The Chemical Abstracts Service Generic Chemical (Markush) Structure Storage and Retrieval Capability. 1. Basic Concepts

WILLIAM FISANICK

Chemical Abstracts Service, P.O. Box 3012, Columbus, Ohio 43210

Received October 23, 1989

The concepts and strategies presented in this paper provide a new way of viewing and representing chemical structures for both specific and generic substances. Groups such as ethyl and phenyl have generic group nodes in the connection tables as alternatives to their specific atom designations. A set of 11 generic group nodes is used. The generic group nodes are hierarchical and can be further qualified by attributes such as the carbon range and ring size. Colloquial generic groups such as alkyl and aryl are likewise mapped onto the hierarchical generic group nodes and appropriate attributes. The fragments of variable groups in the input representation are attached to the appropriate portion of the structure as simple alternatives at a designated point of variability. Points of variability are sometimes algorithmically shifted in the structure to simplify further processing. The resulting topological representation enables searching of specific or generic substances with structure queries consisting of all generic groups, all specific atom groups, or any combination of specific atom and generic groups, and with or without expressed variability.

I. INTRODUCTION

During the past several years, there has been a considerable amount of activity in the research and development of computer systems for the storage and retrieval of topological representations of generic chemical substances (i.e., structures that imply more than one specific substance). Of special interest to researchers has been the handling of generic structures in patents. These are often called Markush structures after Eugene Markush, an American scientist who brought a court case in 1925 that led to the acceptance of such structures in patent claims by the U.S. Patent Office.

Approaches reported for the handling of generic structures are largely extensions of the topological approaches used to handle specific substances. Many of these approaches were discussed at a conference held in 1984 in Sheffield, England,[1] and in a symposium at the American Chemical Society (ACS) meeting in Anaheim, CA, in 1986. The presentations at the latter meeting were summarized in a paper by Barnard.[2]

The first system to provide for topological search and retrieval of generic structures was developed by Ernst Meyer and his colleagues in the 1960s.[3] Although restricted in the type of generic substances that can be handled, this system has demonstrated the feasibility of topologically based generic structure search and retrieval. More recently, the research efforts of M. F. Lynch and his colleagues at the University of Sheffield[4-14] have provided a foundation for other researchers in this area. GENSAL, a formal language developed by the Sheffield group for representing generic (Markush) structures, is also currently being tested and evaluated for use by others.[15] Kudo and Chihara have also proposed a method for handling generic chemical structures.[16] In the past few years, a collaborative effort in this area has been made by a group consisting of Derwent Publications Ltd., a company that indexes and abstracts patents; Telesystemes-Questel, a company that provides the DARC structure-handling software; and the French Patent Office (INPI).[17-19] This effort led to the MARKUSH DARC system that was recently made available. Finally, Tokizane et al. have recently reported a method to store and search chemical structure information with generic expressions.[20]

Chemical Abstracts Service (CAS) is also developing a computer-based storage and retrieval capability for generic chemical substances. CAS's interest in computer-based substance searching began in the late 1960s when we developed a batch substructure search system based on the structure or topology of a specific chemical substance as represented in a connection table.[21] During the early 1970s, efforts were focused on the development of techniques for searching the nomenclature representation of specific chemical substances.[22,23] In 1977, work began on topologically based techniques for online substructure searching of the specific substances in the CAS Chemical Registry; this led to the CAS ONLINE substructure search system, which is available today through STN International.[24-26]

In the early 1980s, research began on the use of topologically based techniques for the handling of generic chemical substances. This research led to the capability for formulating generic queries for searching specific substances on the CAS ONLINE Registry File on STN since 1985 and to research into storing and searching generic chemical substances, currently under development. Some of the design requirements and features of this latter capability were presented at the Sheffield conference and the ACS symposium mentioned previously.[27,28] Aspects of the overall design and search strategy are included in U.S. Patent 4 642 762.[29] The initial application of CAS's generic substance handling will be a search service on Markush structures from patents.

This paper presents the basic concepts of the CAS generic substance handling capability. Additional papers on the file substance description language and the search techniques and capabilities are planned.

II. NATURE OF SPECIFIC AND GENERIC SUBSTANCES

Chemists and other scientists represent chemical substances in the literature with varying degrees of specificity. Various conventions are used to represent structures of specific substances, but all involve the specification or implication of specific atoms (real atoms) and their connecting bonds. The atoms are those given in the periodic table, and the bonds are the traditional chemical bonds such as a single, double, or triple. In Figure 1, structure A represents the specific substance 2-chloro-4-propylpyridine. Such specific substances make up the CAS Chemical Registry System. Generic substances, on the other hand, connote more than one substance or a set of specific substances. This set of substances can be represented by a generic or variable structure. Structure B in Figure 1 is a generic structure that implies an open-ended

0095-2338/90/1630-0145$02.50/0

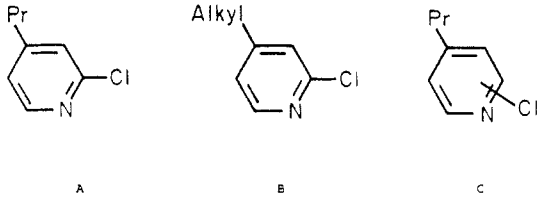

Figure 1. Specific (A) and generic (B, C) substance representations.

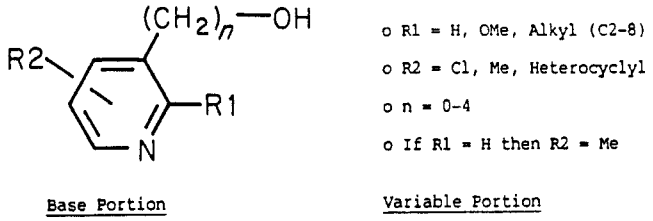

Figure 2. Example Markush formulation.

set of such specific substances as 2-chloro-4-methylpyridine, 2-chloro-4-ethylpyridine, 2-chloro-4-propylpyridine, and 2-chloro-4-isopropylpyridine. Structure C implies the closed set 2-chloro-4-propylpyridine and 3-chloro-4-propylpyridine (monochloro is assumed).

An important use of generic structures in the literature is for the compaction of the representation of a set of specific substances. This is especially true of Markush structures found in patent claims, where the entire set of the implied specific structures is considered to be covered by the claim. Another common use for such compaction is for expressing data for a set of substances. In such cases, the base structure (common substructure) is usually shown with variable substituents, and the data for the specific values of the substituents are given in tabular form.

Generic structures have also been used to express ambiguity about a specific substance. For example, structure C in Figure 1 can be used to express a situation where it has not been determined whether the specific substance is the 2-chloro- or the 3-chloro- derivative.

The most complex generic structures are probably the Markush structures found in patents. Such structures can represent very large sets of specific substances. Kaback has illustrated a structure that implies more specific substances than are in the entire CAS Registry File of more than 9 million substances![30] The conventions used to describe generic structures in patents are very extensive, although they lack standardization. Such Markush structures often are described by both a structural diagram component(s) and a textual component. An example of a Markush formulation is illustrated in Figure 2.

The structure diagram in Figure 2 is used to express the base portion of the generic substance. The text on the right side of the figure gives the definitions of the explicit variable groups and the inter- and intravariable group logic and conditions. A variable group is a set of structural fragment possibilities at a particular position(s) on the base portion of the structure or on a parent variable group. These fragments are described by structure diagrams, nomenclature or other descriptive terms, and/or line formulas. Three main types of variable groups can be distinguished: a fixed position of attachment substituent such as R1; a nonfixed or variable position of attachment substituent such as R2; and a repeating group such as the methylene repetition in the example.

