60) envelopes the other names, toluene (GN 3–2), ethylbenzene (GN 3–28), xylene (GN 3–3), and diethylbenzene (GN 3–13). The name for GN 3–60 (SN 3–2) is given three other S numbers, SN 3–61, 3–1399, and 3–1480, according to which substructural feature of the compounds associated with the name is used to classify it. Dialkyl benzene (C10–13) (SN 3–8) is included in a more general group, dialkylbenzene (C3–36) (SN 3–7). Trialkyl (C1–4) benzene (SN 3–9) and "tri- or tetramethylbenzene" (SN 3–10) have a common member, trimethylbenzene.

From the viewpoint of the specificity of locants, *p*-dibromobenzenes (SN 3–31) is a specific name, dichlorobenzene (SN 3–18) is a generic name, and dichloro-*m*-xylene (SN 3–23) is a hybrid name.

The classification in the Handbook is based not on simple structural characteristics but rather on chemical families. For example, thiols and thioethers are treated as if they were derivatives of alochols and ethers. Similarly, cyclic anhydrides of acyclic acids are located in the same subsections for corresponding acyclic substances.

The List contains indefinite names as well as definite ones. The present system makes further divisions of the so-called indefinite names into ambiguous indefinite names and unambiguous indefinite names. Tetramethylammonium salt is an example of the former because there is no structural information on its anionic portion. Trialkylamine is an example of the latter, which may be treated as if it were a definite name if we adopt an appropriate method of representation for the group of compounds. Thus, it can be treated as if it were a definite name because the term "alkyl" has a very restricted structural meaning, C*n*H2*n*+1 when it is not substituted.

OUTLINE OF THE SEARCH SYSTEM

The most frequent type of queries we anticipate are the questions to ask whether or not a specific substance in which a searcher is interested is contained in the Handbook either in its specific name or in a generic name of which the substance may be a member. Also, a searcher may want to know what other members of an isomeric family of the substance are listed in the Handbook.

The present search system was designed to meet such requirements. To achieve this end, it is necessary to devise several representations of structure at different levels and from different scopes of specificity and to make allowances for generic search. Three representations were used, which are not completely orthogonal, as will be explained in a later section.

The system has a data base consisting of a notation file and a name file whose records are logically linked through the G numbers (GN) with each other. The notation file contains three representations, Q, R, and S, for each name. A query is formulated in the form of a specific structural formula by the searcher, who then specifies whether he wants an exact hit or a family hit by inputting a Markush indicator code.

On accepting the query, the system (i) generates a set of necessary representations according to the same generating procedure as used to build the notation file, (ii) searches through the data base, and finally (iii) displays the set of the relevant names and their associated G numbers.

Figure 1 illustrates what a searcher can retrieve when he inputs a specific query, 1,2-dichloropropane, or a generic query, dichloropropane. In the first case, only 1,2-dichloropropane (SN 2–65, GN 2–81) is retrieved. On the other hand, in the second case, the same structure is input and is converted internally to the generic structure by setting a value of the Markush indicator. The resulting answer includes 1,3-dichloropropane (SN 2–64, GN 9–576) as well as the 1,2-isomer (see also Table IV).

As explained below, a certain kind of generic query can be accepted which would correspond to such a name as "chlorinated (0–2) and/or brominated (0–1) alkane (C1–3)".

Table II. Aspects in the Judgement on the Relevancy

query	structures in an assumed structure file[a]			
	n-C4H10	IsoC4H10	C4H10	C*n*H2*n*+2 (*n* = 3–5)
1 *n*-C4H10	M	U	M	M
2 IsoC4H10	U	M	M	M
3 C4H10	M	M	M	M
4 C*n*H2*n*+2 (*n* = 3–5)	M	M	M	M
5 C*n*H2*n*+2 (*n* = 2–4)	M	M	M	M
6 C*n*H2*n*+2 (*n* = 1–3)	U	U	U	M
7 C*n*H2*n*+2 (*n* = 5–7)	U	U	U	M
8 C*n*H2*n*+2 (*n* = 1,2)	U	U	U	U
9 C*n*H2*n*+2 (*n* = 6,7)	U	U	U	U

[a] M = matched and U = unmatched.

CRITERION OF JUDGMENT ON RELEVANCY

The strategy in designing the system was based on the notion that its most important role is to give users an appropriate warning as to whether the chemical substances of interest are or are not subject to regulations by laws. Therefore, it should intend to attain a complete recall, the complete relevancy being of second importance. If noise is not excessive or its elimination is too time- and cost-consuming, incompletness in relevancy may be permitted.

Table II shows the hypothetical results of application of the criterion. It is assumed that the system file consists of four names: butane, isobutane, C4 (saturated hydrocarbon of C4H10), and C3–5 (a group of saturated hydrocarbons in which the number of carbon atoms are three to five). When a query is butane or isobutane itself, C4 and C3–5 are relevant, and they are not relevant for each other. Obviously, for a query of C4, all are relevant, and for C6–7, none are relevant. For both C1–3 and C5–7, C3–5 is relevant because they are partly overlapped.

In cases where indefinite names are involved, the situation becomes more complicated as shown in Table III. Indefinite names are always relevant for each other. Such a consideration is extremely important in designing the present system which is required to attain the complete recall of relevant names.

FUNCTION OF MARKUSH INDICATOR

The conception of the Markush indicator is one of the most important special devices in the present system. In the prototype system, the Markush indicator is set either as specific or generic. Other values of the indicator are saved to cope with more complicated situations, for example, in the treatment of patent documents. Although its function will be shown in the examples in the later sections of this paper, Figure 1 and Table IV illustrate the features of the function of the Markush indicator.

