

The Grammar of Markush Structure Searching: Vocabulary vs Syntax[†]

EDLYN S. SIMMONS

Marion Merrell Dow Inc., Cincinnati, Ohio 45215-6300

Received November 16, 1990

Markush structures are the language of chemical patents, comprising a vocabulary of chemical fragments organized by a syntax of structural relationships. Chemical information specialists have long sought the perfect notation system for Markush structures in patent documents. Fragmentation codes describing the atoms, rings, and functional groups that make up the vocabulary of Markush structures have long been used by established retrieval systems such as Derwent's Chemical Patents Index and IFI/Plenum's CLAIMS Uniterm and Comprehensive Databases. Topological search systems describing the syntax of Markush structures have recently been developed. This paper will discuss the grammar of Markush structures and will compare the vocabulary-based approach of fragmentation codes with the syntax-based approach of topological indexing systems.

A Markush structure is the generic expression for a class of chemical substances, the format conventionally used in patents, and consists of a molecular skeleton bearing one or more variable substructures with lists of alternative definitions for the variable portions of the molecule. Since, in a general sense, any form of expression is a language and languages are governed by systems of grammar, the retrieval of chemical information from patents is dependent upon the grammar of Markush structures.

In the beginning of the science of chemistry, before the atomic nature of matter was known, each form of matter was named by its discoverers in the same arbitrary way they named plants and animals. As it was discovered that conversions were possible from one form of matter to another, more systematic names were constructed to express the relationships between them.¹ Various chemists introduced names for features that conferred the chemical or physical properties of different compounds. The definitions did not always agree. To this day there are several different definitions for basic terms like "aromatic" and "acid", and the various meanings are used in different contexts. In the 18th and 19th centuries enough progress was made in the science of chemistry that the international community of chemists was able to agree on the identity of the elements that make up more complex compounds and to agree on a symbol for each element. The names and symbols of the elements correspond roughly to an alphabet. Chemists discovered that there are substructures like ring systems and functional groups that occur in many different compounds and contribute to the structure and properties of the whole molecule. It is possible to think of the substructures as words that contribute to the idea the whole compound represents. The substructures are connected in the skeleton of the molecule according to rules of bond formation and stereochemistry that correspond roughly to the rules of syntax for combining words into sentences that express entire concepts.

Chemists have never been able to agree on the one and only valid system of nomenclature. Several systems are used concurrently along with the vestiges of unsystematic names. The unreliability of chemical terminology is offset somewhat by the use of structural diagrams in chemical notation systems. Each chemical notation system acts as a language whose purpose is to convey the identity of chemical substances. Several conventional methods for drawing the skeletons of chemical substances grew up, but as long as only a few scientists were working with a small number of substances, they

could deal with the diversity of nomenclature and notation. Now that hundreds of thousands of chemists work with millions of chemical substances, dealing with the diversity is more difficult.

It was during the development of chemical notation that the world's patent systems developed. The developing patent law was sympathetic to the fledgling science of chemical notation. If a person of ordinary skill in the art understood what was expressed in a patent application, its author could use any notation he or she wished. This is a reasonable attitude for a system that deals with the new and nonobvious: if an invention is new there may well be no terminology for it yet. If an inventor has discovered a new genus of compounds that share similar properties, he or she is welcome to describe it by using any kind of nomenclature and notation that will clearly delineate the claimed invention and enable a skilled person to practice it. Generic descriptions of compounds became common in patents as soon as the structures of chemical substances and the structural relationships among them were discovered. The Markush format gained its name and began to grow in popularity after the United States Patent Office published a decision that the format was acceptable in 1925.² Patents include generic descriptions that are not Markush structures. For example, the patentee can refer to a genus of aliphatic alcohols and the intended meaning will be understood without the need for a drawing. But since applicants for patents are required to be very precise in describing exactly which compounds are covered by a claim, nearly all modern chemical patents use Markush structures to define a class of compounds and exclude all others.

The courts have ruled that the inventor is his own lexicographer.³ Because the inventor defines the terms to be used and the relationships among them, each Markush structure is a unique, self-contained notation system. The drawing of a Markush structure and the accompanying textual disclosure form a dictionary that defines the vocabulary of the chemical idea the Markush structure represents. The Markush structure identifies all the substructures that can be used in the invention and provides the geometrical relationship among the substructures: that is, the Markush structure provides a vocabulary of chemical words and a connecting syntax for a single genus of compounds. The vocabulary can be in terms of real atoms and bonds, generic structural terms like alkyl or halogen, structural drawings, and functional terms not related to the skeleton of the molecule, like "an electron withdrawing group". Each generic term is to be understood as the synonym for each of the individual substructural terms encompassed within it, and vice versa. Methyl, whether stated or shown as CH₃, is the synonym for C_{1-n}-lower alkyl.