The structural fragments represented in a Markush structure (base or variable group fragments) are either specific (real atom) groups such as the Cl, OMe, and pyridine or generic groups such as the alkyl and heterocyclyl. A generic group is an implicit variable group in that it connotes a class or set of real-atom groups. For example, alkyl implies fragments

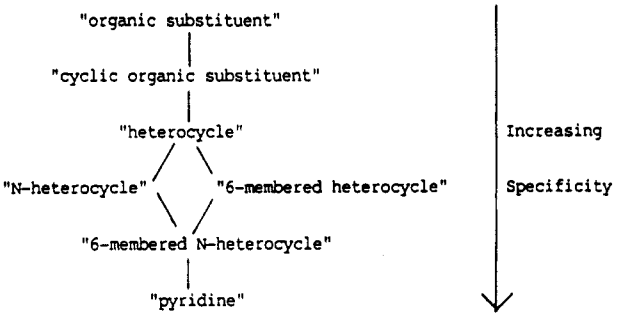

Figure 3. Generic group hierarchies.

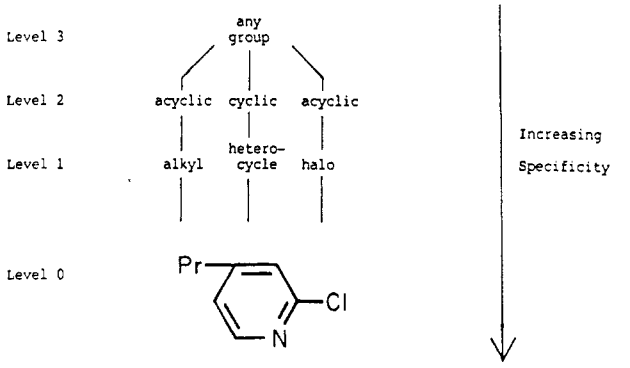

Figure 4. Hierarchical generic group relationship.

such as Me, Et, Pr, *i*-Pr, and Bu. The alkyl set is open-ended; in principle, there is no limit to the number of real-atom groups that may be constructed.

Generic groups may be qualified, as by "C2–8" restriction of two to eight carbons in the alkyl group in Figure 2. They are typically expressed in Markush structures as nomenclature terms.

The logic relationship among the fragments in the variable groups may be implicit or explicit. For example, R1 at the ortho position or 2-position of the pyridine ring system in Figure 2 may be a hydrogen atom, a methoxy atom group, or an alkyl group, limited to between two and eight carbon atoms. The comma implies the OR operator. Intervariable group logic is usually more explicit. For example, if R1 is a hydrogen atom, then the possibilities for R2 are limited or restricted to methyl. The words "if" and "then" express the logic.

The language used to describe generic groups in patents uses an uncontrolled vocabulary that expresses varying degrees of specificity about the structure of the group. These degrees of specificity can correspond to a hierarchical relationship. One such hierarchical relationship is illustrated in Figure 3. For example, a generic group may be defined as broadly as "organic substituent". More specific than this would be a group like "cyclic organic substituent", which implies a ring system. "Heterocycle" implies a ring system and at least one non-carbon node. For a "6-membered N-heterocycle", the heteroatom is a nitrogen and the ring system contains six nodes. The real-atom moiety, pyridine, denotes a specific molecular skeleton, including the bond types and values.

Both specific and generic substances can be viewed from a perspective of a hierarchy of generic groups, where such groups are expressed or implied. Figure 4 illustrates a four-level hierarchy for 2-chloro-4-propylpyridine (structure A in Figure 1). The specific groups (level 0), Pr, pyridinediyl, and Cl, are members of the alkyl, heterocycle, and halo generic groups (level 1), respectively. These groups are, in turn, members of the acyclic, cyclic, and acyclic groups (level 2), respectively. All of the level 2 groups are related to the very broad group "any group" (level 3).

Relative to this hierarchy, all groups for structure A in Figure 1 are at level 0; structure B has groups that are 1–0–0 with respect to the hierarchy and is more generic than structure A. Structure C in Figure 1 can be resolved into two alternative specific structures that are 0–0–0 with respect to the hierarchy. Other possible generic substances such as Pr–heterocycle–Cl (0–1–0), alkyl–heterocycle–halo (1–1–1), acyclic–acyclic–Cl (2–2–0), and Pr–any–Cl (0–3–0) are also related to structures A, B, and C and each other.

So, specific and generic substance representations of varying degrees of specificity can be related to each other on the basis of hierarchical generic groups, provided, of course, that such groups are appropriately defined.

III. GOALS AND REQUIREMENTS

The goals and requirements for the CAS generic substance handling capability are based, to a large extent, on user needs obtained from a survey conducted by CAS marketing staff and from discussions with experts on patent information. The goals and requirements have already been discussed in detail.[27] A summary of the main goals and requirements is given in the following paragraphs.

(A) User Interface. An important requirement for the system's input capabilities has been to use, as far as possible, the same structure language for input of both queries and generic file substances. This language is consistent with the STN query structure language used in searching STN structure-searchable files. This simplifies review of Markush structures retrieved from a search because they contain the same language as a query structure. Also, the compatibility with STN structure query-framing makes it easier for users to shift from specific to generic structure searching. Details on the query and generic file structure language will be given in a later paper.

(B) Query vs File Substance Types. For the retrieval of substance information in patents, users have indicated that they need to access information about both specific and generic substances. This has led to a requirement that a query structure be executable on both the specific structure files and the generic structure file, as far as possible. This one-time query-framing simplifies the user interaction.

Users have also indicated they would like to search with a range of query types against both specific and generic structure files. These query types can be distinguished on the basis of whether or not further substitution (open valences), variable groups, and/or generic groups are provided for in the query structure. Queries with open valences have traditionally been called substructure queries, and those with variable groups have been called Markush structure queries.

Structure queries can also be divided into specific and generic structure queries. Specific structure queries are those that do not have open valences, variable groups, or generic groups. The intended target for this query type is a single specific file structure or a generic file structure that inherently contains the specific structure. A generic structure query contains open valences, variable groups, and/or generic groups. The intended target of this query type is a set of specific file structures or a set of generic file structures that inherently contain one or more specific structure possibilities for the query.

Both specific and generic structure queries, including generic structure queries that are combinations of open/closed valences, variable groups, and generic groups, are provided for in searching against generic file structures. These query types are also currently being applied to the specific substances in the structure-searchable files on STN.

(C) Search Matching Criteria. The basic criteria for matching query structures against structures in the file are important to the design of a generic substance handling capability. The criteria for the query/file structures cases discussed above have been described before.[20,27] Tokizane et al. have further extended the possibilities to include "superstructure" matching, which involves the matching of a file substructure embedded in a query structure.[20]

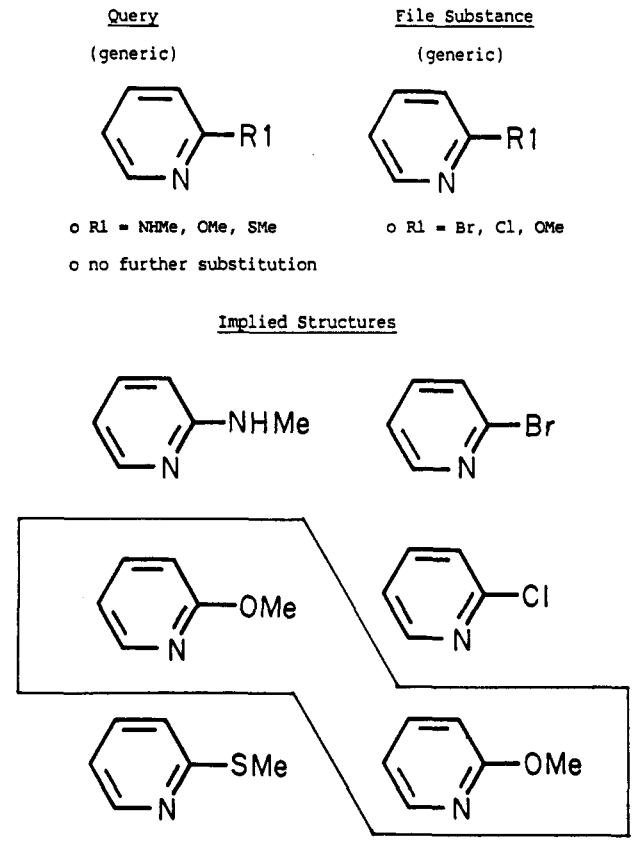

Figure 5. Overlap matching criterion.