Table IV summarizes the results of the four generic and four specific queries by using the four different isomers of dichloropropane, with and without activating the Markush indicator, respectively. It should be noted that only the 1,2- and 1,3-isomers are registered in the Handbook. If the Markush indicator is set specific, the system searches only for a specific

a

```
QUERY No. (0 FOR END) ? 1
Bond value (-1 for end) ? 1
 Atom (0,0 for end) ? 1
 Atom (0,0 for end) ? 2
 Atom (0,0 for end) ? 3
 Atom (0,0 for end) ? 4
 Atom (0,0 for end) ? 0
 Atom (0,0 for end) ? 3
 Atom (0,0 for end) ? 5
 Atom (0,0 for end) ? 0
 Atom (0,0 for end) ? 0
Bond value (-1 for end) ? -1


The connection table.

 1  2 ( 1 )
 2  1 ( 1 ) 3 ( 1 )
 3  2 ( 1 ) 4 ( 1 ) 5 ( 1 )
 4  3 ( 1 )
 5  3 ( 1 )                    (OK), no ? OK
Heteroatom (0 for END) ? CL
  No. (0 for End) ? 1
  No. (0 for End) ? 5
  No. (0 for End) ? 0
Heteroatom (0 for END) ? 0
 1 2 3 4 5
CL . . .CL   (Y)/N ? Y

Markush indicator  (specific or generic).
    Generic  (Y)/N ? N

Input end.
The searching began.
Searched =  369    Found =  1
Retrieved names :
 1            20081        1,2-Dichloropropane
END.
```

b

```
QUERY No. (0 FOR END) ? 2
Bond value (-1 for end) ? 1
 Atom (0,0 for end) ? 1
 Atom (0,0 for end) ? 2
 Atom (0,0 for end) ? 3
 Atom (0,0 for end) ? 4
 Atom (0,0 for end) ? 0
 Atom (0,0 for end) ? 3
 Atom (0,0 for end) ? 5
 Atom (0,0 for end) ? 0
 Atom (0,0 for end) ? 0
Bond value (-1 for end) ? -1


The connection table.

 1  2 ( 1 )
 2  1 ( 1 ) 3 ( 1 )
 3  2 ( 1 ) 4 ( 1 ) 5 ( 1 )
 4  3 ( 1 )
 5  3 ( 1 )                    (OK), no ? OK
Heteroatom (0 for END) ? CL
  No. (0 for End) ? 1
  No. (0 for End) ? 5
  No. (0 for End) ? 0
Heteroatom (0 for END) ? 0
 1 2 3 4 5
CL . . .CL   (Y)/N ? Y

Markush indicator  (specific or generic).
    Generic  (Y)/N ? Y

Input end.
The searching began.
Searched =  369    Found =  2
Retrieved names :
 1            90576        1,3-Dichloropropane
 2            20081        1,2-Dichloropropane
END.
```

Figure 1. Retrieval by means of 1,2-dichloropropane: (a) specific search for 1,2-dichloropropane; (b) generic search for dichloropropane.

Table III. Relevancy for Indefinite Structures

query	structures in a hypothetical structure file[a]					
	C2H5Cl	C2H5Br	a chloride	a bromide	C2H5–?	CH3–?
1 C2H5Cl	M	U	M	U	M	M
2 C2H5Br	U	M	U	M	M	M
3 a chloride	M	U	M	M	M	M
4 a bromide	U	M	M	M	M	M
5 C2H5–?	M	M	M	M	M	M
6 CH3–?	M	M	M	M	M	M
7 CH3–O–CH3	U	U	U	U	U	M

[a] M = matched and U = unmatched.

Table IV. Role of the Markush Indicator in Searching for Isomers of Dichloropropane

	building up of the query		
query[a]	input structure[a]	setting of Markush indicator	answer
1,1-	1,1-	specific	no hit
1,2-	1,2-	specific	1,2-
1,3-	1,3-	specific	1,3-
2,2-	2,2-	specific	no hit
dichloropropane	1,1-	generic	1,2- and 1,3-
dichloropropane	1,2-	generic	1,2- and 1,3-
dichloropropane	1,3-	generic	1,2- and 1,3-
dichloropropane	2,2-	generic	1,2- and 1,3-

[a] 1,1-, 1,1-dichloropropane (not in the Handbook); 1,2-, 1,2-dichloropropane (GN 2-81, SN 2-65); 1,3-, 1,3-dichloropropane (GN 9-576, SN 2-64); 2,2-, 2,2-dichloropropane (not in the Handbook).

Table V. Examples of Interpretation of the Query by the System

input structure	Markush indicator	interpretation result
1,2-dichloropropane	specific	1,2-dichloropropane
	generic	dichloropropane
1,3-dichloropropane	specific	1,3-dichloropropane
	generic	dichloropropane
o-xylene	specific	*o*-xylene
	generic	xylene
α-methylsytrene	specific	2-phenylpropene
	generic	phenylpropene
β-methylstyrene	specific	1-phenylpropene
	generic	phenylpropene
3-phenylpropene	specific	3-phenylpropene
	generic	phenylpropene
1-bromo-2-chlorobenzene	specific	1-bromo-2-chlorobenzene
	generic	bromochlorobenzene

formula that the user inputs. Thus, only if the query is the 1,2- or 1,3-isomer is a corresponding specific hit obtained. On the other hand, when the Markush indicator is set to be generic, the system interprets the query as if it were a generic formula. Thus it becomes immaterial which one of the four isomers the searcher inputs as his query, and the same set of names (and GN's) are retrieved as shown in the last column of Table IV.

