(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 Serive 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 be compared, with an emphasis on the user's interaction with the software. There are fundamental dissimilitudes 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.

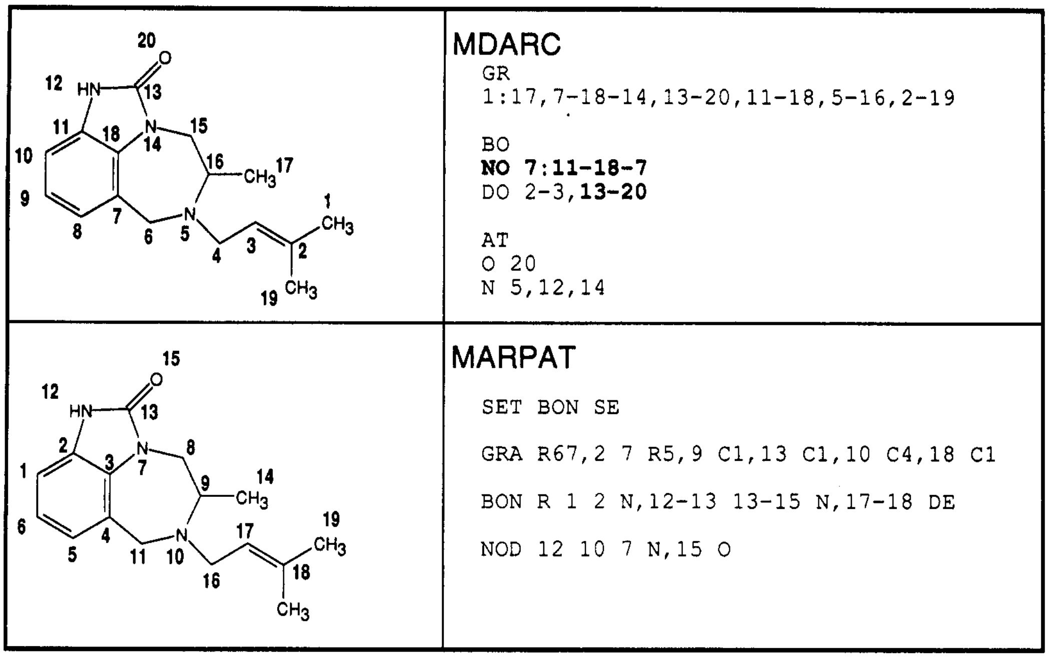
Graphical Front-Ends. In order to overcome many of the limitations of text uploading, a number of companies have developed graphical front-ends for query construction and uploading.[2,3] Table I summarizes some of the features of these packages.

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 ChemConnection 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.

0095-2338/91/1631-0053$02.50/0



MDARC

GR
1:17,7-18-14,13-20,11-18,5-16,2-19

BO
NO 7:11-18-7
DO 2-3,**13-20**

AT
O 20
N 5,12,14

MARPAT

SET BON SE

GRA R67,2 7 R5,9 C1,13 C1,10 C4,18 C1

BON R 1 2 N,12-13 13-15 N,17-18 DE

NOD 12 10 7 N,15 O

Figure 1. Query formulation for uploading. The text strings necessary for generation of the indicated query. Note that bold text indicates differences from the usual input for the exact compound databases (for DARC: EURECAS, POLYCAS, UPCAS; for STN: REG, LREG, Beilstein).

Table I. Front-End Software for MARPAT and Markush DARC Structure Input

	Telecomm.	Mac	MS-DOS	MARPAT	M-DARC	Notes	Available in N. America from:
STN Express	✗	✗	✗	✓		Imports SMD files	Chemical Abstracts Service 2450 Olentangy River Rd. Columbus, OH 43210
DARC Chemlink	✗		✗		✓	Available with several levels of Tektronix emulation	Questel, Inc. 5201 Leesburg Pike, Suite 603 Falls Church, VA 22041
ChemTalk Plus	✗		✗	✓–		Imports MOL files	Molecular Design, Ltd. 2132 Farallon Dr. San Leandro, CA 94577
MOLKICK			✗	✓–	✓–	TSR program	Springer-Verlag 175 Fifth Ave. New York, NY 10010
ChemConnection		✗		✓–		'Desk Accessory'	SoftShell International Ltd. 2754 Compass Dr., Suite 375 Grand Junction, CO 81506

✓ : capable of all aspects of query construction
✓– : capable of many, but not all, aspects of query construction

creating REG File structure input, but it has some difficulty with atom strings pictured in unusual ways (e.g., O_2N- instead of -NO_2). ChemConnection's producer, SoftShell, will soon include the ability for search and display of ROSDAL strings used for the searching of Beilstein on Dialog.