Many of the features commonly used in Markush structures

[†] Presented at a symposium on Markush Structure Files and Searching: Status Report, at the 200th National Meeting of the American Chemical Society, Washington, DC, Aug 29, 1990.

United States Patent [19]

Archer

[11] Patent Number: 4,851,417

[45] Date of Patent: Jul. 25, 1989

[54] 9-SUBSTITUTED
6H-PYRIDO[4,3-B]CARBAZOLES

[75] Inventor: Sydney Archer, Delmar, N.Y.

[73] Assignee: Rensselaer Polytechnic Institute,
Troy, N.Y.

[21] Appl. No.: 198,976

[22] Filed: May 26, 1988

[51] Int. Cl.⁴ A61K 31/475; C07D 471/04

[52] U.S. Cl. 514/285; 546/70

[58] Field of Search 546/70; 514/285

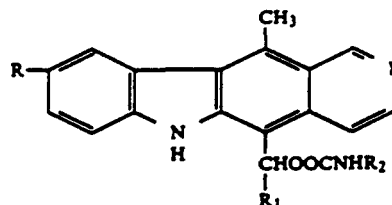
Primary Examiner—Donald G. Daus

Assistant Examiner—Andrew G. Rozycki

Attorney, Agent, or Firm—Notaro & Michalos

[57] ABSTRACT

Compounds comprising the series of 9-substituted 5-

hydroxymethyl-11-methyl-6H-pyrido[4,3-b]carbazole
N-alkyl or aryl carbamates of general structure:

where R=H, lower alkoxy, OH, aryloxy

R₁=H or lower alkylR₂=lower alkyl or aryl.

7 Claims, No Drawings

Figure 1. US Patent 4,851,417.

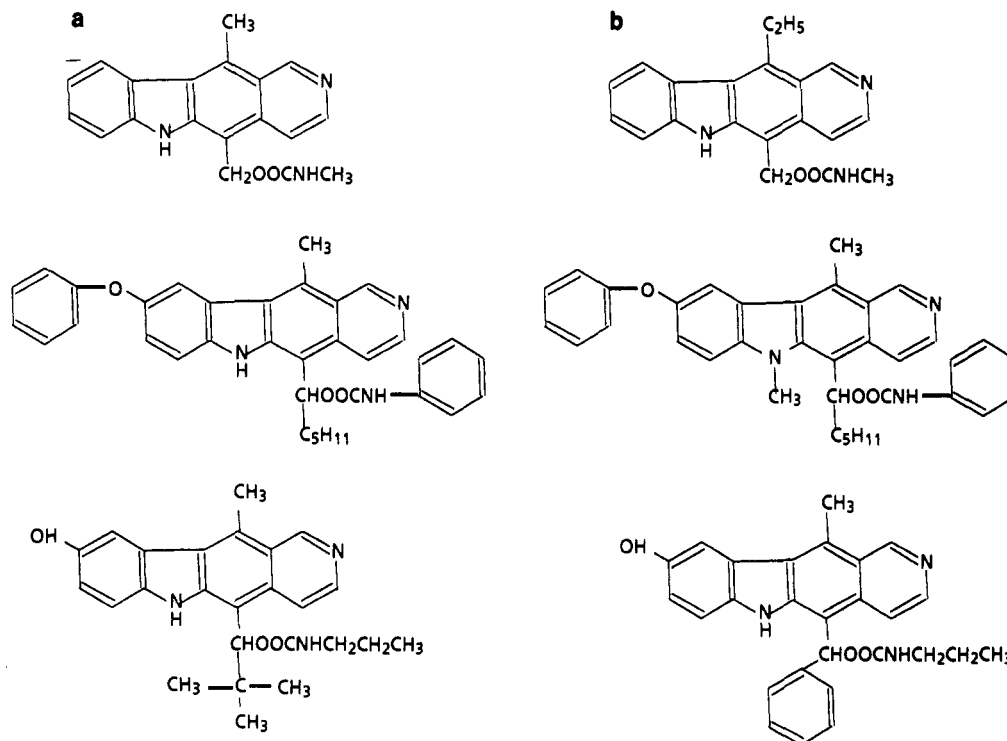


Figure 2. (a) Three embodiments of US 4,851,417. (b) Three compounds not encompassed by US 4,851,417.

are expressed in ways that are familiar from ordinary chemical notation or have migrated into common use as the Chemical Abstracts Service and other database producers have introduced generic query languages for databases of specific structures. A typical Markush claim, as shown on the first page of US 4,851,417⁴ (Figure 1), has a structural diagram with ordinary symbols for the atoms and bonds that are required in every embodiment and generic symbols, especially the letter "R", for variable substructures. The traditional format for Markush notation identifies optional substituents as members of a Markush group, using terminology like "R is selected from the group consisting of hydrogen and lower alkyl", but it is acceptable in most patent offices to list the members of the group as simple alternatives, for example, "R is hydrogen or lower alkyl".