Table V shows some other examples of interpretation of a query formula according to the value of the Markush indicator by the system. The level of generality in searching for family compounds is such that the substances in the answer are limited to a close proximity of the query compound as defined in the Handbook. In other words, if the query compound is 1,2-dichloropropane, its positional isomers as well as "dichloropropane", if they are listed in the Handbook, are retrieved. A more generic form of queries, such as "trialkyl (or alkenyl, C2-5) benzene" and "tri- to hexachlorobiphenyl", can also be accepted.

STRUCTURE REPRESENTATIONS IN THE SYSTEM

(1) Introduction. As mentioned above, the notation system or the structure representation system is used to treat variety of generic formulas or names used in the Handbook. A notation for a query is generated, analogous to those in the notation file, and they are compared with each other during searching. Because it is the internal representation of a structural formula and because all processing is automatically performed, the user does not always need to know about it.

Generally, a connection table is considered as an alternative form of a specific structural formula and is very useful for computer manipulation, but it is not always the best solution for effective searching. Its preciseness and individuality

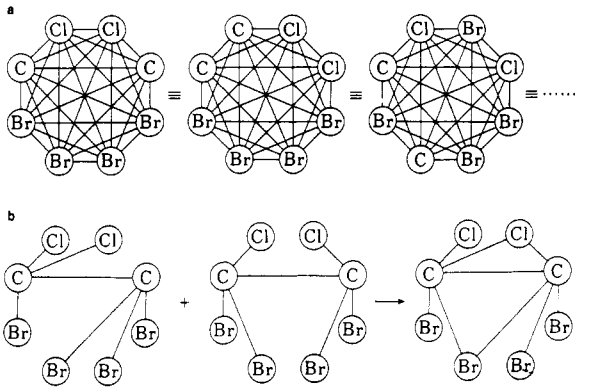


Figure 2. Application of the principle of the colored complete graph to dichlorotetrabromoethane: (a) a colored complete graph; (b) synthesis into a single graph.

sometimes make searching in terms of generic formulas difficult. In order to take care of ambiguous names at variety of levels, it would be necessary to combine multiple representations, which can be used selectively, depending on the extent of generality in searching that is required. Redundancy of the content in such a notation system would be counterbalanced by the easiness and efficiency of processing, but too redundant representations would lead to too large a data base. One must then compromise between the two restrictions, and this depends on the capability of the computer system. On the basis of such considerations, an unusual notation system has been designed which is aimed at handling the names in the Handbook.

(2) Principle of Colored Complete Graph (PCCG). In the List there are such names as to denote a large group of specific structures. Therefore, it is desired that a tool be devised to treat many such structures collectively in terms of as few structural representations as possible. The principle of colored complete graph (PCCG) is useful for this purpose.

Let us consider a complete graph, the number of whose members is N. In the complete graph, all of the N nodes have $N-1$ edges and are connected to all other nodes at least once and not more than once. If the N nodes can be distinguished by their properties, such as the kind of chemical elements, the graph is said to be colored. A chemical structural formula is a kind of colored graph in which different chemical elements (carbon, oxygen, etc.) correspond to different colors. A graph which is not colored also can be treated as a kind of a colored graph, which is monochromatic. Figure 2a shows an example of colored complete graph. It is derived from the structure of dichlorotetrabromoethane, where the number of nodes (N) is 8 and the number of colors is 3 (carbon, chlorine, and bromine). A colored complete graph can overlay any structure graph that has the same constitution as the complete graph. This is the principle of colored complete graph (PCCG). By this, for example, all structural isomers of dichlorotetrabromoethane are synthesized into a single colored graph as shown in Figure 2b.

The principle can be extended so that all the complete graph for C3 may cover the graphs for C1 and C2 if hydrogen atoms are not explicitly described. Thus, a skeleton of butanol covers those of propanol, ethanol, methanol, butane, propane, ethane, and methane.

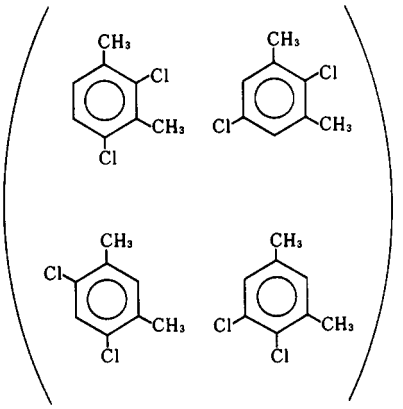
(3) S, Q, and R Representations. Three representations, S, Q, and R, are used internally in the present system which are considered to be able to cover all the generic levels that appear in the Handbook. To give the reader some sort of association, S, Q, and R denote approximately specific, qualitative, and ring structure representations, respectively. They are not perfectly mutually orthogonal, but some redundancy among them is desirable for efficient processing. The S and Q rep-

Table VI. Bond Order Parameter Values

	acyclic	cyclic
no bond	0	0
single bond	1	5
double bond	2	6
triple bond	3	7
special bond[a]	4	8

[a] Bonds other than simple bonds are called special bonds.

Chart I. Dichloro-*m*-xylene (GN 9-1869, SN 3-23)



resentations are composed of their component lists (S and Q lists) and their connection tables (S and Q connection tables).

Components of the S representation are atoms themselves, and those of the Q representation are atoms, atomic groups, and special valence bonds. The R representation is a partial set of attributes of ring systems as will be shown in Table VII. All connectivities inside or outside of ring systems are described with the Q and S representations. The three types of representations are designed to complement one another completely.