In my hands the Macintosh version of STN Express (v 2.01), which I had intended to use in processing queries for this paper, was virtually unusable for structure searching with frequent crashes and freezes, though this may be due in part to my system configuration. But, even ignoring the bugs, the Macintosh implementation of Express is poor. I was, however, able to import into Express a SMD File created with ChemDraw.

In spite of their lack of full MARPAT/M-DARC compliance, both Molkick[3] and ChemConnection have certain advantages. They are multifunctional tools and can be used in conjunction with familiar terminal-emulation packages.

Graphical Input Online. Interactive online input can sometimes be used efficiently if a partial text structure is initially uploaded. It does, however, suffer from relatively slow operation (improved somewhat by 9600-bp access, but still limited by sluggish mainframe response) and from the obvious disadvantage of paying connect hour charges during query input. With STN, the online interface is completely different from that of Express. I cannot comment on the DARC Chemlink vs DARC online consistency as I have no experience

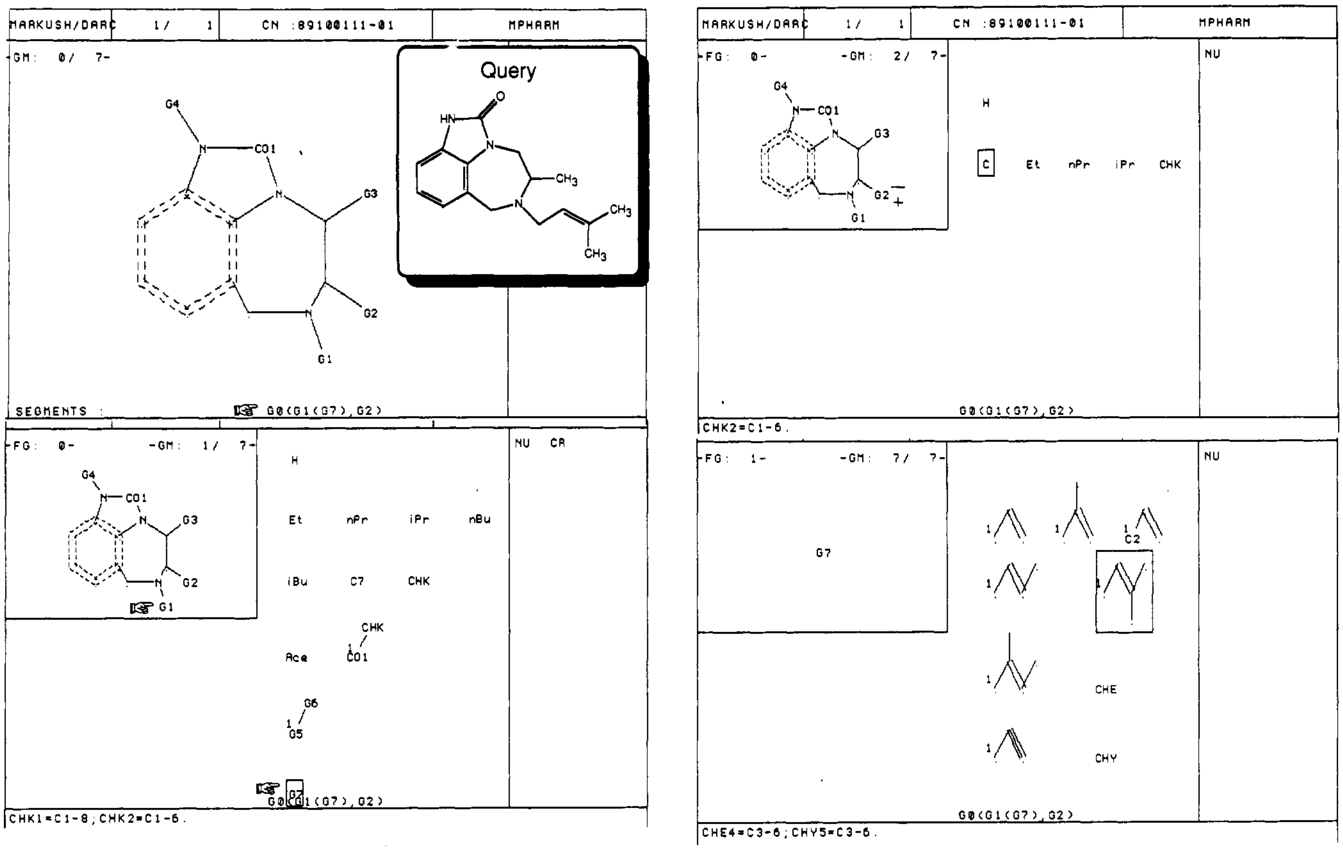


Figure 2. Markush DARC (MPHARM) answer resulting from the indicated query. Note the highlighting of the query structure.