Because a generic file structure is, in principle, a set of alternative specific structures, a criterion for specific structure overlap is needed as well as the more traditional embedment or containment criteria. This is illustrated in Figure 5 for the case of a generic query structure for a generic file structure, with retrieval occurring because the containment criterion is satisfied at the overlap of a pair of appropriate specific structures. With the overlap criterion, parts of an inherent query structure may not be located in different inherent file structures. In this case, exact matching of an inherent query structure is required and the structures at the overlap are identical; i.e., both are 2-methoxypyridine. Only one overlap pair of an inherent query structure and an inherent file structure needs to considered for retrieval to occur. Total overlap, or the equivalency of all structures in the two sets, might be an optional criterion, but this does not seem to be as important as the minimum overlap. The total overlap criterion could also be used for registration purposes, i.e., to verify the equivalency of two generic substances. However, for Markush structures in patents, which can imply very large sets of specific substances, such a total overlap condition would be very difficult to determine, and there are no plans to include such a criterion in the CAS generic substance handling capability, at least initially.

IV. OVERALL STRATEGY

A straightforward approach to handling generic substances would be to generate and store the sets of specific structure possibilities implied in each generic structure. The technology needed for storage, search, and retrieval would be essentially that which is currently being used to handle specific substances.

Query Group	File Substance Group	Match Group
1. "2-pyridinyl" Level 0	"heterocyclyl" Level 1	"heterocyclyl" Level 1
2. "any cycle" Level 2	"heterocycle" Level 1	"any cycle" Level 2
3. "alkyl" Level 1	"propyl" Level 0	"alkyl" Level 1
4. "any group" Level 3	"propyl" Level 0	"any group" Level 3

Figure 6. Information level normalization.

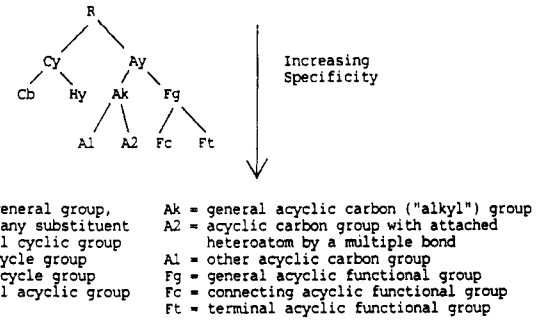

Figure 7. CAS generic group hierarchy.

However, such a strategy would not be feasible for a significant number of generic substances. The presence of a single generic group (e.g., aryl) in the substance can lead to an open-ended set of specific substances. Even when there is only real-atom variability, large sets of specific substances are still possible.

Other approaches to matching generic query structures to generic file structures require the resolution of the crux of the generic substance handling problem: the correct matching of different levels of structure information, that is, the correct matching of real-atom groups with generic groups (and vice versa) and generic groups with generic groups. For example, "heterocyclyl" in a query structure must match "2-pyridinyl" in the file structure (and vice versa); "heterocyclyl" in a query must match "any cyclic group" in the file structure (and vice versa). To allow for matching of different levels of information, the search algorithm must rectify the appropriate query and/or the file structure portions to the same information level so that a proper comparison can be made.

Our basic strategy for this rectification process is to perform the matching on the more generic level of the two levels being considered, if they are different. This is done by generating the target generic level of information for the more specific group in question so that a same level comparison can be made. Figure 6 illustrates the comparison of several sets of query and file substance groups with the information levels being those given for the groups shown in Figure 3. For the first example shown in Figure 6, the comparison between 2-pyridinyl and heterocyclyl is made on the basis of the information content of the heterocyclyl rather than 2-pyridinyl.

Matching on the more specific level of the two different levels being compared would be very difficult. For example, consider the matching of a query with a real-atom pyridine group and a file substance with a generic heterocycle group. To match at the real-atom level, the structures of all the ring systems implied by the term "heterocycle" would need to be generated, e.g., aziridine, pyrrole, furan, pyridine, pyrazine, and quinoline. These ring systems would be alternatives for the purposes of matching and, hence, pyridine would match the generated pyridine. However, such a set of ring systems is open-ended, and their generation is not very realistic. Even if this set were limited to all known heterocyclic systems, such as those known to the CAS Registry, it would be a large and unwieldy set.

Welford's TOPOGRAM program[6] generates real-atom fragments contained in several important generic groups. These fragments are used in a screening step in the searching of generic substances. However, due to the essentially open-ended vocabulary of generic groups, especially as found in the patent literature, the development of schemes to handle all types of generic groups would be a time-consuming and difficult task. If all the generic groups cannot be handled, then the capability becomes dependent on the nature of the queries, e.g., the system correctly handles "alkyl" but not "heterocycle".

The mapping of more specific information to more generic information is much easier to accomplish on a total recall basis. However, in matching on the more generic level of two groups, precision may be lost in the specific to generic mapping for a group(s). To improve the precision, CAS's strategy is to incorporate assigned sublevels, in addition to the main levels, in the matching. These sublevels correspond to qualifications or attributes of the generic groups. One example of a generic group attribute is the carbon range of a group such as alkyl.

This basic strategy for solving the crux of the generic substance handling problem requires several supporting concepts that are described in the following section.

V. BASIC CONCEPTS

(A) Generic Group Hierarchy. All queries and file substances in the CAS generic substance handling capability are viewed and manipulated in terms of a controlled vocabulary of generic groups that is hierarchical in nature. This hierarchy is illustrated in Figure 7. The hierarchy categorizes chemical structures or portions of structure into fairly broad classes. The most general group, R, corresponds to any structural fragment. R also corresponds to more than one of the groups below it on the hierarchy. At the next, more specific, level, a fragment is classified as acyclic (Ay) or cyclic (Cy). Note the mutual exclusivity at this level and the remaining levels, so that a fragment with both an acyclic portion and a cyclic portion would be split into an acyclic fragment and a cyclic fragment. A cyclic fragment is either a carbocycle (Cb) or a heterocycle (Hy). An acyclic fragment is either an alkyl carbon chain (Ak) or a functional group (Fg), which is defined as a heteroatom and its attached hydrogens. This definition of a functional group is a very simple, albeit fairly generic, one due to the need for easy recognition by a user or a computer algorithm. Functional groups are either connecting (Fc) or terminal (Ft).

Carbon chains (Ak) are divided, somewhat arbitrarily for the purpose of mutual exclusivity, into chains that are bonded to a heteroatom by a multiple bond (A2) and into other chains (A1) that are bonded only to heteroatoms by a single bond or are not bonded to a heteroatom. Thus, A2 includes the acyl family of fragments; Al the alkyl, alkenyl, and alkynyl families.

The main features of this generic group hierarchy are the mutual exclusivity at each level, the ability to completely describe, albeit generically, any substance, and the fairly simple definitions of the groups. The set of hierarchical generic groups (HGG) has a number of important uses as part of the generic substance handling capability:

- They are the controlled vocabulary to which all of the colloquial generic group vocabulary is mapped.
- They are the basis for a new, fully generic structure representation that is superimposed on all query structures, specific file structures, or generic file structures.
- They are the boundaries upon which query and file substances are manipulated, as in the rectification of explicit variability.