(4) S Representation. The role of the S representation is to describe a structure (or structures) as specifically as possible. Its components are atoms, and its connection table is the same as the conventional atomic connection table in most cases. For example, *p*-dibromobenzene (SN 3-31, GN 3-52) has two bromine atoms and six carbon atoms in its list. Its connection table describes the connectivities between all pairs of atoms in the list. For instance, there are two acyclic bromine–carbon single bonds and no bromine–bromine bonds. Table VI is the list of the bond order parameter values used to describe connectivities. The bond order parameter of a bond in a benzene ring is arbitrarily given a value of 8. The use of digital codes for this parameter coalesces subtle chemical distinctions of fractional bond orders and simplifies computer handling.

Let us take another example. The S list of dichloro-*m*-xylene (SN 3-23, GN 9-1869) is composed of two chlorine atoms and six carbon atoms, but its S connection table is the logical sum of those of 2,4-, 2,5-, 4,5-, and 4,6-dichloro-1,3-dimethylbenzene (Chart I). This processing becomes possible on the basis of the principle of colored complete graph (PCCG) as exmplained above. The graph can be described by using the set of the bond order parameters given in Table IV, but it is often too bulky because the graph is complete at the expense of compactness. Therefore, in the practical system an alternative technique has been invented, by which several specific formulas can be synthesized into a single formula; for example, "α or β-chlorostyrene" (GN 3-34, SN 3-49) (Chart II) is expressed by a single formula. Similarly, "*p*-1 (or 3)-menthene" (GN 3-2247, SN 3-2648) (Chart III) is expressed by a single, though complicated, formula. Broadened coverage of substances in the S representation which results from the use of PCCG is restricted by a sophisticated combination with the Q and R representations.

Chart II. α- or β-Chlorostyrene (GN 3-34, SN 3-49)

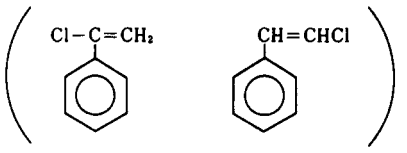
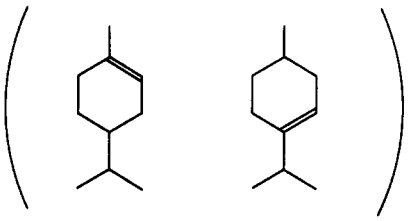


Chart III. *p*-1(or 3)-Menthene (GN 3-2274, SN 3-2648)



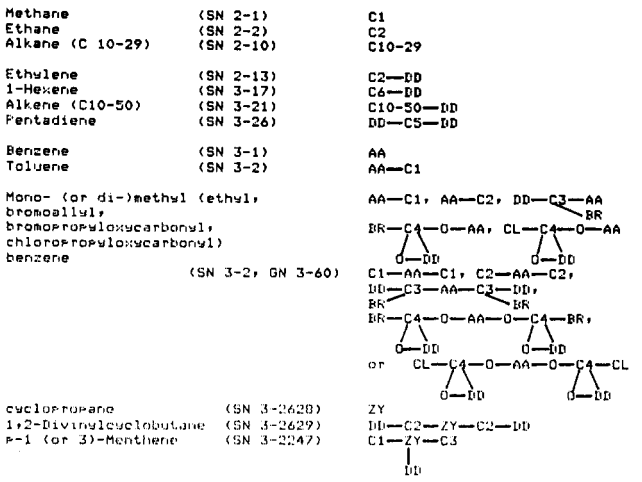


Figure 3. Examples of Q representations.

(5) Q Representation. The Q representation is a set of a kind of connection tables, and in making comparisons during the search it strictly requires an exact match. Therefore, while the S representation covers many structures simultaneously as explained for the example of dichloro-*m*-xylene, the Q representation rigidly restricts the framework which the structures allowed by the S representation can span. Whereas the S representation of dichloro-*m*-xylene (SN 3-23), though it rejects the ortho and paraisomers, covers even benzene (SN 3-1), it is only the positional isomers of dichloroxylene that the Q representation of dichloro-*m*-xylene defines; it rejects all others, including the other benzene derivatives which are chlorinated and methylated at nuclear positions. Thus, the Q representation of dichloro-*m*-xylene rejects 2,3-dichlorotoluene and 2-chloro-*m*-xylene.

Different degrees of specificity are possible in a qualitative description of structures. One could devise a texture in which multiple levels of qualitativeness may be handled by different representations. When more complicated objects such as patent specifications are to be treated, such a texture may become necessary. However, in the present system, we use only one kind of grammar in the Q representation, and in this sense the multiplicity of Q representation itself is single. The Q list has its components which include "ZZ" for a heterocyclic ring (without differentiation among various heterocyclic ring systems), "ZY" for a carbocyclic ring other than the benzene ring, "AA" for a benzene ring, and other symbols for acyclic carbon atom groups, acyclic heteroatoms, and/or a metal atom. Other components are "DD" for a double bond and "TT" for a triple bond. Some kinds of the components form a collection of components, such as Br2 and DD2, if they are attached on the same component. Namely, the connection

Table VII. Attributes of a Ring System in the R Representation

the number of ring members (10 for quinoline)
the number of all possible circuits (containing enveloping rings; 3 for quinoline)
the number of heteroatoms (1 for quinoline)
whether P, N, S, O, As, Sn, and/or any other heteratom is contained or not
the size of all possible circuits (6, 6, 10 for quinoline)

table of the Q representation is a kind of component connection table.[6] Figure 3 shows some examples of "structural formulas" at the level of the Q representation. It should be noted that a set of ten Q representations is necessary for GN 3–60 because of a variety of structures refered to by its name.

(6) R Representation. All connectivities of a specific formula or a group of specific formulas are described in the S representation. The framework of the formula(s) is restricted in the Q representation in which each ring system is treated as if it were an atomic group (ZZ, ZY, or AA) as explained in the preceding subsection. This means that all the necessary information to describe any organic structure is included in the two representations. However, for the sake of efficient searching, some important attributes of ring systems are recorded in the R representation. They are shown in Table VII.