with the former. For interactive online input, I prefer DARC, as STN "MENU" input does not respond to mouse clicks, but requires that the user coordinate mouse movement and keyboard input (e.g., "F", move the mouse, then "T" to draw a bond from point to point).

BONDING CONVENTIONS

A significant difference between the two systems is in the area of structural conventions. Some of these are summarized in Table II. CA has for the most part used the conventions that have been successfully used for the 10 million substances in the Registry File, while DARC has taken a different path (one might say a wrong turn!).

The text input of Figure 1 should look familiar to those conversant with DARC and the CA Registry File. The input for MARPAT is a superset of the commands used for REG/BEIL input, and consequently, queries can be constructed in any of the appropriate STN files and processed in any other of those files. For the structure in Figure 1, the input is identical with that for REG or BEIL. In fact, if superfluous MARPAT-specific commands are input for a REG or BEIL structure search, they will be ignored, but retained for subsequent processing in MARPAT. Conversely, queries created in MARPAT can be crossed into the other STN structure files. This shows forethought, and sensitivity to user needs, on the part of the file designers. But there is at least one unsettling note in that cyclopentadienyl–metal complexes are treated somewhat differently in MARPAT than in REG.

M-DARC on the other hand introduces many entirely new bonding conventions, used to my knowledge nowhere else on this planet. Some of these relate to the always troublesome areas of tautomeric/aromatic/normalized bonds and others to organometallic coordination compounds. While a few of the conventions do seem more intuitive than those of CA, many of them make little sense and some are counterintuitive. Compare, for example, in Table II the M-DARC entries for the amidine **2** (C-NH_2 bond normalized) with amide **3** (C-NH_2 bond single). Compare acid **1** (one single bond C-O, one double bond C=O) with amidine **2** (both carbon–nitrogen bonds normalized). These examples seem at odds with the stated goals that[4] "The rules should be as few in number, and as simple, as possible".

QUERY TRANSLATION

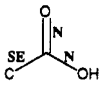
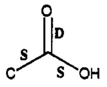
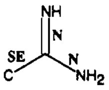
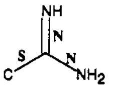
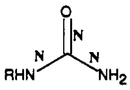
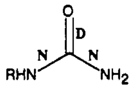
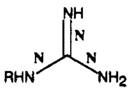
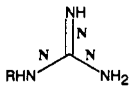
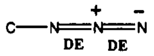
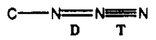
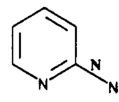
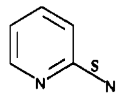
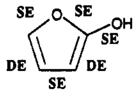
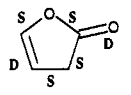
Processing of the query of Figure 1 in the MPHARM (Pharmsearch) file of M-DARC gives one answer shown in Figure 2. Quick visual identification of the answer is not easy, as the answer spread over four screens. The use of the "VIEW FOCUS" command does provide for some measure of search-term highlighting. At the bottom of screen 1, the "G0(G1(G7),G2)" indicates that the query is matched by fragments in these G-groups which can be displayed sequentially after a simple carriage return. Note that what was retrieved was an exact answer (i.e., no superatoms). This is because there is no translation of specifics into generic terms (e.g., the ring methyl is NOT translated into a lower alkyl—CHK LO).

It is just this automatic translation that produces 65 answers from MARPAT, the first answer of which is shown in Figure 3. This record is retrieved (without the pointers to the relevant search terms) from the so-called "spin-off" group. The imidobenzodiazepine generates "Hy" (heterocycle) while the methyl and pentenyl groups generate two "Ak", alkyl superatoms. The C=O group generates both =O and the tautomeric -OH.

Lack of search-term highlighting provides for answers that are difficult to match with their respective queries. Readability also suffers from a lack of grouping in the variable display. For example, for "VAR G5": does the "(1-2)" refer to the N or to the O?

