The Chemical Abstracts Service Generic Chemical (Markush) Structure Storage and Retrieval Capability. 2. The MARPAT File†

TOMMY EBE,* KAREN A. SANDERSON, and PATRICIA S. WILSON

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

Received November 6, 1990

The MARPAT File is a structure-searchable database containing Markush structures found in patents abstracted in *Chemical Abstracts*. The file was introduced in late 1989 on STN International, the Scientific and Technical Information Network, and access was extended gradually during 1990 concurrent with implementation of service and support enhancements. The file became available to all STN accounts on September 30, 1990. The MARPAT File offers online access via structure-based queries to the generic substance representations found in the chemical patent literature. A single query language is used to build structure queries that are searchable in any STN structure file. A unique capability permits searcher-specified cross-matching of specific query nodes with generic nodes included in the MARPAT File structures.

INTRODUCTION

The development of computer systems for the storage and retrieval of topological representations of generic (Markush) chemical structures in patents has been of special interest to researchers in recent years. Many approaches for handling Markush structures were presented at a 1984 conference on this subject held in Sheffield, England,[1] and at a 1986 American Chemical Society symposium in Anaheim, CA. Barnard summarized the latter presentations in 1987.[2]

Chemical Abstracts Service (CAS) began research in the early 1980s on topologically based techniques for handling Markush structures, leading to U.S. Patent 4,642,762 covering the overall design and search strategies that were developed.[3] A previous paper presented the basic concepts of the CAS generic substance handling capability.[4] This paper presents the database content and the features available for searching and displaying retrieved information in the MARPAT File. Additional papers on the file substance description language and search techniques and capabilities are planned.

COVERAGE

The new Markush structure search service developed by CAS is called the MARPAT File. Until 1989, there had been two structure-searchable files on STN: the REGISTRY File and the BEILSTEIN File, each of which contains *specific* structures, i.e., each structure represents exactly one chemical substance (Figure 1). With the introduction of the MARPAT File, CAS has enhanced and extended the powerful structure-search capability to access the chemical patent literature directly via the *generic* and *prophetic* substances represented by the Markush structures presented in chemical patents (Figure 2). Markush structures, as opposed to specific structures, use variables to represent more than one chemical substance. Variables define both finite sets of specific substances and infinite sets with the use of generic terms, e.g., "alkyl", "N-containing heterocycle". Many of these substances are not prepared or tested, but only hypothesized to be possible based on the claimed technology, hence the term "prophetic".

The MARPAT File covers the Markush structures presented in patents that have publication dates of January 1, 1988, or later and that are covered as basics in any of the 80 sections of *Chemical Abstracts*. A CA basic patent is the family member first received by CAS. This includes patents from 26 countries and two international organizations. When widely released in September 1990, MARPAT contained more than 21 000 citations and more than 65 000 Markush structures. Each biweekly update offers access to approximately 500 new chemical patents containing about 1500 new Markush structures.

Markush structures presented in the patent claims are routinely included. Markush structures from the patent disclosures are included if there is no Markush structure in the claims or if the Markush structure presented in the disclosure is broader than any presented in the claims. MARPAT contains Markush structures that are representations of organic or organometallic molecules. Alloys, metal oxides, inorganic salts, intermetallics, and polymers are not included at this time. Future enhancements will extend file coverage to include polymers.

COMPLEMENTARY WITH REGISTRY AND CA FILES

The MARPAT and REGISTRY Files are complementary in several ways. The specific substances reported in the chemical literature are contained in the REGISTRY File, while the MARPAT File contains the generic and prophetic structures that are found in the chemical patent literature. These two files are also complementary in that they use the same STRUCTURE command to build the structure query, which may then be used in any of the structure-searchable files on STN. There is no need to rebuild the structure to search in a different file. All of the attributes needed to search MARPAT may be included when the structure is initially created. Structure query attributes needed for MARPAT but not for REGISTRY are simply ignored in REGISTRY.

The MARPAT and the CA Files are also complementary. They are both bibliographic document-based files and share the same file accession number, the CA abstract number. This makes it particularly convenient for transferring answer sets between MARPAT and CA. One reason for transferring a MARPAT answer set might be to further narrow the scope of the answers by combining the answer set with text terms and searching in the CA File. A reason for transferring a CA answer set might be to display in MARPAT the Markush structures for patent citations that were retrieved by a text search in the CA File.