In certain cases, a combination of the Q and R representations provides the necessary and sufficient information on ring systems, and the S representation becomes redundant.

(7) Combination of the Representations. As shown in Figure 3, the Q representation of toluene (GN 3–2) is AA–C1, and the compound can be fully described even without using the S representation. There are many cases in which the combination of the Q and R representations can completely describe the compound(s) that corresponds to a name in the Handbook. Namely, either the Q or R representation of the combination or the S representation only is redundant and can be discarded in such cases to reduce the size of the data base without loss of the system capability. In practice, the S representation is eliminated when it is entirely redundant to improve the search speed. The size of the S list is set equal to zero, indicating that the matching step of the S representations will be skipped.

For searching patent documents, however, such a redundant S representation may be useful to meet the requirement of the detailed substructure searching.

There are cases, of course, in which a name represents so many substances of different types that the single set of the three representations is still insufficient to describe it. In such cases, more than one set must be used. "Mono- (or di-) methyl (ethyl, bromoallyl, bromopropylooxycarbonyl, or chloropropyloxycarbonyl) benzene" (SN 3–2, GN 3–60; see Table I and Figure 3) is an example.

There are other cases where S representation is omitted in building the data base. An example is 1,1-dialkyl (C4–20) vinylidene (GN 2–29, SN 2–22) for which the S representation will be added. In such a case, only the Q representation is used at the expense of relevancy to a small extent. Another example arises in the case of branched alkylbenzene (C3–34) (GN 3–22, SN 3–4) for which a large number of S representations are needed because a complete colored graph would also include normal alkyl chain substituents which are difficult to exclude.

THE DATA BASE

As mentioned above, the data base consists of a notation file and a name file.

The notation file stores sets of G numbers and the corresponding representations for names. The representations are searched, and the G numbers are retrieved. On the other hand, the name file stores sets of G numbers and corresponding

Table VIII. Constitution of the Notation File

byte	range	content
1	8	G number
9	10	reserved
11		the size of the Q list
12		the size of the Q connection table
13	14	the number of ring systems (e.g., two for 1-phenylnaphthalene)
15		the size of the S list
16		the size of the S connection table
17		the Q list, the Q connection table, representation (2 bytes/ring system), the S list, and the S connection table

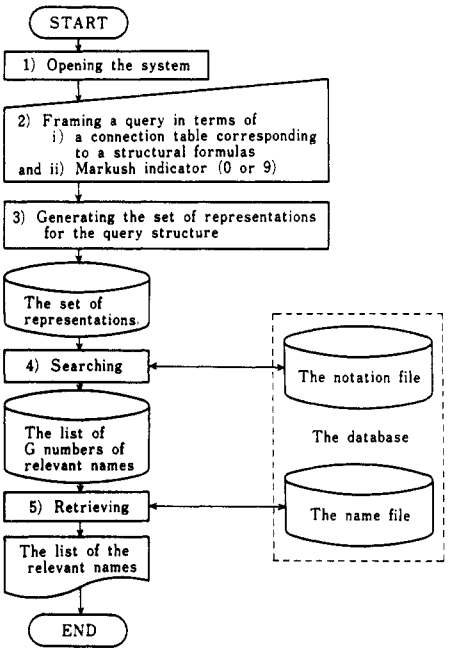


Figure 4. Searching procedure through the system.

names. G numbers are searched and the names are retrieved whose G numbers matched.

Although the system covering all the 20000 names has been designed to have the notation file consisting of three separated subfiles for the three types of representations, in the prototype system a set of three representations is packed in a 512-byte record, and the notation file consists of these records. The search is performed sequentially record by record. The first eight bytes are used for the G number; i.e., GN 2–1 is, for example, written as 20001 in the record. The ninth and tenth bytes are saved for future use. The next (the eleventh) byte is for the size of the Q list (the number of its components), and the twelfth is for the number of bonds in the Q connection table. Including these, all the elements have the allocated memory space as shown in Table VIII.

The data elements of the name file are G numbers, and the names are replicated from the Handbook. For one name, a 512-byte record is assigned.

SEARCH PROCEDURE

A general flow of the search is illustrated in Figure 4 and explained as follows. (1) The system, after initialization, sends some prompting messages step by step. (2) According to prompting by the messages, the user may describe a connection table representing a structural formula and set the Markush indicator to be either specific or generic. (3) The query is reformulated as a set of representations. If the indicator is generic, a parameter is set to skip the steps of searching the S representation. (4) The notation file is searched for the same set of representations as the query. Whenever a set of rep-

Table IX. Sizes of the S Lists for 1-Bromo-2-chlorobenzene and Bromochlorobenzene as Queries Using the Same Structure

Markush indicator	Q representation	size of the S list	components of S list
specific	BR-AA-CL	8	one bromine, one chlorine, and six carbon atoms
generic	BR-AA-CL	0	none

Table X. Other Examples of the Conversion in the System

specific formula	result of conversion
butane	C4
2-methylpropane	C4
benzene	benzene
cyclohexane	carbocyclic ring system (6)
naphthalene	carbocyclic ring system (6, 6, 10)
azulene	carbocyclic ring system (5, 7, 10)
furane	heterocyclic ring system (O5)
pyridine	heterocyclic ring system (N6)
quinoline	heterocyclic ring system (6, N6, N10)
sodium	metal
calcium	metal
1-butene	C4 with a double bond
2-butene	C4 with a double bond
1,3-butadiene	C4 with two double bonds

resentations which match the query is found, its G number is stored. They are the G numbers of relevant names. (5) In order to retrieve the relevant names, the name file is searched for the G numbers.