Markush format is not used exclusive for chemical structures. A patent claim could be directed to "a method for treating a disease selected from the group consisting of AIDS and AIDS-related complex". That is expressed in Markush format as well. In the context of chemical substances, a Markush structure is simply a shorthand expression for a group of specific compounds. It encompasses every individual compound that can be constructed by varying its optional substructures according to its rules of syntax, the individual compounds being referred to as embodiments.

It is important to recognize the difference between a search for a Markush structure and a substructure search of the kind chemists are accustomed to performing in the CAS Registry File. Substructure searches preserve the syntax of the molecule and allow additional vocabulary. A Markush structure is

Class/subclass	Class/Subclass Title
514	DRUG, BIO-AFFECTING AND BODY TREATING COMPOSITIONS
514/1	DESIGNATED ORGANIC ACTIVE INGREDIENT (DOAI) CONTAINING
514/183	• Heterocyclic carbon compounds containing a hetero ring having chalcogen (i.e., O,S,Se or Te) or nitrogen as the only ring hetero atoms DOAI
514/277	•• Hetero ring is six-membered consisting of one nitrogen and five carbon atoms
514/279	••• Polycyclic ring system having the six-membered hetero ring as one of the cyclos
514/284	•••• Tetracyclo ring system having the six-membered hetero ring as one of the cyclos
514/285	••••• Plural hetero atoms in the tetracyclo ring system (e.g., acronycines, etc.)

Figure 3. United States patent classification of US 4,851,417.

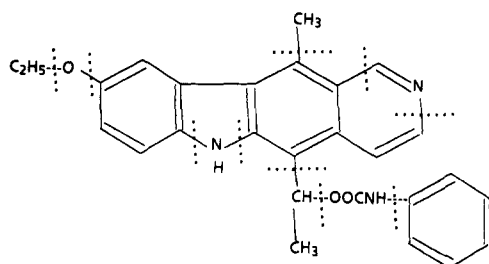


Figure 4. IFI/Plenum fragmentation system.

Searchable Codes for Required or Permitted Fragments

Functional Group Fragments

40026	CN C=N
40044	CNO ₂ Carbamic acid, carbamate
40304	HN Secondary amine
40305	HO Hydroxy
40417	O Ether

Ring Fragments

40551	Benzene ring
36729	C ₄ N,C ₅ N,C ₆ ,C ₈ Generic ring
30035	Carbocyclic ring
34210	Fused or bridged ring
34236	Heterocyclic ring

Figure 5. Fragmentation coding for US 4,851,417 (IFI/Plenum Uniterm Database).

Searchable Codes for Required or Permitted Fragments

Functional Group Fragments

30306	CN C=N (1 Possible)
30385	CNO ₂ Carbamic acid, carbamate (1 Possible)
32742	HN Secondary amine (1 Possible)
32745	HO Hydroxy (1 Possible)
33697	O Ether (1 Possible)

Ring Fragments

34701	Benzene ring
36729	C ₄ N,C ₅ N,C ₆ ,C ₈ Generic ring
30035	Carbocyclic ring
34210	Fused or bridged ring
34236	Heterocyclic ring

Negation Codes

Must Codes: Required Groups For Restricting Retrieval in Searches Where These Fragments Are Forbidden

Functional Group Fragments

30063	Amine
30305	CN C=N
30384	CNO ₂ Carbamic acid, carbamate
32741	HN Secondary amine
33775	Uncommon functional group

Ring Fragments

34211	Fused or bridged ring
34276	Nitrogen in ring
34237	Heterocyclic ring
34263	Maximum ring unsaturation
36989	Uncommon ring

Configuration Terms

34198	Functional Group on aliphatic carbon
37745	Three carbon atoms between functional groups

Figure 6. Fragmentation coding for US 4,851,417 (IFI/Plenum Comprehensive Database).