QUERY STRUCTURE-BUILDING FEATURES

Some of the structure-building features that CAS developed for searching MARPAT are equally valuable in searching the other structure-searchable files on STN. Most of these features became available late in 1989. The variable point of attachment of substituents on a ring system allows the con-

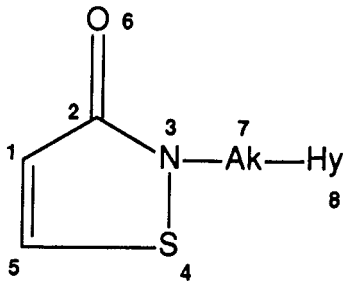
† Presented at the 200th National Meeting of the American Chemical Society, Washington, DC, Aug 29, 1990.

0095-2338/91/1631-0031$02.50/0

Chart I

```
=> file registry
FILE 'REGISTRY' ENTERED AT 15:42:50 ON 09 JUL 90
 ...

=> str
 ...                                [Build the query structure.]
:dis sia
```




```
NODE ATTRIBUTES:
MLEVEL  IS ATOM    AT   1       [The ring nodes are set at MLEVEL ATOM to    ]
MLEVEL  IS ATOM    AT   2       [match only Markush structures that contain  ]
MLEVEL  IS ATOM    AT   3       [specific (real) isothiazolone rings, and not]
MLEVEL  IS ATOM    AT   4       [generically equivalent HY nodes.  MLEVEL has]
MLEVEL  IS ATOM    AT   5       [no effect in the REGISTRY File.             ]
GGCAT   IS LOC     AT   7
GGCAT   IS SAT     AT   7
GGCAT   IS MCY     AT   8
GGCAT   IS LOQ     AT   8

GRAPH ATTRIBUTES:
RSPEC I
NUMBER OF NODES IS   8
:end

L1              STR

=> s l1 ful sss                    [Search the query, L1.]
FULL SEARCH INITIATED  15:43:26
FULL SCREEN SEARCH COMPLETED -   1106 TO ITERATE
 94.3% PROCESSED    1043 ITERATIONS            6 ANSWERS
100.0% PROCESSED    1106 ITERATIONS            6 ANSWERS
SEARCH TIME: 00.00.26

L2            6 SEA SSS FUL L1

=> d l2 1-6 rn cn                  [Display the CAS Registry Numbers® and]
                                   [chemical names for the six specific  ]
                                   [substances that were retrieved.      ]

L2  ANSWER 1 OF 6
COPYRIGHT (C) 1990 AMERICAN CHEMICAL SOCIETY

RN   113526-86-6
CN   3(2H)-Isothiazolone, 5-benzoyl-2-(3-pyridinylmethyl)-
     (9CI)  (CA INDEX NAME)

L2  ANSWER 2 OF 6
COPYRIGHT (C) 1990 AMERICAN CHEMICAL SOCIETY

RN   113526-85-5
CN   3(2H)-Isothiazolone,5-benzoyl- 2-(2-pyridinylmethyl)-
     (9CI)  (CA INDEX NAME)

L2  ANSWER 3 OF 6
COPYRIGHT (C) 1990 AMERICAN CHEMICAL SOCIETY

RN   113526-84-4
CN   3(2H)-Isothiazolone, 5-benzoyl-2-(2-thienylmethyl)-
     (9CI)  (CA INDEX NAME)

L2  ANSWER 4 OF 6
COPYRIGHT (C) 1990 AMERICAN CHEMICAL SOCIETY

RN   113526-83-3
CN   3(2H)-Isothiazolone, 5-benzoyl-2-(2-furanylmethyl)-
     (9CI)  (CA INDEX NAME)
```