Alkyl	117		
		1-3 C	11
		1-4 C	31
		2-4 C	1
		2-5 C	1
		1-6 C	15
		2-6 C	3
		3-6 C	5
		5-7 C	1
		1-8 C	3
		3-8 C	2
		1-9 C	2
		"lower"*	13
		"higher"**	22
		no range	7
			117
		(*) Includes 1-10 C	
		(**) Upper end of range > 10	
Alkenyl	12		
Aralkyl	10		
Alkoxy	12		
Other***	60		

(***) Includes such items as aryl, acyl, and heterocycle

Figure 8. Generic groups in 79 Markush patent claims.

- They are the main levels upon which groups of different information levels are normalized for matching purposes.
- They are involved in screen fragments and screen searching, including a new screening technique—screen searching is an initial search step that eliminates many irrelevant file structures from further searching via the matching of a set of characteristic structural fragments called screens.
- They function as connection table nodes and are used in atom-by-atom (iterative) searching.

The assignment of HGGs for colloquial generic groups such as those described for Markush structures in patents can be illustrated for the groups given in Figure 3. "Organic substituent" is an R, and "cyclic organic substituent" is Cy. All of the heterocycle groups (heterocycle, N-heterocycle, 6-membered heterocycle, 6-membered N-heterocycle) are Hy with the distinction among them being made at the attribute sublevels. Pyridine is a real-atom group, not a generic group in itself. However, when it is matched with other heterocycle groups such as those given in Figure 3, it would have an alternative description as an Hy with appropriate attributes.

The availability of some fairly generic HGGs, especially the catch-all R group, allows for an HGG corresponding to any colloquial generic group definition, although the most precise HGG should be used to avoid search precision failures. However, caution must be used in interpreting the colloquial expressions. For example, "a ring system containing a benzene" should be assigned to Cy because there is the possibility of a heterocycle. Similarly, "a substituent containing a benzene" should be assigned to R because the presence of acyclic parts in the substituent cannot be excluded.

Although the HGG definitions handle the various types of colloquial generic group definitions from Markush structures in patents, they were not designed solely to handle such groups. Ease of use and mutual exclusivity of the groups were important considerations in the design. Figure 8 gives a summary of the generic groups found in a sample of 79 Markush claims that was derived from a sample of 206 patents from eight countries.[27] On the basis of these results, there is a predominance of alkyl and related generic groups in the patent literature. For example, alkyl, alkenyl, aralkyl, and alkoxy accounted for 151 (71.6%) of the total 211 colloquial generic groups in the sample. Because of this skewing, the use of

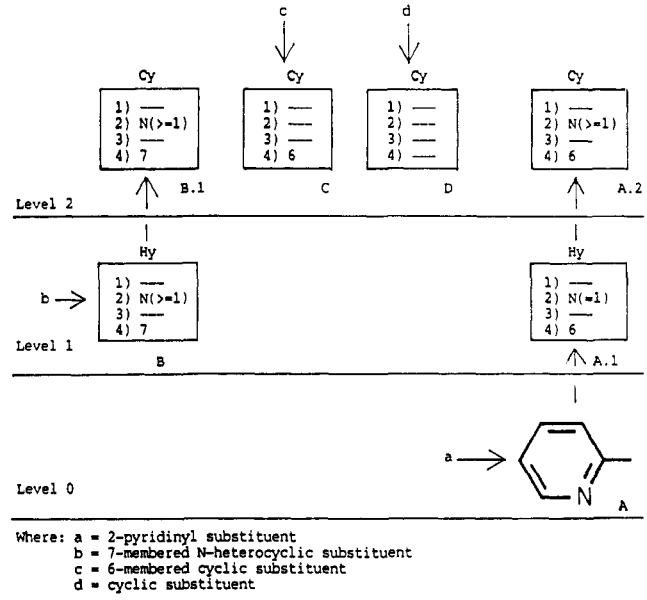

Figure 9. Matching of generic groups with attributes.

additional HGGs for the parent A1 was considered. A16, which is an A1 with six or fewer carbons, and A17, which is an A1 with seven or more carbons, were the prime candidates. A1 would be used for cases where no carbon range is specified. For example, in matching A16, A17, or a real-atom group such as *n*-butyl (⇒ A16) against A1, the A1 level would be used, i.e., the most generic level. Note that for the alkyl carbon ranges given in Figure 8, approximately half (67) have six or fewer carbons. Thus, A16 and A17 would seem to be a reasonable division of A1. However, special techniques are needed to handle the overlap ranges such as C(1–9). This additional complexity, plus certain aspects of screen generation and searching, led to a decision not to extend the hierarchy with A16 and A17, at least in initial versions of the capability. Distinction among carbon ranges is to be handled via the attribute sublevels of the A1 HGG.

Others have used or are considering using generic group definitions similar to the HGGs as part of their generic substance handling capability. The MARKUSH DARC system developed by Telesystemes uses a set of 21 generic groups called "superatoms" that also can be qualified by attributes.[17-19] Some examples of superatoms are CHK (alkyl, alkylene), ARY (carbocyclic aromatic), and UNK (unknown group). More recently, Tokizane et al. described the use of generic group nodes in conjunction with the expression and searching of structure attributes as bit-map vectors, but did not include a specific list of the nodes.[20] Lynch et al. have been experimenting with the representation and searching of specific and generic substances using generic group nodes of approximately the same specificity, e.g., R for ring, N for nonring, RC for a solely carbon ring, and RZ for a ring containing one or more heteroatoms.[14]

The various functions of the generic group hierarchy will be described more fully in later sections of this paper and in subsequent papers in this series.

(B) Generic Group Attributes. The sublevels or attributes of the HGGs are meant to improve the matching precision relative to the use of only the HGGs. Figure 9 illustrates the matching of the real-atom group, 2-pyridinyl (a), and the colloquial generic groups "7-membered N-heterocyclic substituent" (b), "6-membered cyclic substituent" (c), and "cyclic substituent" (c). Assume the set of attributes includes the kind and number of heteroatoms and the ring size (attribute 2 and 4 in Figure 9). A direct comparison between 2-pyridinyl (a) and "7-membered N-heterocyclic substituent" (b) cannot be made because they are on different levels of

specificity. Thus, the real-atom structure (A) of 2-pyridinyl is mapped to A.1 with the HGG Hy and the appropriate attributes and compared with the HGG version of (b), B, on level 1. If only the HGGs are considered in the matching, A.1 would match B because both are Hys. However, if the ring size attributes are also compared, there would be no match, i.e., six vs seven. Some other comparisons are as follows: (d) (via D) would match (a) (via A.2), (b) (via B.1), and (c) (via C). Note that in comparing an attribute, a specific value is considered to match a "no value". Also, to accomplish a comparison with (c), (a) and (b) need to be mapped to level 2; (c) (via C) would not match (b) (via B.1) because of the different values in the ring size attribute.

Within the framework of the generic group hierarchy and the definitions of the HGGs, a general-purpose set of attributes should reflect the qualifications of colloquial generic groups whenever possible. Such attributes should also be readily generated from corresponding real-atom groups and represent fairly general characteristics of the real-atom groups. The basic definitions set for CAS's initial generic substance handling capability include the following attributes (the possibility of using additional attributes is currently being investigated):

- attachment node (AN)—the number and kind of non-H atoms within the generic groups attached to nodes outside the generic group
- bonds (BD)—the number and type/value of the bonds within a generic group
- chain atoms (CA)—the number and type of acyclic atoms within a generic group
- charge (CH)—the collective charge on the atoms within a generic group
- degree of connectivity (DC)—the number and types of connectivities for the atoms within a generic group
- element count (EC)—the number and kinds of atoms, including H, within a generic group
- fusion atoms (FA)—the number and kind of fusion atoms within a ring generic group
- generic group (GG)—the number and kind of HGG groups to which the R group is restricted
- ring atoms (RA)—the number and kind of skeletal atoms within a ring generic group
- ring count (RC)—the number of rings within a ring generic group
- ring size (RS)—the size and number of each size of the ring within a ring generic group
- substructure (SS)—a real-atom substructure that further describes or qualifies a generic group
- text qualifier (TX)—a text statement that further describes or qualifies a generic group

Some of these attributes are restricted to certain generic groups. For example, those that deal with rings are restricted to ring generic groups. Many attributes have definitions similar to those of some of the general substructure search screens for the structure searchable file on STN, e.g., RC, DC, and EC. The use of attributes corresponding to more specific substructure screen classes such as augmented atoms (AA) and atom sequences (AS) did not seem to be appropriate, because colloquial generic groups are seldom qualified by something as specific as an augmented atom or an atom sequence.