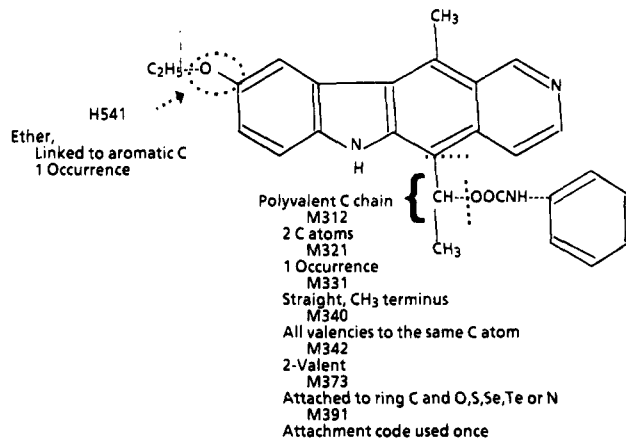


Figure 7. Derwent fragmentation system.

Searchable Codes for Required or Permitted Fragments (28 of 50)

D022	1 Carbocyclic ring substituent; ≥ 8 to ring fusion
D023	2 Carbocyclic ring substituents
E330	Fused ring system; N sole heteroatom(s); ≥ 4 rings
G100	Uncondensed benzene as the sole carbocycle
H401	Total of 1 OH
L462	=N-C(=O)-O- linked thru N to aromatic ring C
L463	=N-C(=O)-O- not described by previous codes
M122	Benzene linked to other aryl C by group coded below
M141	Rings linked by -O-
M210	C ₁₋₆ 0- or 1-valent C chain
M211	1-Carbon 0- or 1-valent C chain
M231	Unbranched 0- or 1-valent C chain
M233	Tertiary C in 0- or 1-valent C chain
M240	0- Or 1-valent C chain attached to ring C
M272	0- Or 1-valent C chain attached to O
M273	0- Or 1-valent C chain attached to N
M311	1-Carbon ≥ 2-valent C chain
M321	M31: code used once
M331	Straight ≥ 2-valent C chain with aliphatic branch
M340	All valencies of ≥ 2-valent C chain attach through 1 C
M373	≥ 2-valent C chain attached to ring C and (N, O, S, Se or Te)
M511	1 Fused ring heterocyclic system
M530	0 Aromatic ring systems
M531	1 Aromatic ring systems
M532	2 Aromatic ring systems
M540	0 Alicyclic ring systems
04518	Pyrido[4,3-b]carbazole

Negation Codes

Essential Group Codes: Required Groups For Restricting Retrieval in Searches Where These Fragments Are Forbidden

K0	A group coded K: or L: always present
L4	A U-C(=T)-(N, O, S, Se or Te) moiety always present

Figure 8. Fragmentation coding for US 4,851,417 (Derwent Chemical Patents Index).

normally closed to further substitution. It is for this reason that many Markush structures in patents refer to groups that are "optionally substituted". Whereas references to closely


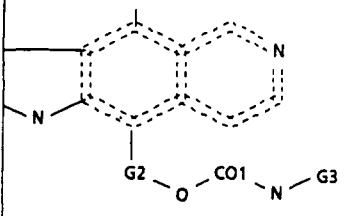
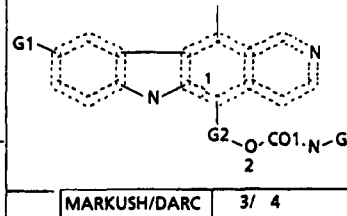
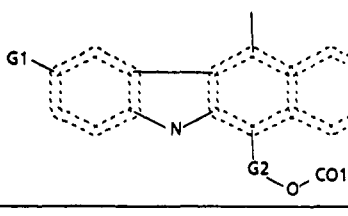
MARKUSH/DARC	3/ 4	CN:8937-23101	WPIM
-FG: 2- -GM: 4/ 4- 		H CHK C	
MARKUSH/DARC	3/ 4	CN:8937-23101	WPIM
-FG: 0- -GM: 3/ 4- 		CHK Ph C	
?			
MARKUSH/DARC	3/ 4	CN:8937-23101	WPIM
-FG: 0- -GM: 2/ 4- 		CHK 1 2 G4	
?			
MARKUSH/DARC	3/ 4	CN:8937-23101	WPIM
-FG: 0- -GM: 1/ 4- 		H 1 O-CHK 1 O-Ph O-ARY	
7			

Figure 10. Markush DARC structure record for US 4,851,417 (WPIM Variable Groups).

years of the 20th century, a great many organizations tried to create fragmentation codes for indexing compounds, including those in patents.⁶ The original systems were proprietary, and when individual companies gave up their attempts to index the patent literature for themselves, the commercial systems that replaced them were made available to subscribers at a substantial cost. With the exception of the CLAIMS Uniterm Database and the Pharmsearch and MARPAT Files that were introduced in 1989 and 1990, respectively, databases with comprehensive indexing of Markush structures are still restricted to subscribers.