Chart I (*Continued*)

```
L2  ANSWER 5 OF 6
COPYRIGHT (C) 1990 AMERICAN CHEMICAL SOCIETY

RN   68944-05-8
CN   3(2H)-Isothiazolone, 2-[1-(2-oxo-1-pyrrolidinyl)ethyl]-
     (9CI)  (CA INDEX NAME)

L2  ANSWER 6 OF 6
COPYRIGHT (C) 1990 AMERICAN CHEMICAL SOCIETY

RN   68943-96-4
CN   3(2H)-Isothiazolone, 2-(1-piperidinylmethyl)-
     (9CI)  (CA INDEX NAME)

=> file ca
FILE 'CA' ENTERED AT 15:44:43 ON 09 JUL 90
 ...

=> s l2                          [Search the Registry Numbers in the REGISTRY]
L3           2 L2                [answer set (L2) in the CA File to retrieve ]
                                 [the citations that reference them.         ]

=> d l3 1-2 an ti pi             [Display the title and patent information   ]
                                 [for the two citations that were retrieved. ]

L3  ANSWER 1 OF 2
COPYRIGHT (C) 1990 AMERICAN CHEMICAL SOCIETY

AN   CA108(17):150464s
TI   Preparation of 5-acyl-3(2H)-isothiazolones as fungicides
     and plant growth regulators
PI   EP 246735  A1  25 Nov 1987

L3  ANSWER 2 OF 2                                [Both of these patents were  ]
COPYRIGHT (C) 1990 AMERICAN CHEMICAL SOCIETY     [published prior to the start]
                                                 [of MARPAT coverage.         ]
AN   CA90(7):54934j
TI   3-Isothiazolones as biocides
PI   US 4105431  8 Aug 1978

=> file marpat
FILE 'MARPAT' ENTERED AT 15:45:24 ON 09 JUL 90
 ...

=> s l2 ful sss                  [Search the structure query contained in the]
                                 [REGISTRY answer set (L2) in MARPAT.         ]
FULL SEARCH INITIATED  15:45:50

FULL SCREEN SEARCH COMPLETED -     47 TO ITERATE
 93.6% PROCESSED      44 ITERATIONS            6 ANSWERS
100.0% PROCESSED      47 ITERATIONS            6 ANSWERS
SEARCH TIME: 00.00.21

L4            6 SEA SSS FUL L1

=> d 1-6 ti cs                   [Display the title and corporate source for ]
                                 [the six citations that were retrieved.     ]
L4  ANSWER 1 OF 6
COPYRIGHT (C) 1990 AMERICAN CHEMICAL SOCIETY

TI   Preparation and testing of isothiazolone 1,1-dioxide
     derivatives with psychotropic activity
CS   American Home Products Corp.

L4  ANSWER 2 OF 6
COPYRIGHT (C) 1990 AMERICAN CHEMICAL SOCIETY

TI   Epoxide stabilizers for isothiazolone microbiocides
CS   Rohm and Haas Co.

L4  ANSWER 3 OF 6
COPYRIGHT (C) 1990 AMERICAN CHEMICAL SOCIETY

TI   Heterocyclic acyl sulfonamides useful as herbicides and
     plant growth regulants, and their compositions and use
CS   du Pont de Nemours, E. I., and Co.

L4  ANSWER 4 OF 6
COPYRIGHT (C) 1990 AMERICAN CHEMICAL SOCIETY
```

Chart I (*Continued*)

```
TI   Preparation of isothiazolones as antibacterial agents
CS   Rohm and Haas Co.

L4  ANSWER 5 OF 6
COPYRIGHT (C) 1990 AMERICAN CHEMICAL SOCIETY

TI   Ortho ester-stabilized isothiazolone compositions
CS   Rohm and Haas Co.

L4  ANSWER 6 OF 6
COPYRIGHT (C) 1990 AMERICAN CHEMICAL SOCIETY

TI   Wood preservatives containing transition metal
     carboxylates and isothiazolones
CS   Mooney Chemicals, Inc.
```

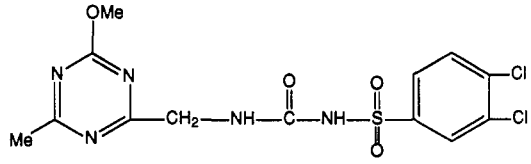
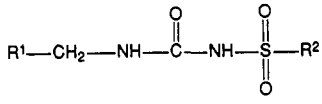


Figure 1. Specific structures in the REGISTRY and BEILSTEIN Files.



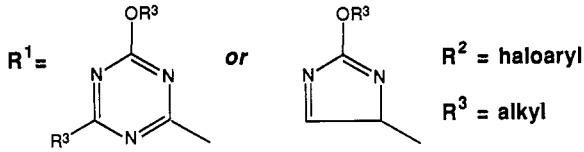


Figure 2. Generic (Markush) structures in patents.

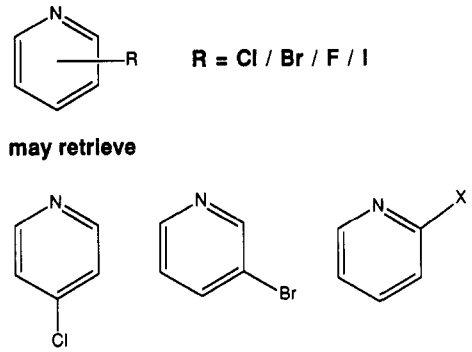


Figure 3. Variable points of attachment (VPA).

struction of a single structure query that may retrieve any of the possible positional isomers (Figure 3). This eliminates the need to build separate structure queries to retrieve such isomers.

Another new structure-building feature is a fourth generic node to complement the three cyclic nodes that were introduced in 1985. These cyclic generic nodes are: CY, CB, and HY, representing cyclic, carbocyclic, and heterocyclic rings, respectively. The newly introduced generic node is AK, representing any carbon chain. The AK node may be used to retrieve generic as well as specific carbon chains, e.g., alkyl, methyl, octyl (Figure 4).