A STRUCTURAL FORMULA AS A QUERY

A query always consists of an appropriate specific structural formula and a value of the Markush indicator. The structural formula is described in terms of atoms and connectivities between them in the same way as in the data base by using the bond order parameter set shown in Table IV.

Examples of input structural formulas are shown in Figures 1 and 5, Figure 5 being for the case of 1-bromo-2-chlorobenzene. The bond order parameters of the six bonds in a benzene ring is eight because they are cyclic special bonds. When it is entered as a query, there are two situations, depending on the setting of the Markush indicator. If it is set as specific, the three representations will be generated, whereas if it is set as generic, the size of the S list is set equal to zero, and the S connection table will not be generated (Table IX).

Table X shows other examples of the conversion of the query to Q representations that are performed in the system.

OUTPUT OF THE SYSTEM

Examples of output results are shown in Figures 1 and 5. Relevant names are displayed together with their G numbers. If the GN–SN concordance table were also used, the system would display the corresponding S numbers. In the same way, these numbers can be connected with the Chemical Abstracts Service Registry Numbers.

EXPERIMENTAL RESULTS

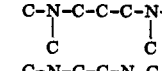
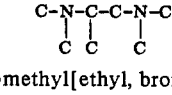
Some test runs of the pilot system containing about 300 names yielded the results shown in Tables XI–XIV, which will illustrate some features of functions of the system. In these cases, the recalls are all perfect.

Table XI gives typical search results for various compounds. The example of butanes shows that when a user sets the Markush indicator to "specific", he will retrieve the same answer regardless of which isomer he puts in. The "no hit" designation in the tables shows that the query compound is not covered by the Handbook, which is a very significant information indicating that the compound is not regulated by the law.

When the query is toluene or xylenes, the long name of GN 3–60 (SN 3–2), "mono-(or di)methyl (ethyl, bromoallyl,bromopropyloxycarbonyl, or chloropropyloxycarbonyl) benzene",

Table XI. Retrieval of Acyclic Saturated Hydrocarbons

		buildup of the query		retrieved				
no.	query	input structure	Markush indicator	SN	GN	names	recall	relevancy
1	*n*-C4H10	C-C-C-C	s	2-4	2-4	butane	100	100
2	C4H10	C-C-C-C	g	2-4	2-4	butane	100	100
3	IsoC4H10		s	2-4	2-4	butane	100	100
4	C4H10		g	2-4	2-4	butane	100	100
5	*n*-C10H22	*n*-C10H22	s	2-10	2-10	alkane (C10-29)	100	100
6	C10H22	*n*-C10H22	g	2-10	2-10	alkane (C10-29)	100	100
7	*n*-C29H60	*n*-C29H60	s	2-10	2-10	alkane (C10-29)	100	100
8	C31H64	*n*-C31H64	g	no hit				
9	toluene	toluene	g	3-2	3-2	toluene		
				3-2	3-60	*a*	100	100
10	xylene	*o*-xylene	g	3-2	3-3	xylene		
				3-2	3-60	*a*	100	100
11	*o*-xylene	*o*-xylene	s	3-2	3-3	xylene		
				3-2	3-60	*a*	100	100
12	dibromobenzene	dibromobenzene	g	3-31	3-52	p-dibromobenzene	100	100
13	p-dibromobenzene	p-dibromobenzene	s	3-31	3-52	p-dibromobenzene	100	100

no.	query	GN	SN	name given in the Handbook
14	*n*-octylamine	2-133	2-147	monoalkyl (or alkenyl, C5-28) amine
		2-176	2-160	*N,N,N*-trialkyl (or alkenyl, at least one of the alkyl or alkenyl group is C8-24; others are H or Cl-5)
15		2-155	2-173	*N,N,N',N'*-tetramethyl-alkylene (C2-4) diamine
		2-2378	2-171	*N,N,N',N'*-tetraalkyl (C1-3 and C50-150) diaminopropane
16		2-2378	2-171	*N,N,N',N'*-tetraalkyl (C1-3 and C50-150) diaminopropane

[a] Mono(or di)methyl[ethyl, bromoallyl, ((bromopropyl)oxy)carbonyl, or ((chloropropyl)oxy)carbonyl]benzene.

Table XII. Retrieval of Propylbenzenes

no.	query	answer SN	answer GN	answer names (as given in the Handbook)	recall	relevancy
1	propylbenzene	3-3	3-21	*n*-alkylbenzene (C3-36)		
		3-4	3-22	branched alkylbenzene (C3-36)	100	100
2	isopropylbenzene	3-4	3-22	branched alkylbenzene (C3-36)	100	100
3	*n*-propylbenzene	3-3	3-21	*n*-alkylbenzene (C3-36)		
		3-4	3-22	branched alkylbenzene (C3-36)	100	50

Table XIII. Retrieval of Halogenated Alkanes

query:	chlorinated (0-2) and/or brominated (0-1) alkane (C1-3)				
answer: SN	GN	name (as given in the Handbook)	C	Cl	Br
2-1	2-1	methane	1	0	0
2-2	2-2	ethane	2	0	0
2-3	2-3	propane	3	0	0
2-40	2-35	chloromethane	1	1	0
2-41	2-36	methylene chloride	1	2	0
2-44	2-39	methyl bromide	1	0	1
2-48	2-59	bromochloromethane	1	1	1
2-53	2-53	ethyl chloride	2	1	0
2-54	2-54	dichloroethane	2	2	0
2-58	9-518	ethyl bromide	2	0	1
2-63	2-80	monochloropropane	3	1	0
2-64	9-576	1,3-dichloropropane	3	2	0
2-65	2-81	1,2-dichloropropane	3	2	0
2-67	2-73	1-bromopropane	3	0	1
2-68	2-76	isopropyl bromide	3	0	1
2-70	9-370	1-chloro-3-bromopropane	3	1	1
2-71	9-1247	1-bromo-3-chloropropane	3	1	1
2-72	9-2007	1-bromo-3-chloropropane	3	1	1