Interpreting 65 answers such as these could be very time consuming. The best approach to reviewing answers is probably to trust in the almighty Fisanick[5] algorithm and take

Table II. A Comparison of Bonding Conventions in MARPAT and Markush DARC[a]

	MARPAT	M-DARC
1		
2		
3		
4		
5		
6		
7		

[a] MARPAT: DE, double exact; N, normalized; SE, single exact. M-DARC: D, double; N, normalized; S, single; T, triple.

advantage of MARPAT's ability to display CA bibliographic information, as well as the CA File's graphical abstract. This is somewhat analogous to performing a WPI/L coding search, and taking a look at Derwent's documentation abstract with its accompanying graphic (unfortunately not yet available online).

MARPAT—Match Level. Fortunately, the translation can be controlled using the "Match Level" (MLE) command. This is a powerful facility whereby translation can be turned on or off for various parts of the molecule. In our query, for example, imposing "Match Level Atom" (MLE ATOM) for all the ring atoms (i.e., MLE ATOM 1 2 3 4 5 6 7 8 9 10 11 12 13) would require an exact atom-for-atom match in the ring, resulting in only imidazobenzodiazepines (having either superatom or "real" alkyls attached). Likewise, imposing "Match Level Atom" for the isopentenyl chain would require that those five atoms be present with the given connectivity and bonding.

Note that although the match level command can be issued for only a part of the ring (e.g., the benzene portion), the software will actually impose this match level for the entire ring system. That is, "MLE ATOM 1 2 3 4 5 6" will in effect impose "MLE ATOM" for the entire ring system, even though the intent was to require an actual benzene but to permit fusion to any dinitrogen-containing heterocycle. Note, however that although "MLE ATOM 1" will give the same results as "MLE ATOM 1 2 3 4 5 6 7 8 9 10 11 12 13", the former search will be much slower, so that it is advantageous to use MLE ATOM for each ring atom.

Even considering its limitations, the translation facility represents a tremendous advance in commercial Markush searching. It places MARPAT at a considerable advantage over M-DARC. In M-DARC, lack of translation means the searcher must give careful consideration to query formulation. For example, in the current implementation, a search for an attached alkyl group must include a query superatom (e.g., CHK) and an exact atom chain (e.g., C*): the former to retrieve superatom answers, and the latter to retrieve "real" atom chain answers.

Table III. Groups Used in Query Construction: A Comparison of MARPAT's Generic Groups with Markush DARC's Superatoms

MARPAT Generic Groups / Special Nodes			M-DARC Superatoms		MARPAT (MP) / M-DARC (MD) differences
Ak	Ak	SAT	CHK	alkyl	
	Ak	UNS	CHE	alkenyl	
	Ak	UNS	CHY	alkynyl	no MP group specific for alkyne
Cy	Cb		CYC	cycloaliphatic, monocyclic or polycyclic	MD - no aromatic† rings MP - any bonding
	Cb	UNS	ARY	carbocyclic, monocyclic or polycyclic	MD - ≥1 benzene MP - ≥ 1 unsaturations
	Hy	MON UNS	HEA	heterocyclic, monocyclic	MD - ≥ 1 aromatic† rings MP - can optionally specify HIQ/LOQ*
	Hy	MON	HET	heterocyclic , monocyclic	MD - no aromatic† rings, any non-HEA monocyclic ring MP - can optionally specify HIQ/LOQ*
	Hy	PCY	HEF	heterocyclic, fused	MP - can optionally specify HIQ/LOQ*
	X		HAL	halogen	
	M		MX	metal	
			ACT	actinide	
			ACY	acyl	
			A35	group IIIa-Va	
			DYE	dye	
			LAN	lanthanide	
			PEG	polymer end group	
			POL	polymer, polypeptide residue	
			PRT	protecting group	
			TRM	transition metal	
			UNK	undefined group	
	Q			any atom except carbon or hydrogen	
	A		XX	any atom except hydrogen	

† By the M-DARC convention, "aromatic" is defined by the Hückle rule (i.e., $4n + 2$ π electrons in a cyclic path), so pyrrole, imidazole, etc. are considered to be aromatic. This is in contrast to the rules used to cstablish normalized bonding. * HIQ—more than 1 heteroatom, LOQ—exactly 1 heteroatom.