It is likely that the fragmented nature of Markush structures themselves suggested the use of fragmentation codes. Markush structures *are* fragmentation systems. The ideal way to devise a fragmentation system for indexing patents would be to identify every possible molecular substructure and give each one a name or code number that will be used in the retrieval system. Since there is an infinite number of possible chemical fragments, the creators of each database were forced to decide which fragments ought to be searchable. The designers of each

fragmentation code created a vocabulary of code terms for substructures that occur frequently in patents and a set of rules for applying them, in effect, another notation system. An indexer simply decides on a code-by-code basis whether the substructure it defines is permitted in a particular published chemical structure and enters each code into a record. To ensure that the record will be retrieved whenever an overlapping structure is searched, the indexer has to expand the generic term in the Markush structure into all possible specific combinations and permutations. A few of these embodiments of the Markush structure in U.S. 4,851,417 are illustrated in Figure 2. In this patent this procedure involves combining the groups in the definition of R, i.e., the 16 isomers of C₁₋₅ alkoxy groups, hydrogen, hydroxy, and aryloxy (which means phenoxy and maybe naphthyloxy), with each of the optional meanings of R₁ and R₂.

To retrieve references from the database, a searcher does a similar analysis with the chemical structure to be searched. He or she visualizes all the embodiments, identifies all possible code terms, and then combines the terms into an appropriate

through all single bonds that connect a carbon atom with either a heteroatom or a multiply bonded carbon atom, including those in most ring systems.

The fragments are then matched with appropriate 5-digit code numbers from the fragment code thesaurus, shown in Figure 5. Commonly occurring ring systems and functional groups have unique code numbers. Uncommon rings and functional groups are given generic fragmentation codes based on the empirical formula of the group.

The difficulty with the pure fragmentation system used for the Uniterm Database is that there are so many patents with common functional groups that most searches based only on structure retrieve an unmanageable number of false drops, and searches have to be limited with nonstructural codes from the thesaurus of general terms.

IFI/Plenum has a more specific fragmentation code system for its Comprehensive Database.⁷ Using the same system for fragmenting molecules, the Comprehensive indexing system counts the number of times common functional groups occur and assigns a different numerical code to each number of occurrences that are possible. Codes for fragments that are permitted, including those that are required, are designated as "POSSIBLE" codes.

Limiting retrieval by negating codes for forbidden substructures is not possible in a file of Markush structures. Simply negating the codes for substructures that are forbidden in the query structure would eliminate records for prior art generic structures that have a Markush group containing permitted substructures as alternatives to forbidden ones. False drops from a database of Markush structures are records that require fragments that are forbidden in the query structure.

In the Comprehensive Database, IFI/Plenum assigns a second set of codes to common fragments that are required in all embodiments of a Markush structure. These are designated as "MUST" codes. These codes are not designed to help searchers retrieve the patent records that contain them. They are designed only to prevent retrieval of the records when they would be false drops. The false drops are eliminated by searching for the presence of the POSSIBLE codes that are either required or permitted, and then negating from the answer set the records that contain a MUST code for any forbidden fragment. This cuts down enormously on the number of false drops, especially in searches for compounds with unusual substructures like the ring system in U.S. 4,851,417. There are a great many individual and hierarchical MUST codes, and many of them must be negated for precise retrieval, but, as shown in Figure 6, only a few apply to any particular patent.

The fragmentation code used for nonpolymeric compounds in Derwent Publications Ltd.'s Chemical Patents Index⁸ is not based purely on vocabulary. The Derwent chemical fragmentation code has no clear-cut code generating system. As indicated in Figure 7, many of the code terms define not only a functional group but its connection in the molecule. The CPI code system has no single code for ethers, for example. It has a series of codes for the numbers of ether groups bonded to carbon atoms in heterocyclic rings, aromatic rings, alicyclic rings, and aliphatic chains. Some code terms define only the context of a fragment in the molecule. The ethylene chain in the illustrated molecule is defined by seven different codes. Whereas the codes that describe specific aspects of the ethylene chain would be helpful in restricting search results to a specific compound, most of the code definitions are too detailed to use when searching for a generic structure.

There are a great many codes that apply to one or more embodiments of the Markush structure, especially to carbon chains of various lengths and degrees of branching. Figure 8 shows some of the codes applied to U.S. 4,851,417. In 1972 the Derwent chemical fragmentation code was modified to

eliminate false drops due to the nonspecific nature of many individual codes, and since that time all ring systems that were previously searchable only through generically defined codes also have a specific Ring Index Number. The Derwent system also has a set of codes assigned to required fragments and negated to eliminate records where forbidden groups are required. These were introduced in 1970, with additional negation codes added to the system in 1981. Derwent calls them Essential Group codes and uses a much smaller set than IFI. This combination of vocabulary and syntax allows a searcher to retrieve a relatively smaller number of false drops based on structure alone than is possible from the Comprehensive Database.