Table XIV. Retrieval of Alkane, Alkene, and/or Alkadiene

query:	alkane, alkene, or alkadiene (C5-6)	
answer: SN	GN	name (as in the Handbook)
2-5	2-5	pentane
2-6	2-6	hexane
2-16	2-19	pentene
2-17	2-22	1-hexene
2-23	2-30	ethylene oligomer (3-8 mer)
2-24	2-31	propylene oligomer (2-10 mer)
2-26	2-20	pentadiene

appears as an answer in the two cases irrespective of the requested query.

The example of *n*-octylamine illustrates the system capability of being able to handle such complicated entry as SN 2-176. "*N*,*N*,*N*′,*N*′-Tetramethyl-alkylene (C2-4) diamine" (SN 2-155, GN 2-173) is included in an answer for one query, 1,3-bis (dimethylamino) propane, but not for another query, 1,2-bis (dimethylamino) propane.

Table XII shows an example of the results in whcih the relevancy is not 100%. Thus, the branched alkylbenzene retrieved for the query, *n*-propylbenzene, is not relevant. This occurred because the S representation for the name "branched alkylbenzene (C3-36)" was not made for the reason described in section 7 above.

Table XIII illustrates the case of combination of variable numbers of different chemical elements.

Table XIV is an example of a variable number of double bonds in a molecule.

DISCUSSION

A Markush formula is a very convenient tool for describing a group of similar chemical substances, but it is not always convenient to be searched for. In order to solve this problem, this paper offers a possible and perhaps useful method to search a substance file or list containing Markush formulas as well as specific ones to look either for generic or for specific formulas by means of a connection table of a structural formula and Markush indicators. With the built-in Markush indicator mechanism, the user is freed from groping for uncertain image of generic formulas and from investigating the special notation system. The combination of different types of representations, which complement one another, would provide efficient searching of specific and generic names as shown in Table I. There are instances in which the Markush indicator may appear not to be always functioning because both a specific query and its corresponding generic query give the same answer. One reason that this occurs is that no structural isomers are possible in the case of the simplest structures, e.g., methane. The other reason comes from special nature of the names in the original Handbook, e.g.: (1) only generic names are found in the Handbook that correspond to the query structure as in the case of "butane"; (2) only one structural isomer is found in the Handbook that corresponds to the query structure, e.g., *p*-dibromobenzene.

The examples given in the Experimental Results section show that the recall is perfect in all cases although the relevancy was not 100% in one case, for which we have a good reason as explained earlier. The primary objective of the present system, therefore, has been achieved, i.e., that one can be assured as to whether a query compound or compounds are under regulation of the law. Even when one gets irrelevant

```
QUERY No. (0 FOR END) ? 3
Bond value (-1 for end) ? 1
 Atom (0-0 for end) ? 1
 Atom (0-0 for end) ? 2
 Atom (0-0 for end) ? 0
 Atom (0-0 for end) ? 7
 Atom (0-0 for end) ? 8
 Atom (0-0 for end) ? 0
 Atom (0-0 for end) ? 0
Bond value (-1 for end) ? 8
 Atom (0-0 for end) ? 2
 Atom (0-0 for end) ? 3
 Atom (0-0 for end) ? 4
 Atom (0-0 for end) ? 5
 Atom (0-0 for end) ? 6
 Atom (0-0 for end) ? 7
 Atom (0-0 for end) ? 2
 Atom (0-0 for end) ? 0
 Atom (0-0 for end) ? 0
Bond value (-1 for end) ? -1


The connection table.

 1  2 ( 1 )
 2  1 ( 1 ) 3 ( 8 ) 7 ( 8 )
 3  2 ( 8 ) 4 ( 8 )
 4  3 ( 8 ) 5 ( 8 )
 5  4 ( 8 ) 6 ( 8 )
 6  5 ( 8 ) 7 ( 8 )
 7  2 ( 8 ) 6 ( 8 ) 8 ( 1 )
 8  7 ( 1 )                (OK), no ? OK
Heteroatom (0 for END) ? BR
  No. (0 for End) ? 1
  No. (0 for End) ? 0
Heteroatom (0 for END) ? CL
  No. (0 for End) ? 8
  No. (0 for End) ? 0
Heteroatom (0 for END) ? 0
 1 2 3 4 5 6 7 8
BR . . . . . .CL    (Y)/N ? Y

Markush indicator  0 (specific) or 9 (generic).
      9 (Generic)  (Y)/N ? N

Input end.
The searching began.
Searched =  369    Found =  1
Retrieved names :
 1            30066          Chlorobromobenzene
END.
```

Figure 5. Retrieval of 1-bromo-2-chlorobenzene.

hits, their number is always less than five.

It is considered that the strategy developed here can be extended for other more general cases such as patent information by increasing the degree of multiplicity of representations and by incorporating multiple functions of the Markush indicator.

Although the tactics discussed in this paper would be sufficiently powerful even for other types of generic names in the Handbook which are not mentioned in the present paper, it is evident that there are many difficult problems to solve on a wide variety of Markush claims in patent specifications, including indefinite and a theoretically infinite number of structures. Therefore, before we can deal with the patent information in a completely satisfactory way, it would be necessary to give careful consideration at least on the following points: (1) more generic terms such as alkyl and halogen should be taken care of; (2) the Q representation should have a greater degreeof multiplicity; (3) the Markush indicator should be able to discriminate between more than two situations; (4) a more complicated query must be accepted by using logical operations (AND, OR, and so on).