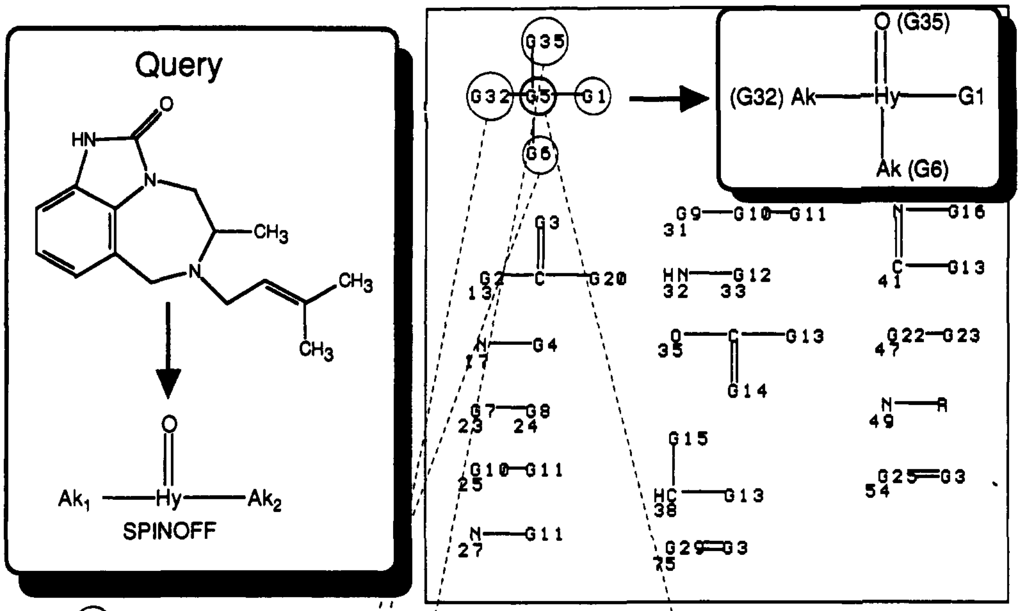
SUPERATOM COMPARISON

The Markush DARC superatoms are comparable to MARPAT's generic groups and system-defined nodes (all will be referred to subsequently as superatoms). Some of these represent closed sets (e.g., metals, halogens) while others represent infinite sets (e.g., carbocycles). These are further defined by attributes which relate to branching, chain length, isolation of single rings, etc.

MARPAT takes a hierarchical approach to superatoms. This can be seen in Table III, groups used in query construction, which summarizes the different approaches of the two systems. For example, while M-DARC defines CHE as an alkenyl group, MARPAT uses Ak with the attribute UNS to cover both alkenyl and acetylenic groups. Even with MARPAT's superatom-attribute approach, there tend to be more narrow M-DARC superatoms categories which can be specified in M-DARC.

There are other superatoms that might have been included, but were not by either group. For example in MARPAT, a metallocyclobutane query will be translated into a saturated monocyclic heterocycle having one heteroatom (HY MON SAT LOQ). This means that most retrieved answers will not contain a metal. Perhaps "metallocycle" should be added to the superatom list.

Other differences are also highlighted in Table IV. MARPAT specifies only that groups are saturated or unsaturated and avoids the aromatic/nonaromatic distinction made by M-DARC. One vestige of Derwent's fragment coding system seems to be the aromatic designation as applied to



```
VAR G1      = 13 / 54 / Hy<EC (1-3) N, AN (1-) N, BD (1-) D, AC (1),
              RS (1) M5 (1) X7> (SO) / 75
VAR G2      = NH / 17
VAR G3      = O / S
VAR G4      = alkyl
VAR G5      = Hy<EC (5-) C (-1) O (1-2) N (-1) S (0) OTHERQ, AR (1-),
              AC (2-), RS (1-) E6 (1-) E7>
VAR G6      = H / alkyl (SO) / alkoxy / cycloalkyl<(3-8)> / OH / NO2 /
              CN / X / CHO / CO2H / 23 / 25 / 31 / 35 / 38 / 41 /
              alkenyl (SO) / A
VAR G7      = C(S) / C(O) / S(O) / SO2 / 32-112 33-24
VAR G8      = A / alkyl (SO) / aryl (SO) / heteroaryl (SO) /
              alkoxy (SO) / aryloxy (SO) / heteroaryloxy (SO)
VAR G9      = C(S) / C(O) / S(O) / SO2
VAR G10     = NH / 27
VAR G11     = alkyl / alkylcarbonyl
VAR G12     = C(O) / S(O) / SO2
VAR G13     = A / alkyl (SO) / aryl (SO) / heteroaryl (SO) / alkoxy (SO
VAR G14     = O / S
VAR G15     = OH / SH
VAR G16     = OH / NH2
VAR G20     = H / alkyl (SO) / A / alkenyl / NH2 / 47 / aryl (SO) /
              heteroaryl (SO)
VAR G22     = NH / 49
VAR G23     = Ak<BD (0) T> / alkylcarbonyl (SO X) /
              Cb<EC (6) C, AR (1-), BD (6) N, AC (1), RS (1) E6> (SO)
VAR G25     = Hy<EC (5-6) C (1-2) N (-1) O (-1) S (0) OTHERQ, AR (0),
              BD (ALL) SE, AC (1), RS (0-) E5 (0-) E6 (0) OTHER> (SO)
VAR G29     = Hy<EC (1-3) N, AN (1-) N, BD (1-) D, AC (1),
              RS (1) M5 (1) X7> (SO)
VAR G32     = H / alkyl (SO) / alkoxy / cycloalkyl<(3-8)> / OH / NO2 /
              CN / X / CHO / CO2H / 31 / 35 / 38 / 41 / alkenyl (SO) /
VAR G35     = H / OH / alkoxy / acyloxy
```