The carbon range is the most prevalent qualification used for generic groups in the patent literature, especially for alkyl and related groups (see Figure 8). This is handled through the EC attribute with the range of values being specified. The attribute-matching algorithm will allow for the appropriate range overlap in the comparison. For example, consider the comparison of the real-atom group, *n*-butyl, with the unqualified generic group, alkyl, and the qualified generic groups, alkyl (C5–10) and alkyl (C2–6). All of the groups would be mapped to the A1 HGG, and the EC attributes would be 4 C, (5–10) C, and (2–6) C, respectively. The 4 C is included in (2–6) C but not (5–10) C, so a false drop is eliminated. (5–10) C is a match with (2–6) C because they overlap at five and six carbons. A minimum overlap criteria is used to determine a range match; i.e., only one member of each range set needs to be in common for a match to occur, all other things being equal.

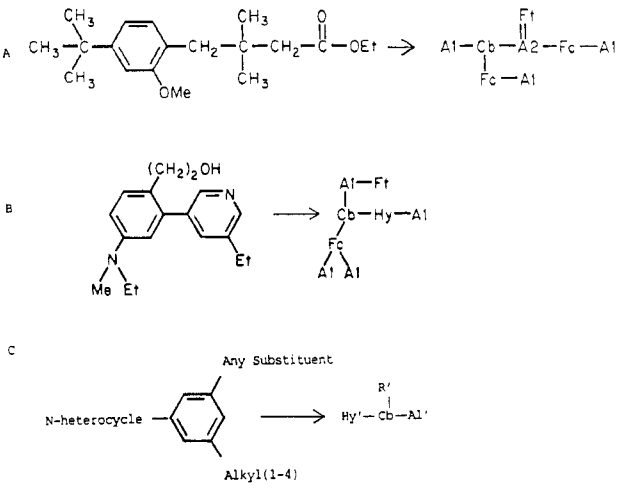

Figure 10. Mapping of specific and generic structures to generic (HGG) structure.

The SS, TX, and GG attributes are intended for use primarily with the R HGG. The SS attribute conveys structural information in an unbounded generic group such as for the thiophene substructure in "a substituent containing a thiophene ring". The TX attribute is used in cases where the structural definition is very vague and for which other attributes are difficult to generate, e.g., for a definition such as "electron-withdrawing group". TX attributes are not included in automatic matching; i.e., the matching is performed only at the HGG level. However, the TX data are to be free-text searchable, and its display in the retrievals can be used to help manually eliminate false drops. The GG attribute indicates the number or number range and the kind of more specific HGGs implied in R when they can be determined. This helps to increase the precision of R group matching.

The parameter lists of GENSAL[5] and the structure attributes described by Tokizane et al.[20] are similar to CAS's attribute set. Tokizane and his colleagues also generate attributes for real-atom groups for the purpose of matching against generic groups.

Examples of attribute specification and syntax will be given in a subsequent paper on the file substance description language; the searching of attributes will be discussed in a subsequent paper on search capabilities.

(C) Specific, Generic, and Composite Structures. The internal representation of structures in CAS's generic substance handling capability is based on several concepts.

As mentioned previously, all queries and file substances are described in terms of the generic group hierarchy. This hierarchy provides for an algorithmic mapping of a chemical structure, specific or generic, to a fully generic structure in which the nodes of the structure correspond to HGGs. Figure 10 illustrates the mapping for two specific structures and a generic structure without expressed variability. The various real-atom groups are mapped into the corresponding HGG node. Similarly, the colloquial generic groups for a file substance are mapped to the most precise corresponding HGG node, either through a term-to-HGG fragment dictionary or manually by the input specialist. The bonding between the

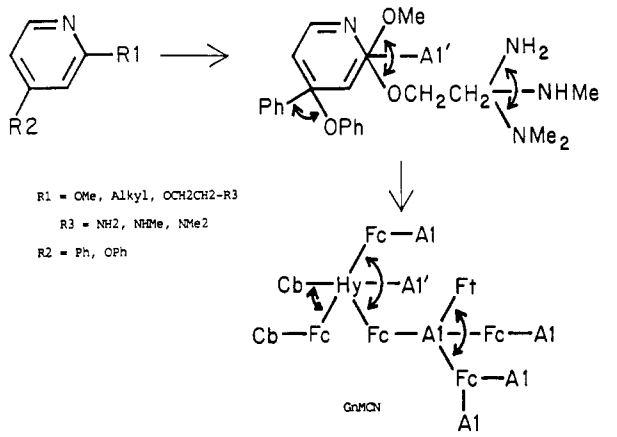

Figure 11. MCN (SpMCN and GnMCN) structures of a generic substance.

HGG nodes is the same as the bonding between the corresponding real-atom groups in the structure. As part of the HGG node generation process, the attributes for the HGG nodes are also generated. For example, for the "$(CH_3)_3C$–" group in structure A, attributes such as an EC of four carbons, a BD of three chain single bonds (non-hydrogen), and a DC of an atom with a degree of three are generated. In structure C, the colloquial generic group, "alkyl (1–4)" is mapped to A1 with the EC of 1–4 carbons attribute, "N-heterocycle" to Hy with the EC of one or more nitrogens, and "any substituent" to R without attributes (i.e., nondefault attributes). The nodes of the HGG structure are flagged with a "prime" if they arise from a colloquial generic group; i.e., they are distinguished from nodes that arise from a real-atom group (see structure C).

With respect to graph theory, the HGG structure is a reduced graph of the original graph representing a specific or generic substance. The HGG definitions are used for the reduction of the edges and vertices of the original graph.

The expressed variability that often occurs in generic substances is handled by a structure concept called a multiple connectivity node (MCN). In an MCN structure, the fragments of a variable group are attached to the base structure at the point of variability to form a fully connected graph of all possible structural fragments. The nodes to which the variable group fragments are attached are appropriately flagged. Such nodes are allowed to have connections that can exceed normal valences. There are two types of MCN structures: a specific one (SpMCN) that directly corresponds to the input generic structure and a generic one (GnMCN) that is derived from the SpMCN and corresponds to the fully generic HGG structure. Figure 11 illustrates an SpMCN and a GnMCN structure of a generic substance. The arc symbol is used to indicate the branches in the structure that belong to a particular variable group. In an MCN structure, the logic aspect of the variability is implicit: at each point of variability, the flagged branches (those within an arc) are simple alternatives. For example, in the SpMCN structure of Figure 11, the logic implication for the 4-position of the pyridine is that there is a Ph or an OPh substituent, but not both.

A complete set of fragment possibilities is always attached at a given point of variability, even though additional logic parameters may exclude some of the possibilities. For example, suppose the statement "If R1 = OMe Then R2 = Ph" is part of the specification of the generic substance given in Figure 11; i.e., the combination of R1 = OMe and R2 = OPh is not allowed. This narrowing of the scope of the substance is not reflected in the formal part of the connection table but rather in associated data elements. Thus, the same SpMCN and GnMCN structures in Figure 11 would be used for the structure with the additional conditional restriction.