Derwent originally had several slightly different fragmentation code systems for various technologies. In 1981 the code systems were merged and additional terms were added. The resulting improvement in precision was offset by the complex time ranging it imposes on retrieval strategies. Despite the improvements made in the fragmentation code over the years, and in part because of them, users of Derwent's fragmentation code in the mid-1980s had a number of complaints:

The code was difficult to learn and difficult to apply, especially since a code strategy required several time-ranged levels of specificity.

It resulted in too many false drops, especially in the older portions of the file that were indexed before the more specific codes were introduced.

It was impossible to reconstruct the structure of the indexed compounds from the coded record, and after a copy of the patent document or its abstract was obtained, it was difficult to recognize the compounds that caused the patent to be retrieved in the document.

New users of the code did not have sufficient access to training and support.

Many Derwent subscribers looked at the relatively simple topological systems for searching the CAS Registry File on CAS Online (now STN International) and Questel and asked that Derwent replace the fragmentation code with a system that could match Markush structures to a structural drawing.

Derwent joined with Telesystemes Questel and the French Patent Office, INPI, to develop the Markush DARC system.⁹ Derwent's WPIM structure file for the WPIL Database and INPI's Pharmsearch with the corresponding structure records in the MPHARM File were introduced early in 1989. Markush DARC accepts topological input to search a structure file, but it does not simply allow the user to draw any structure he or she wants to search. Markush DARC is another form of chemical notation with its own vocabulary and rules of syntax. Figure 9 shows the parent group display for U.S. 4,851,417 from the WPIM File. You can see that this record looks much more like the original Markush structure than does a list of alphanumeric codes.

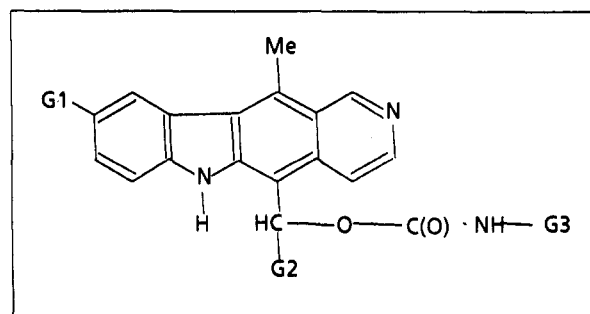
But not all structures are this similar to the original when shown in Markush DARC notation. For example the variable groups for this record, shown in Figure 10, are not exactly recognizable from the structure in the patent. That is due in part to the fact that topological indexing systems are not completely compatible with the fragmented nature of the Markush format. Indexers have to provide syntax to Markush structures in order to make them searchable in Markush DARC notation. It is also because Markush DARC has a vocabulary of fragment codes called "superatoms" for generically described groups, and the notation for the superatom does not necessarily correspond to the notation in the patent document. For the time being, at least, superatoms are not interchangeable with the real substructures they represent.

Patents are indexed independently by Derwent and INPI for the two structure files.¹⁰ In this case, the parent group in

L8 ANSWER 1 OF 1
 COPYRIGHT (C) 1990 AMERICAN CHEMICAL SOCIETY

AN ***112:35839*** MARPAT

MSTR 1



VAR G1 = H / alkoxy<(1-5)> / OH / OPh / (SC OMe)

VAR G2 = H / alkyl<(1-5)>

VAR G3 = alkyl<(1-5)> / Ph / (SC Me)

MPL: claim 1

Figure 12. MARPAT structure record for US 4,851,417.

the corresponding MPHARM record looks the same as in WPIM, but Figure 11 shows that the variable groups were interpreted differently by INPI's indexers. Markush DARC will not retrieve this patent if the searcher modifies the parent structure by specifying an ethoxy group where G1 is shown: both databases have indexed the ether by designating only methoxy, the patent's only example of alkoxy, and the CHK superatom bonded to oxygen. If the searcher specified an isopropyl group in place of G3, he or she would retrieve this patent in MPHARM but not in WPIM, because the indexers have translated the vocabulary of the patent into Markush DARC vocabulary differently in the two databases. On the other hand, if one were to search for an analogous molecule where the heterocyclic ring and the carbamate were separated by more than the one carbon atom specified by this patent, this patent would be retrieved because both indexers have destroyed part of the patent's syntax by using the superatom CHK to represent the optionally substituted methylene group shown in the parent structure record as G2.

While Derwent was working with Telesystemes and INPI on their topological search system, Chemical Abstracts Service was also developing a topological system for indexing Markush structures.¹¹ The MARPAT File introduced in 1990 was never intended to replace an earlier system; it is an attempt to meet the demand of CAS customers who wish to retrieve more than the examples from patents indexed in *Chemical Abstracts*.