Figure 3. MARPAT retrieval for the indicated query, answer 1 of 65.

five-membered heterocycles like furan and thiophene. This is somewhat confusing since the bonding in these systems is not defined to be normalized, as one might have expected for aromatic systems.

For heterocycles, MARPAT uses the HIQ/LOQ attribute to distinguish those with exactly one heteroatom, while in M-DARC no query distinction is possible. While M-DARC has no HIQ/LOQ distinction, but systems have similar implementations of other attributes, as can be seen in Table IV. Aside from HIQ/LOQ, the major difference is in attributes related to chain carbon number. M-DARC uses a high, mid, and low count while MARPAT has only high and low.

SEARCH ALGORITHM

The search software in each case is based on the corresponding "exact" structure-search system, which involve a preliminary screening step followed by an atom-by-atom (superatom-by-superatom) search. Inherent in a Markush system is a substantially more complex and CPU-intensive atom-by-atom search. This explains the M-DARC's CPU time limits and pricing structure, which encourages users to forego this step.

M-DARC uses the screening process (RE) developed by Dubois,[6-10] based on what he calls FRELs. These are locally limited fragments defined about a central atom, branching out two levels. There is an inverted file of these FRELs, which are generated by algorithm. This contrasts the STN method which maintains a fixed screen list of structural features, each having a moderate level of appearance in the file.[11]

DARC is especially good at detecting unusual local features or sets of these features. By stopping at the RE stage, it is possible to retrieve similar answers which might differ in substructure from the query. For example, a RE answer might have a fragment (FREL) contained in a seven-membered ring, while the query had that feature in a six-membered ring. In this way a search can be expanded beyond a simple substructure search. In some instances, unanticipated but relevant answers can be retrieved.

Table IV. A Comparison of Superatom Attributes

MARPAT Generic Group Categories			M-DARC Attributes		
HIC	High Carbon	≥7 carbons	HI	High carbon	>10 carbons
----	----	----	MID	MID carbon	7–10 carbons
LOC	LOw Carbon	≤6 carbons	LO	LOw carbons	≤6 carbons
HIQ	High Q	≥2 heteroatoms	----	----	----
LOQ	LOw Q	1 heteroatom	----	----	----
SAT	SATurated	all bonds SE	----	----	----
UNS	UNSaturated	≥1 non-SE bonds	----	----	----
BRA	BRAnched	≥1 atoms in group bonded to ≥2 other atoms in group	BR	BRanched	SAME AS MARPAT
LIN	LINear	all atoms in group bonded to ≤2 atoms within group	STR	STRaight chain	SAME AS MARPAT
PCY	PolyCYclic ring system	a bond or atom is common to ≥2 rings	FU	FUsed	SAME AS MARPAT
MCY	MonoCYclic ring	no bond or atom is common to >1 ring	MON	MONocyclic ring	SAME AS MARPAT

The usefulness of the RE retrieval is mitigated, however, in M-DARC as the final RE list can be either larger or smaller than the apparent result of combining FREL lists. This appears in part to be due to lumping into the final RE answer certain structures with as yet ungenerated inverted FREL files. Despite this limitation, it is still worthwhile to examine the RE for similar answers of interest.

For the searcher, the advantage of the Registry File's screen system is the ability to manually input screens from the screen dictionary. This is not now permitted in MARPAT, which maintains a somewhat different screen set. Its strength is in rendering processable some queries which would otherwise exceed system limits. Unlike DARC, CAS screening is not useful in uncovering other substances having similar unusual structural features.