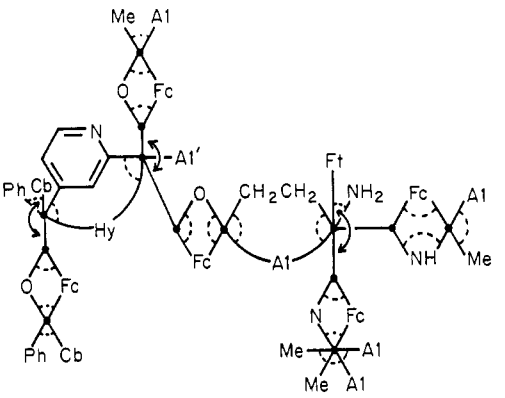

Figure 12. Internal representation of a composite (CpMCN) structure.

Similarly, when a specified number of fragments is distributed over several substituent positions, the complete set of fragments is still attached at each position. For example, suppose the specification for the generic structure in Figure 11 contained only the R1 and R2 variable groups and R1 = R2 = (1)Cl; i.e., a total of one Cl is substituted at one of the two possible positions. This would result in a SpMCN with two alternative fragments, Cl and H, at each of the R1 and R2 positions. Thus, MCN structures may have a considerable amount of redundancy with respect to storage requirements of variable group fragments—two chloro atoms instead of one for this last example. Schemes that use pointers to require the one-time storage of a particular group are possible. However, the tradeoff is that such redundancy tends to minimize the complexity of the search process and allows for the phasing in of search features such as the rectification of conditional logic. Ignoring conditional logic and occurrence counts of fragments in searching a MCN structure could lead to precision failures but not recall failures.

Another feature of the MCN structure is that the sets of alternative fragments are in logical nesting levels that can be envisaged as concentric circles emanating from the base portion of the structure. For example, in Figure 11 the R1 and R2 fragments are on the same level/circle, level 1, whereas R3 (a nest of an R1 fragment) is on the next level/circle, level 2. With appropriate node-specific flagging, the generation of the inherent specific structures and/or simple generic structures (i.e., where the input generic groups are passed through) can be achieved by a fairly simple "pruning" algorithm operating on the SpMCN structure.

While the tandem processing of the SpMCN and GnMCN structures of generic substances can meet the stated objectives of CAS's capability, the actual internal representation of generic substances is a composite of the two, i.e., a CpMCN. The CpMCN structure can be viewed as a superimposition of the GnMCN on the SpMCN or vice versa. The basic concept for the CpMCN is that the HGGs of the GnMCN are alternatives for the corresponding real-atom groups in the SpMCN, regardless of whether they are part of an original variable group or not. For example, the R2 group (Figure 11) can be considered to be a set of the following fragments: Ph, Cb, OPh, FcPh, OCb, and FcCb. Similarly, the Hy in the GnMCN is an alternative for the pyridine in the base portion of the SpMCN. Thus, in the CpMCN, all the real-atom groups belong to a variable group consisting of, as a minimum, the real-atom group and its HGG node counterpart. The only group that does not have a counterpart is a colloquial generic group. As mentioned previously, corresponding real-atom groups are not generated for such a group.

Figure 12 illustrates the CpMCN for the generic substance in Figure 11. The dots are path connectors (null nodes). Solid arcs indicate the fragments belonging to a particular variable group; they are shown at both ends for a connecting (i.e.,

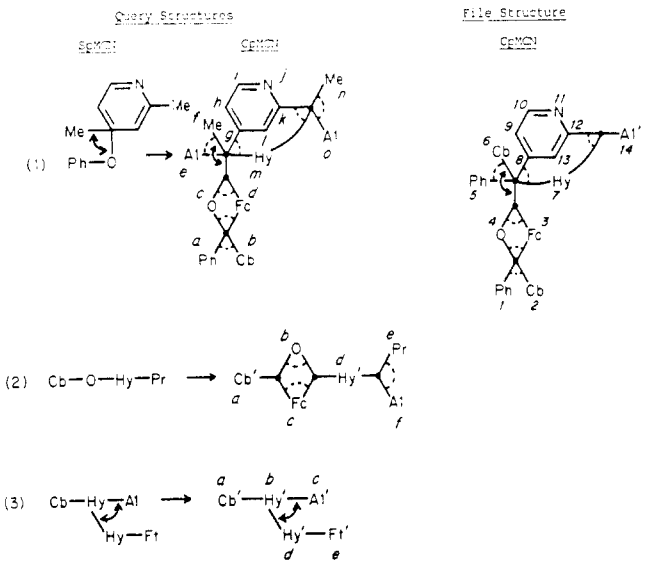

Figure 13. Query SpMCN/CpMCN structures vs file substance CpMCN structure.

nonterminal) group. Dashed arcs indicate additional points of variability created in building the CpMCN so as to avoid full replication of all possible fragments. For example, for R2, rather than storing FcPh, OCb, and FcCb as the alternative for the OPh fragment, the graph gives Fc and O as alternative nodes to which is attached Ph or Cb as alternatives. The CpMCN structure fully encodes all possible real-atom and HGG paths that are expressed or implied in the input generic substance, including all possible combinations of real-atom and HGG paths. The paths are delineated on the basis of HGG boundaries. Note that the A1′ has no real-atom counterpart, but is a member of a variable group since it is a fragment of such a group in the generic substance specification (Figure 11).

By use of the same basic algorithms, a CpMCN is also generated for query structures to allow for the correct matching against generic file substances. Similarly, a CpMCN can be generated for a specific substance. The only difference in a CpMCN for a specific substance and one for a generic substance is that the former cannot have points of expressed variability or primed HGG nodes. CpMCNs for specific substances are currently being used in searching the structure-searchable files on STN when the query structure contains a generic group. In such cases, the CpMCNs are generated "on-the-fly" for candidate substances for the final atom-by-atom matching.

The CpMCN structures for specific or generic substances can match specific or generic query structures, where appropriate. Figure 13 illustrates the matching of several query CpMCN structures against a file CpMCN structure. The file structure nodes are labeled with numbers and the query nodes with letters. The primed HGGs are input as part of the query or file structure; the unprimed HGGs are derived from a corresponding real-atom group that was input as part of the query or file structure.

Query structure 1 consists of all real-atom groups and has a single point of expressed variability (i.e., a methyl or phenoxy on the 4-position of the pyridine ring). This query structure matches the file structure via the pyridine ring, the phenoxy alternative in both the query and file structures, and the derived A1 for the 2-methyl in the query onto the A1′ in the file structure. A query to file node correspondence is $a \rightarrow 1$, $c \rightarrow 4$, $g \rightarrow 8$, $h \rightarrow 9$, $i \rightarrow 10$, $j \rightarrow 11$, $k \rightarrow 12$, $l \rightarrow 13$, and $o \rightarrow 14$. An attribute match would apply to the initial $o \rightarrow 14$ match (i.e., methyl vs alkyl) as a further qualification of the node match. For example, if the alkyl (A1′) had a carbon count of 4–10, the query "methyl" (carbon count of 1) would not match and the o vs 14 node would be a mismatch.

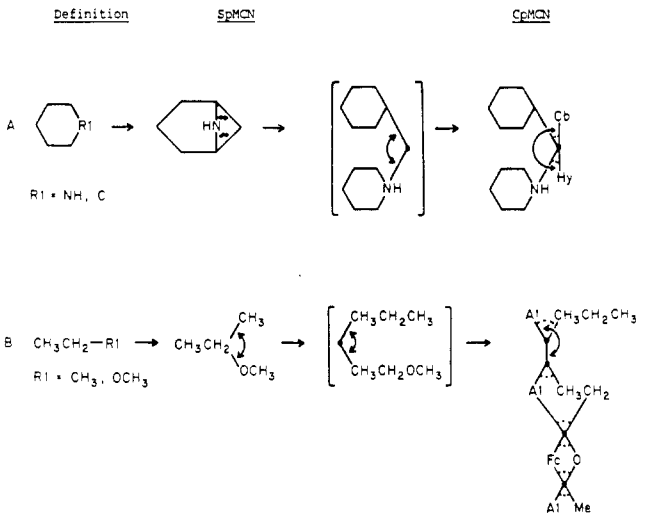

Figure 14. Structure rearrangements via point-of-variability shifts (cyclic and acyclic).