As can be seen in Figure 12, the notation used for the MARPAT system is more compact than the notation in either the patent or the Markush DARC system. Unlike the Markush DARC records in Figures 9-11, the structure shown in a MARPAT record is not necessarily what the computer searches. MARPAT has a very small vocabulary of generic group symbols that includes, in addition to the usual generic symbols for variable atoms, one symbol for aliphatic groups and three for rings. The software is designed to recognize all aliphatic groups and rings and to match them against the generic groups, so that G1 will match ethoxy even though it is not shown in the structure record or exemplified in the patent. The computer accomplishes this by dividing the query structure and each candidate-indexed structure into real atoms, and "original generic groups", reducing the structure to simple combinations of heteroatomic, cyclic, and aliphatic "rolled-up

generic groups". Matching is done by comparing the resulting reduced structure graphs. The searcher specifies the Match Level for the search to guide the software in matching the real atoms and rolled-up generic groups of the query structure with the real atoms, original generic groups, and spinoff generic groups in the indexed structures. This extremely powerful retrieval system introduces new vocabulary, and the syntax is unlike any of the notation systems in common use up to this time. Unfortunately, the software is so good at recognizing specific groups as members of the few generic groups in its vocabulary that much of the syntax in the query can be lost and a large number of false drops can be retrieved.

Are topological indexing systems better for Markush structures than fragmentation codes? The topological systems have been available for only a short time, and only limited comparisons have been made.^{12,13} Preliminary results seem to indicate that some searches can be done more efficiently with topological search systems while others can be completed easily with fragmentation codes and cannot be done at all with these topological systems. It should not be surprising that search queries without a well-defined molecular skeleton are difficult to complete in MARPAT or Markush DARC. Topological systems emphasize syntax, while fragmentation codes emphasize vocabulary. The Markush format allows both to vary. The difficulties encountered in searching for Markush structures are rooted in the complexity of Markush structures themselves, and we should not be surprised if we discover that none of the retrieval systems that are now available provides the ideal combination of simplicity, recall, and precision.

REFERENCES AND NOTES

- (1) Priesner, Claus. How the language of chemistry developed. *Chem. Int.* **1989**, 11 (6), 216-224, 237-238.
- (2) Ex parte Markush, 340 O.G. 839 (1924).
- (3) Lear Siegler, Inc. vs Aeroquip Corp. *U.S. Pat. Q.* **1984**, 221, 1025-1034, and cases cited therein.
- (4) Archer, Sydney. U.S. 4,851,417. Assigned to Rensselaer Polytechnic Institute. July 25, 1989.
- (5) Simmons, Edlyn S. The Paradox of Patentability Searching. *J. Chem. Inf. Comput. Sci.* **1985**, 25, 379-386.
- (6) Hunsberger, I. M.; Frear, D. E. H.; Harmon, R. E.; Smith, E. G. *Survey of Chemical Notation Systems*. National Academy of Sciences-National Research Council: Washington, DC **1964**, Publication 1150.

- (7) (a) Lambert, Nancy. How to Search the IFI Comprehensive Database Online...Tips and Techniques. *Database* 1987, 10 (6), 46-59. (b) Donovan, K. M.; Wilhide, B. B. A user's experience with searching the IFI Comprehensive Database to U.S. Chemical Patents. *J. Chem. Inf. Comput. Sci.* 1977, 17 (3), 139-143. (d) Balent, Mary Z.; Emberger, Jane M. A unique chemical fragmentation system for indexing patent literature. *J. Chem. Inf. Comput. Sci.* 1975, 15 (2), 100-104.
- (8) (a) Norton, P. Central Patents Index (CPI) as a Source of Information for the Pharmaceutical Chemist. *Drug Inf. J.* 1982, 208-215. (b) Kaback, Stuart M. Chemical structure searching in Derwent's World Patents Index. *J. Chem. Inf. Comput. Sci.* 1980, 20 (1), 1-6. (c) Simmons, Edlyn S. The Central Patents Index Chemical Code, A User's Viewpoint. *J. Chem. Inf. Comput. Sci.* 1984, 24 (1), 10-15.
- (9) (a) Shenton, Kathleen E. Graphic retrieval of patent information. Proceedings of the 9th International Online Information Meeting, London Dec 3-5, 1985, pp 43-59. (b) O'Hara, M. P.; Pagis, Catherine. The PHARMSEARCH Database. *J. Chem. Inf. Comput. Sci.* 1991, 31, 59-63.
- (10) Cloutier, Kathleen, A. A Comparison of Three Online Markush Databases. *J. Chem. Inf. Comput. Sci.* 1991, 31, 40-44.
- (11) (a) Fisanick, William. The Chemical Abstracts Service Generic Chemical (Markush) Structure Storage and Retrieval Capability. 1. Basic Concepts. *J. Chem. Inf. Comput. Sci.* 1990, 30 (2), 145-154. (b) Fisanick, William, U.S. 4,642,762. Assigned to American Chemical Society. Feb 10, 1987. (c) Fisanick, William. Requirements for a system for storage and search of Markush structures. In *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; Barnard, John M., Ed.; Gower: Aldershot, U., K., 1984; pp 106-129. (d) Ebe, Tommy; Sanderson, Karen A.; Wilson, Patricia S. The Chemical Abstracts Service Generic Chemical (Markush) Structure Storage and Retrieval Capability. 2. The MARPAT File. *J. Chem. Inf. Comput. Sci.* 1991, 31, 31-36.
- (12) Schoch-Grübler, Ursula. (Sub)Structure Searchers in Databases containing Generic Chemical Structure Representations. *Online Rev.* 1990, 14 (2), 95-108.
- (13) Wilke, Robert N. Searching for Simple Generic Structures. *J. Chem. Inf. Comput. Sci.* 1991, 31, 36-40.