COMBINING STRUCTURE WITH CONCEPTS

MARPAT permits the combining of structure search answers (which correspond to CA accession numbers) with lists generated in the CA File. However, it is not possible to limit the retrieval by concept, such as CA section, etc. while in MARPAT. All concept searching must be done in CA and can be used only as an L# list.

M-DARC does permit retrieval to be limited to "File Segment". For Derwent's WPIM File it is possible to limit by Derwent class (but, "File segment-B" covers both Farmdoc and Agdoc) and by certain other structural and nonstructural concepts such as Grafted Polymer, Polypeptide, Registry, or Ordinary Chemical.

ANSWER DISPLAY

As can be seen from Figure 3 the MARPAT display, spread over three screens, can be tedious and time consuming to interpret. Although in some display formats the MSTR is highlighted, there is no indication as to which of the variables match the query structure. As was indicated above, it is quicker and easier to display part of the CA record with its graphical abstract (e.g., d cbib abs), as is possible in MARPAT.

One feature unique to the MARPAT display is that Markush structures specifically claimed (SC) are distinguished from those merely exemplified (EX). The utility of this feature is somewhat limited, however, in that most of the non-U.S. patents covered by CA are unexamined. Likewise, the lack of an integral patent family facility makes it difficult to follow compounds (or classes) through to granted patents.

M-DARC on the other hand (Figure 2) has query highlighting available (VI FO), though this is undoubtedly easier here where there is no query translation. Unfortunately, the answer display is still spread over multiple screens. For Derwent's WPIM File there is file segment display, indicating at least in which Derwent class the record appears (although it is not possible to distinguish Farmdoc from Agdoc records, both having File Segment-B). However there is no possible display of any other bibliographic data, all of which must be displayed in the EPI/L Files (which contain no graphical images).

SUGGESTED IMPROVEMENTS

A number of improvements could facilitate the use of these files.

MARPAT. As MARPAT is relatively slow, it would be used if background query processing were possible. For example, a mechanism might be arranged whereby one could begin execution of the query, disconnect, and log in 10 min or so later. Alternatively, one might like to switch to the CA or REG Files to do some related searching while MARPAT processing proceeded in the background.

Some limited text-search facility would be useful. Even the ability to limit to CA sections would be useful. The linking of a Markush structure with an activity term (preferably controlled vocabulary) would be useful. The concept of "Role Qualifiers" used in M-DARC (e.g., substance analyzed, catalyst, starting material) could be beneficially applied here. The process of interpreting answers would be expedited if the query "spin-offs" were displayable.

Some provision for searching stereochemistry would be useful. (It is now "display" only.)

M-DARC. A priority item for M-DARC is for query translation, for specifics to generics and vice versa. Any serious Markush search system must have this capability.

A facility for the translation of the CA bonding conventions into those of M-DARC would be immensely helpful to those of us familiar with the CA system. It would also assist in transferring queries from G-DARC to M-DARC.

There is currently no convenient way to specific a variable point of attachment. Attaching G-groups to each position is not satisfactory. Aside from being time consuming, it results in answers having multiple G-groups attached, when only one was desired.

The thoughtful integration of MARPAT into the STN system serves as a laudable example for M-DARC. There is no integration of M-DARC with G-DARC and Questel. Currently it is necessary to logout from Questel, then log into either M-DARC and G-DARC. Likewise it is not possible to easily switch to or from M-DARC to G-DARC. Aside from the use of the expensive and specialized PC package DARC ChemLink, there is no way to phrase a single query and search both M-DARC and G-DARC databases.

Some provision for the direct display of FREL lists would be more useful than the current system where RE display is somewhat unpredictable. An ability to explicitly display the FREL along with its resultant list would also be useful.

CONCLUSION

As criticism is easy, it is important to remember that the commercial implementation of MARPAT and M-DARC represent remarkable accomplishments in an uncharted frontier. Both systems will undoubtedly evolve as patent searchers test the limits of these systems.

Both MARPAT and M-DARC have advantages and disadvantages. The choice of which system to use will depend on the particular problem at hand. An important component in this decision is reported on in this issue by Kathy Cloutier;[12] and that component is database content and indexing policies.

It seems to be a part of the American personality trait to ask but which is the best? If I were pressed to decide which of the systems represents the "cutting edge", the conclusion seems inescapable that, at the present time MARPAT is the more mature product from a software viewpoint. This judgement is based primarily on the user-specifiable selective query translation, and MARPAT's integration with the other STN structure and bibliographic files.

REFERENCES AND NOTES

(1) Meurling, A. CAS Online and DARC: A comparison. *Database* **1990**, *Feb*, 54-63.