EXPERIMENTAL SECTION

The searching program is written in BASIC PLUS and runs on a Disital Equipment Corp. PDP-11/60 at Japan Association for International Chemical Information (JAICI).

ACKNOWLEDGMENT

We are indebted to T. Oshima of Japan Chemical Industry Ecology-Toxicology & Information Center (JETOC) for discussions related to the Handbook. This work was financially supported in part by a Grant-in-aid for Scientific Research from the Ministry of Education, Science, and Culture to which the authors' gratitude is due.

REFERENCES AND NOTES

(1) (a) Ash, J. E.; Hyde, E., Eds. "Chemical Information Systems"; Ellis Horwood Ltd.: Chichester, 1975. (b) Rush, James E. "Handling Chemical Structure Information". In "Annual Review of Information Science and Technology"; Williams, Martha E. Ed.; Knowledge Industry Publications: 1978; Vol. 13, Chapter 8. (c) Howe, Jeffrey W.; Milne, Margaret M.; Pennell, Ann F., Eds. "Retrieval of Medical Chemical Information". *ACS Symp. Ser.* **1978**, *No. 84*.
(2) "NEW CAS SERVICES". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 117.
(3) Lynch, M. F.; Barnard, J. M.; Welford, S. M. "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 1. Introduction and General Strategy". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 148-150. (b) Barnard, J. M.; Lynch, M. F.; Welford, S. M. "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 2. GENSAL, a Formal Language for the Description of Generic Chemical Structures". *Ibid.* **1981**, *21*, 151-161. (c) Welford, S. M.; Lynch, M. F.; Barnard, J. M. "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 3. Chemical Grammaers and Their Role in the Manipulation of Chemical Structures". *Ibid.* **1981**, *21*, 161-168. (d) Barnard, J. M.; Lynch, M. F.; Welford, S. M. "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 4. An Extended Connection Table Representation for Generic Structures". *Ibid.* **1982**, *22*, 160-164.
(4) The Japanese edition has a name index which helps reduce the time if the substance in question is listed as a specific name.
(5) Chemical Products Safety Division, Basic Industries Bureau, Ministry of International Trade & Industry. "Handbook of Existing Chemical Substances", 2nd Ed.; The Chemical Daily Co. Ltd.: Tokyo, Japan, 1981.
(6) (a) Kudo, Yoshihiro; Yamasaki, Tohru; Sasaki, Shin-ichi. "The Characteristic Polynomial Uniquely Represents the Topology of a Molecule". *J. Chem. Doc.* **1973**, *13*, 224-227. (b) Kudo, Yoshihiro; Sasaki, Shin-ichi. "The Connectivity Stack, A New Format for Representation of Organic Chemical Structures". *Ibid.* **1974**, *14*, 200. (c) Kudo, Yoshihiro; Sasaki, Shin-ichi. "Principle for Exaustive Enumeration of Unique Structures Constituent with Structural Information". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 43. (d) Kudo, Yoshihiro; Aoki, Shotaro; Takada, Yoshito; Taji, Toyoaki; Fujioka, Ichiro; Higashino, Kazuko; Fujishima, Hisayuki; Sasaki, Shin-ichi. "A Structural Isomers Enumeration and Display System (SIEDS)". *Ibid.* **1976**, *16*, 50. (e) Sasaki, Shin-ichi; Abe, Hidethugu; Hirota, Yuji; Kudo, Yoshihiro; Ochiai, Shukichi; Saito, Keiji; Yamasaki, Tohru. "CHEMICS-F: A Computer Program System for Structure Elucidation of Organic Compounds". *Ibid.* **1978**, *18*, 211.

Chemical Inference. 1. Formalization of the Language of Organic Chemistry: Generic Structural Formulas

JOHN E. GORDON*[1] and JOYCE C. BROCKWELL

Chemical Abstracts Service, Columbus, Ohio 43210, and Department of Chemistry, Kent State University, Kent, Ohio 44242

Received February 22, 1983

Categorization, syntax, semantics, and history of generic structural formulas (GSFs) are discussed. Their roles in chemical inference, chemical documentation, and chemistry learning are considered in the context of normalization and formalization of languages of structural formulas, chemical equations, and mechanisms. A formal language (ABSF) of homocomposite GSFs and a heterocomposite language employing normalized structural variables (NVSF) are defined and merged. Useful formal operations involving these languages, their expressive power, and their relationship to Markush SFs and the GENOA and GENSAL languages are considered.

Generic structural formulas, i.e., those that denote whole classes of specific structural formulas (SFs), are heavily used in all domains of the chemical literature. Despite this usage and despite applications in other parts of chemistry to be discussed below, they have not been the subject of a general and fundamental linguistic study, with two partial exceptions: Study of the information-handling aspects of Markush SFs has been motivated by their extensive use in patents.[2-4] Lynch and his students have recently begun to describe a more fundamental treatment of generic structural formulas of variable composition.[5] Our work, which concentrates on fixed-composition generic structures, complements Lynch's rather well. This report describes our initial study and attempts to place it in a broad context of the history of SF-class notation and its applications in chemistry and metachemistry beyond chemical information science.

IMPORTANCE OF GENERAL STRUCTURAL FORMULAS

Maturation of a science is generally accompanied by formalization of its languages. Chemistry is currently in this stage; its linguistic development is proceeding in two directions.

0095-2338/83/1623-0117$01.50/0