The second query structure (2) in Figure 13 is a combination of real-atoms and generic groups. This query structure can be located in the file structure with a query to file node correspondence of $a \rightarrow 2$, $b \rightarrow 4$, $d \rightarrow 7$, and $f \rightarrow 14$. The third query structure (3) consists entirely of generic groups and has a single point of expressed variability, i.e., A1′ or Hy′–Ft′ attached to a Hy′ (b). The query to file node correspondence is $a \rightarrow 6$, $b \rightarrow 7$, and $c \rightarrow 14$. However, as mentioned above, the node matching is qualified by attribute matching and, thus, the individual query and file nodes must also have an appropriate set of attributes.

The user may also restrict all or portions of the query structure to the matching against file structure real-atom groups. This will be explained in more detail in a subsequent paper on the search techniques and capabilities.

(D) Structure Manipulation. Colloquial representations of generic substances and queries often do not directly map to CpMCN structures in which there are discrete HGG boundaries. The generic substance handling capability contains a structure manipulation algorithm that shifts the points of variability, as appropriate, so as to lead to CpMCN structures that can be directly processed by the search algorithms. Figures 14 and 15 illustrate some of the types of rearrangements that are used in building the CpMCN structures for generic queries and substances.

In Figure 14, the A sequence illustrates a shift when a variable group fragment is embedded in a ring skeleton. Since the ring HGGs correspond to complete carbocyclic or heterocyclic ring systems, the point of variability is shifted to a null node (a dot connector in Figure 14) to allow for appropriate boundaries for the ring HGGs. As a result of such shifting, the search algorithm need not account for cases with a complex heterocycle–carbocyclic ring system. Note that shifting the point of variability does not change the set of implied specific substances in the generic substance; both the SpMCN and the CpMCN lead to the same implied specific substances, cyclohexane and piperidine. In the B sequence, if the shifting were not performed, there would be an A1–A1 (for CH_3CH_2–CH_3) fragment spanning the point of variability. The shifting establishes a new boundary with a single A1 group for the $CH_3CH_2CH_3$. Again, this simplifies the search algorithm in that it need not handle the complicated case where a HGG spans a point of variability.

Figure 15 illustrates a generic substance that has both acyclic and cyclic fragments for a variable group(s). The cyclic

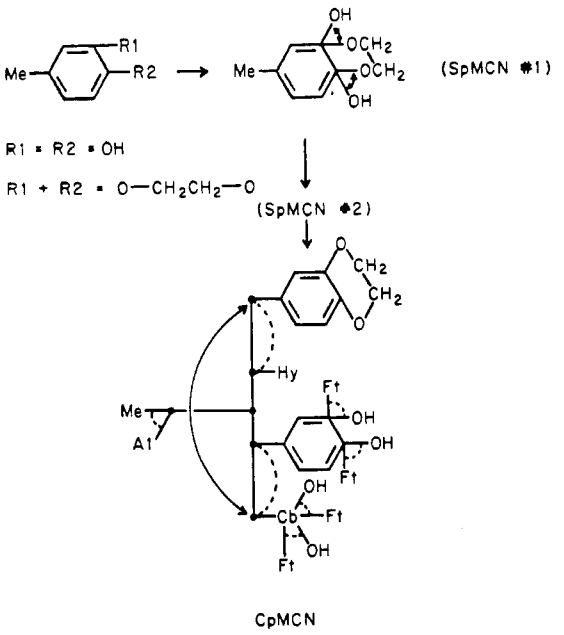

Figure 15. Structure rearrangements via point-of-variability shifts (acyclic–cyclic).

fragment forms a new ring system that corresponds to a different HGG (Hy) than the precursor ring system (Cb). The point of variability is shifted so that the different ring systems and their HGGs can be distinguished. Thus, the large arrow on the left of the CpMCN structure indicates that there are four alternative structural fragments which are attached to methyl and its HGG alternative, including the fused oxygen heterocycle and its HGG counterpart, Hy.

(E) Screen Toggling. The basic search techniques in the CAS generic substance handling capability are extensions of the two-step process being used for substructure searching of structure files on STN: a screen search of predetermined structural characteristics (screens) followed by a precision refining atom-by-atom (iterative) search on the candidate substances passing screen search.[26]

Generic substances, especially the Markush structures found in patents, can be very complex substances, often implying hundreds or thousands of specific substances. Thus, it is important to perform a screening function to eliminate most of the irrelevant answers before more precise, but more time-consuming, search techniques such as atom-by-atom searching are applied.

One of the difficulties in screening generic substances is that they can contain colloquial generic groups (mapped to primed HGGs); such groups have no alternative real-atom groups. This could severely limit the direct use of screens to file substances that do not contain any primed HGGs (only a small number of Markush structures from patents would not have colloquial generic groups). To help overcome this difficulty, CAS's generic substance capability uses a screen-toggling mechanism that enables the system to decide on an individual file substance basis when it is feasible to use the screens derived from real-atom groups in the query. Other aspects of our screens and screening technique are predicated on this mechanism.

Figure 16 illustrates the basic concept of the screen-toggling mechanism. The SpMCN structures, rather than the CpMCN structures, are shown for the sake of simplicity. Real-atom screens are used as sets corresponding to HGG boundaries. Thus, for a substructure query, there are two sets: the pyridinyl set (Hy HGG) and the ethyl set (A1 HGG). File substance A does not contain any primed HGG and, therefore, both the pyridinyl and ethyl real-atom screen sets of the query can be inclusion-matched against the screens of the file substance without a potential for a recall failure. However, for file substance B, which is a relevant answer, the pyridinyl screens cannot be used since there are no real-atom screens for the Hy′ in the file substance; i.e., a recall failure would occur if the query pyridinyl screen set were used. Similarly, for file substance C, the pyridinyl set can be used but not the ethyl set because an A1′ is present.

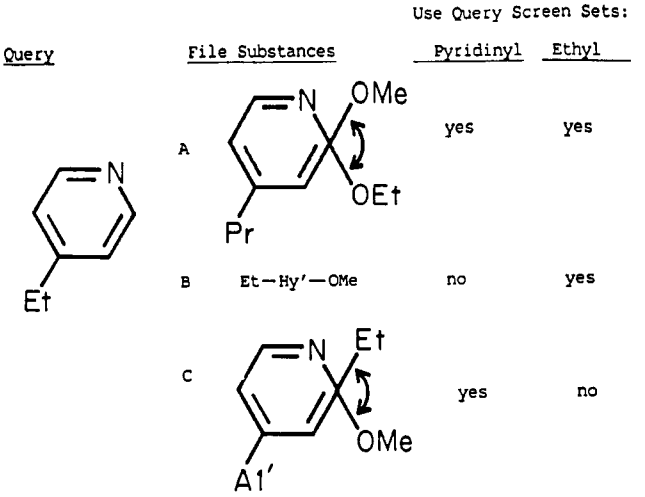

Figure 16. Real-atom screen toggling.

The toggling mechanism uses special switches or diagnostic screens for the file substances based on the primed HGGs and tests these switches with the query screen logic expression. For example, suppose the switch screens for the file substances in Figure 16 are Hy′ and A1′. For substance A, both of the switches would be off (not present); for B, Hy′ is on (present) and A1′ is off; and for C, Hy′ is off and A1′ is on. In the query screen logic expression, the switch screens are used as an alternative for the corresponding real-atom set, i.e., "(pyridinyl screen set OR Hy′) AND (ethyl screen set OR A1′)". This lets the system toggle the use of real-atom screens for the query on the basis of individual file substances. Thus, all three of the file substances in Figure 16 would correctly pass the screening and be candidates for more precise searching.