(2) Warr, W. A.; Wilkins, M. P. Graphics front ends for chemical searching and a look at Chemtalk Plus. *Online* **1990**, *May*, 50-4.

(3) Brueggman, P. Creating chemical structures for online searching with Molkick. *Database Searcher* **1989**, *May*, 22-7.

(4) *Workshop Manual*: *WPI*, *Markush DARC*, Feb 1989 ed.; Derwent Publications Limited: London, 1989.

(5) Fisanick, W. The Chemical Abstract's Service generic chemical (Markush) structure storage and retrieval capability. 1. Basic concepts. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 145-54.

(6) Dubois, J. E.; Laurent, D.; Viellard, H. DARC system. Polymatrix structural description. Writing of formal matrices. *C. R. Hebd. Seances Acad. Sci., Ser. C* **1966**, *263*, 1245-8.

(7) Dubois, J. E.; Laurent, D.; Viellard, H. System of documentation and automation of correlation researches. General principles. *C. R. Hebd. Seances Acad. Sci., Ser. C* **1966**, *263*, 764-7.

(8) Dubois, J. E.; Laurent, D. DARC (documentation and automation of correlation research) system. Population-correlation theory, organization, and description. *C. R. Acad. Sci., Paris, Ser. C* **1968**, *266*, 943-5.

(9) Dubois, J. E.; Mathieu, G.; Peguet, P.; Panaye, A.; Doucet, J. P. Simulation of infrared spectra: an infrared spectral simulation program (SIRS) which uses DARC topological substructures. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 290-302.

(10) Attias, R.; Dubois, J. E. Substructure systems: concepts and classifications. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 2-7.

(11) 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.

(12) Cloutier, K. A Comparison of Three Markush Databases. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 40-44.

The PHARMSEARCH Database[†]

MICHAEL P. O'HARA*

Questel, Inc., 5201 Leesburg Pike, Suite 603, Falls Church, Virginia 22041

CATHERINE PAGIS

INPI, 26 Bis rue de Leningrad, 75008 Paris, France

Received December 19, 1990

PHARMSEARCH, a database produced by the French Patent and Trademark Office (INPI), covers pharmaceutical patents issued by the European, French, and United States patent offices from November 1986 onward. PHARMSEARCH is composed of MPHARM, a structure file searchable using Markush DARC software, and PHARM, the companion bibliographic file. Markush structures claimed in the patent documents are entered into the database as variable generic structures. Specific structures are also included in the database, when they are not part of a Markush structure in the patent document. Chemical index terms describe all moieties of the structure. Indexing also describes the therapeutic activities and preparation processes for the compounds. The indexing policies used in the production of this database are described.

INTRODUCTION

PHARMSEARCH is a pharmaceutical patents database which covers pharmaceutical patents issued by the European, French, and United States patent offices. PHARMSEARCH actually consists of two files: MPHARM, which is a Markush structure file, and PHARM, which is a companion bibliographic file. PHARMSEARCH originally began as a publication in paper form. In 1983, it was acquired by INPI, the French Patent and Trademark Office. After acquiring PHARMSEARCH, INPI began a study to develop automated processes for the production of the publication. As a result of this study, INPI decided to use a graphics structure database to index and record the chemical structures which occur in the pharmaceutical patents. In 1984, INPI joined the Markush DARC development that was already under way at Telesystèmes, the parent of Questel. The actual building of the graphics database began in 1986. The online database was opened to the public in January 1989, making PHARMSEARCH the first Markush structure database to become available for online searching. The searching methodology for Markush structure systems is described in the paper by John Barnard in this issue.[1] This paper focuses on the indexing considerations used in the preparation of this unique database.

PATENT INDEXING CONSIDERATIONS

When preparing indexing rules for patents databases there are two considerations for which the rules must account: the levels of information which are contained in the actual patent documents and the patent law requirements. There are three levels of information in patent documents: a general level, a preferred level, and a specific level. In the actual patent documents, the general level of information is contained in the description and in the independent claims; the preferred level and the specific level of information are described in the description and in the dependent claims. Of these three levels of information in the patent documents, patent law requires consideration of the general and the specific levels of information.

Pharmaceutical patents generally are searched for two basic types of information: the chemical compounds involved and

[†] Presented at the Division of Chemical Information, 200th National Meeting of the American Chemical Society, Aug 29, 1990.

[*] Author to whom correspondence should be addressed.

0095-2338/91/1631-0059$02.50/0