A Comparison of the MARPAT and Markush DARC Software[†]

NORMAN R. SCHMUFF

NTEK Information Services, 9 Forest Drive, Baltimore, Maryland 21228-5028

Received November 17, 1990

MARPAT and Markush DARC are compared, with an emphasis on the user's interaction with the software. There are fundamental dissimilarities in both text and graphical structure point. Other important differences relate to bonding conventions, superatom definition, and search algorithm. The query translation of MARPAT puts it at a significant advantage over M-DARC.

INTRODUCTION

The recent introduction of the MARPAT File along with Markush DARC (M-DARC) brings to two the number of commercially available systems for Markush structure searching. While a recent publication has appeared comparing structure searching using DARC and CAS ONLINE,¹ it seems worthwhile to examine some of the similarities and differences of the corresponding Markush systems. This paper will focus on certain noteworthy software aspects of the two.

INPUT

After deciding on the most appropriate of the three Markush databases (Derwent's WPIM, INPI's Pharmsearch, or MARPAT), the next issue to confront the patent searcher is query construction. This can be thought of as a two-step process: what query do I use and how do I accomplish query input.

The cost conscious searcher will next consider how to build offline at least a partial query for uploading. This can be accomplished either by the use of a nonspecialized program for ASCII text input or by the intervention of a graphical front-end.

Text. Each approach has advantages and disadvantages. Text input will typically involve the creation of a small ASCII file using word processing software, and subsequent uploading with a terminal-emulation package. Both programs will typically be those that are frequently used in a variety of contexts; and consequently, they will be programs with which the searcher is familiar.

A disadvantage is that this approach requires a thorough familiarity with the commands for query construction and

attention to the structure and syntax of these commands. A missing space or a misplaced comma can have serious consequences. This contrasts with graphical input which is considerably more intuitive and gives immediate visual feedback.

Figure 1 compares the text input for the indicated structure, a novel HIV inhibitor. At first glance the requisite text strings seem comparable in size and complexity. There is, however a significant difference. With M-DARC, the query can be numbered in any arbitrary way, while MARPAT requires prior knowledge of how the benzodiazepine ring system will be numbered.

The MARPAT system is not without its advantages. The GRA Rxx... command provides an expeditious method for building polycyclic rings (e.g., GRA R66U6D5 builds the steroid skeleton). Also, the commands are consistent with those used in the other structure-searchable STN files, Registry (REG) and Beilstein (BEIL). On the other hand, "BON R 1 2 N" hardly seems an obvious way to designate the six-membered ring as being aromatic. Overall, for textual query input, M-DARC seems preferable.

Graphical Front-Ends. In order to overcome many of the limitations of text uploading, a number of companies have developed graphical front-ends for query construction and uploading.^{2,3} Table I summarizes some of the features of these packages.

Given the complexity of MARPAT and M-DARC, it is not too surprising that only the front-ends produced by the vendors fully support their respective search software. The disadvantages of DARC Chemlink and STN Express is that both are relatively expensive; and both are specialized packages, currently limited to one system.

My personal experience is limited to the use of the Chem-Connection and STN Express for the Macintosh. I use the former frequently, but mainly as a tool for drawing high-quality chemical structures. It works reasonably well at

[†] Presented Aug 29, 1990, at the Markush Structure Files and Searching Symposium at the 200th ACS National Meeting, Washington, DC.