In practice, the switch screens involving HGG pair fragments are used to provide for greater precision in the screening. Additional details on the screens and search techniques of the generic substance handling capability will be given in a subsequent paper.

VI. DISCUSSION

CpMCN representations allow the desired matching criteria described earlier since they contain all the inherent substances that are indicated by the expressed variability in the original formulation. Also, at any given point of variability there is a simple set of alternatives. So, the atom-by-atom search, if properly embellished with procedures to process these alternatives, can check the entire set of inherent substances. The additional, inherent substances resulting from the implicit variability of colloquial generic groups need not be considered since the matching of such groups is not performed at the real-atom level. Because the matching criteria require only a correct match against just one of the inherent substances, the path tracing can be terminated after the first match is located.

A solution to the crux of the generic substance handling problem, the ability to correctly match real-atom groups against appropriate generic groups and generic groups against appropriate generic groups, is also implicit in the CpMCN definitions of query and file substances. For example, an ethyl in a query will match an A1′ in a file substance via the A1

alternative for ethyl in the query; an A1′ in a query will match an ethyl in a file substance via the A1 alternative for ethyl in the file substance.

The key strategy of matching a real-atom group against a generic group at the generic group's level of specificity, rather than attempting to generate the real-atom counterparts for the generic group for matching, is made more workable by the screen-toggling mechanism. Because there are no real-atom screens corresponding to an input generic group, the mechanism allows the real-atom screens for other portions of the query to be used selectively and so improves the screening precision.

The generic group hierarchy is a crucial part of generic substance handling, because all real-atom and colloquial generic groups are mapped onto it. While the hierarchy may need to be extended or redefined in the future, such features as the simplicity of the definitions and mutual exclusivity of the groups are believed to be required for generic substance handling.

VII. CONCLUSION

The concepts described in this paper were designed to support the searching of a file of Markush structures with queries that consist of either specific or generic structures. The CpMCN and/or its SpMCN and GnMCN counterparts allow searching of specific or generic substances with structure queries comprised completely of generic groups, queries containing only real atoms, or queries that are combinations of real-atom and generic groups, with or without expressed variability.

In many ways the basic concepts and strategies presented here provide a new way of viewing and representing chemical structures, in which the distinction between specific and generic substances is viewed as a continuum rather than having a discrete line of demarcation. These concepts may well provide the basis for other applications and tools that further enhance the user's ability to search chemical substance files and make use of the results.

REFERENCES AND NOTES

(1) Barnard, J. M., Ed. *Computer Handling of Generic Chemical Structures* (Proceedings of a Conference organized by the Chemical Structure Association at the University of Sheffield, England, March 26–29, 1984); Gower: Aldershot, U.K., 1984.

(2) Barnard, J. M. Online Graphical Searching of Markush Structures in Patents. *Database* **1987**, *10*, 27–34.

(3) Meyer, E.; Schilling, P.; Sens, E. Experiences with Input, Translation and Search in Files Containing Markush Formulae. In *Computer Handling of Generic Chemical Structures*; Barnard, J. M., Ed.; Gower: Aldershot, U.K., 1984; pp 83–95.

(4) 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.

(5) 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. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 151–161.

(6) Welford, S. M.; Lynch, M. F.; Barnard, J. M. 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.* **1981**, *21*, 161–168.

(7) 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 (ECTR) for Generic Structures. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 160–164.

(8) Welford, S. M.; Lynch, M. F.; Barnard, J. M. Towards Simplified Access to Chemical Structure Information in the Patent Literature. *J. Inf. Sci.* **1983**, *6*, 3–10.

(9) Welford, S. M.; Lynch, M. F.; Barnard, J. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 5. Algorithmic Generation of Fragment Descriptors for Generic Structures. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 57–66.

(10) Barnard, J. M.; Lynch, M. F.; Welford, S. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 6. An Interpreter Program for the Generic Structure Description Language GENSAL. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 66–71.

(11) von Scholley, A. A Relaxation Algorithm for Generic Structure Screening. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 235–241.

(12) Lynch, M. F.; Barnard, J. M.; Welford, S. M. Generic Structure Storage and Retrieval. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 264–270.

(13) Gillet, V. J.; Welford, S. M.; Lynch, M. F.; Willett, P.; Barnard, J. M.; Downs, G. M.; Manson, G.; Thompson, J. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 7. Parallel Simulation of a Relaxation Algorithm for Chemical Substructure Search. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 118–126.

(14) Gillet, V. J.; Downs, G. M.; Ling, A.; Lynch, M. F.; Venkataram, P.; Woods, J. V.; Dethlefsen, W. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 8. Reduced Chemical Graphs and Their Applications in Generic Chemical Structure Retrieval. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 126–137.

(15) Kolb, A. Generation of the GREMAS Code from Graphical Input of Generic Structures. Presented at the 192nd National Meeting of the American Chemical Society, Anaheim, CA, Sept 1986.

(16) Kudo, Y.; Chihara, H. Chemical Substance Retrieval System for Searching Generic Representations. 1. A Prototype System for the Gazetted List of Existing Chemical Substances of Japan. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 109–117.

(17) Bonnet, J.-C. Going to an Actual Chemical Patent Management System with DARC. In *Computer Handling of Generic Chemical Structures*; Barnard, J. M., Ed.; Gower: Aldershot, U.K., 1984; pp 162–166.

(18) Roussel, J.-C.; Norton, P.; Shenton, K. E.; Roesch, C. Handling Markush Formulae in Patents with the DARC PMS Software. Presented at the 192nd National Meeting of the American Chemical Society, Anaheim, CA, Sept 1986.

(19) Fearns, E. A.; Shenton, K. E.; Langdon, M.; Norton, P. Development of the Derwent Markush Graphics Database for Patents. Presented at the 192nd National Meeting of the American Chemical Society, Anaheim, CA, Sept 1986.

(20) Tokizane, S.; Monjoh, T.; Chihara, H. Computer Storage and Retrieval of Generic Chemical Structures Using Structure Attributes. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 177–187.

(21) Wigington, R. L. Machine Methods for Accessing Chemical Abstracts Service Information. *Proceedings of IBM Symposium on Computers and Chemistry*; IBM Data Processing Division: White Plains, NY, 1969.

(22) Fisanick, W.; Mitchell, L. D.; Scott, J. A.; Vander Stouw, G. G. Substructure Searching of Computer-Readable Chemical Abstracts Service Ninth Collective Index Nomenclature Files. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 73–84.

(23) Dunn, R. G.; Fisanick, W.; Zamora, A. A Chemical Substructure Search System Based on Chemical Abstracts Index Nomenclature. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 212–218.

(24) Farmer, N. A.; O'Hara, M. P. CAS ONLINE—A New Source of Substance Information from Chemical Abstracts Service. *Database* **1980**, *3*, 10–25.

(25) Zeidner, C. R.; Amoss, J. O.; Haines, R. C. The CAS ONLINE Architecture for Substructure Searching. *Proceedings of the 3rd National Online Meeting*; Learned Information: Medford, NJ, 1982; pp 575–586.

(26) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. The CAS ONLINE Search System. 1. General System Design and Selection, Generation, and Use of Search Screens. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93–102.

(27) Fisanick, W. Requirements for a System for Storage and Search of Markush Structures. In *Computer Handling of Generic Chemical Structures*; Barnard, J. M., Ed.; Gower: Aldershot, U.K., 1984; pp 106–129.

(28) Fisanick, W. Chemical Abstracts Service Markush (Generic) Structure System. Presented at the 192nd National Meeting of the American Chemical Society, Anaheim, CA, Sept 1986.

(29) Fisanick, W. Storage and Retrieval of Generic Chemical Structure Representations. U.S. Patent 4 642 762, Feb 10, 1987.

(30) Kaback, S. M. Access All the Information in Patents. *CHEMTECH* **1985**, *15*, 146–151.