Generic group categories (GGCs) were also developed to provide some precision for the generic nodes used in structure queries. Five pairs of GGC values have been defined and may be optionally specified to restrict the scope of a generic node (Figure 5). An HY node, for example, may be qualified by the specification of several GGC values to retrieve only answers

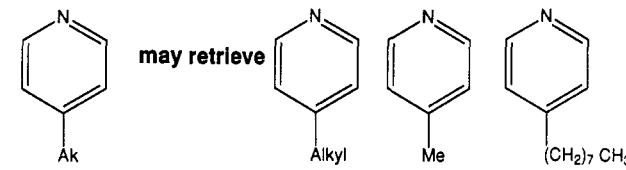


Figure 4. AK generic node.

BRA / LIN	Branched vs. linear
SAT / UNS	Saturated vs. unsaturated
LOC / HIC	1 - 6 carbons vs. 7 or more
LOQ / HIQ	1 heteroatom vs. 2 or more
MCY / PCY	Monocyclic vs. polycyclic

Figure 5. Generic group categories (GGCs).

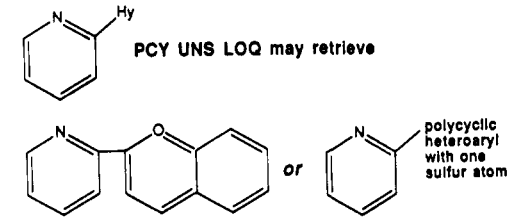


Figure 6. HY restricted by GGC values.

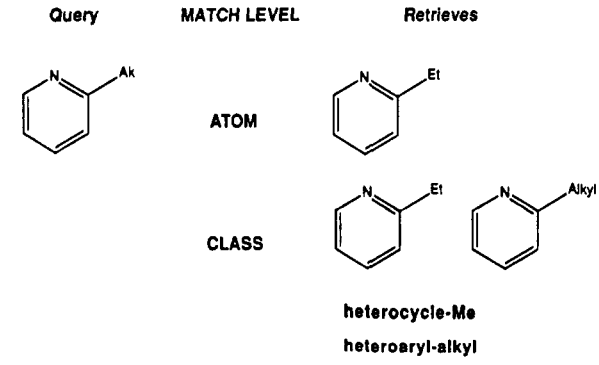


Figure 7. Match level (MLE).

with polycyclic, unsaturated, heterocyclic ring systems with a single hetero atom (Figure 6).

MATCH LEVEL

The ability to specify match level (MLE) is unique to MARPAT and is the heart of searching this Markush file. To understand how it works, consider an AK node at the ortho position of a pyridine ring (Figure 7). MLE CLASS allows the query nodes to match both the real atoms and the equivalent generic nodes in the file structures, e.g., a pyridine

query may retrieve both pyridine and heterocycle in file structures, and an AK may retrieve ethyl, methyl, etc., as well as alkyl in file structures. If MLE CLASS is assigned to all the pyridine ring nodes and to the AK node, the query would not only retrieve the 2-ethyl- and the 2-alkylpyridines, it would also retrieve broader generically equivalent heterocycles, e.g., heterocycle substituted with methyl, heteroaryl substituted with alkyl.

MLE ATOM allows the specification of some or all of the query nodes to match only the real atoms in file structures, e.g., a pyridine query may retrieve only pyridine and an AK may retrieve only ethyl, methyl, etc. in file structures. If MLE ATOM is assigned to all the query nodes, the query would retrieve such specific-atom matches as 2-ethylpyridine. It would also retrieve methyl, propyl, or any real atom carbon-chain substituent at that position. If MLE ATOM is assigned to the pyridine ring nodes and MLE CLASS is assigned to the AK node, then the query would retrieve not only the specific-atom matches, it would also retrieve encompassing generic groups for the AK, e.g., alkyl.

This translation capability, when used in conjunction with the generic group categories, provides searchers with a powerful tool for controlling the precision of their structure search queries.

SEARCH AND DISPLAY OPTIONS

The search options for MARPAT include all of those available in the REGISTRY File: SAMPLE, FULL, RANGE, and SUBSET. Subset searching is particularly useful in the MARPAT File. Broad MLE CLASS answer sets can be easily and quickly refined to more precise, smaller answer sets by performing SUBSET searches using modified queries, e.g., with MLE ATOM specified for some of the query nodes.

The search types that are offered for MARPAT searches include the SubStructure Search (SSS), as well as the new Closed Substructure Search (CSS). The CSS version performs a substructure search that considers all the query nodes to be closed to any further substitution unless a position was specified to be open to substitution. This capability is also quite useful in the REGISTRY File. BATCH search, performed offline, is available for those queries that require more time, or if the searcher does not want to spend time online waiting for the search to complete.

The Markush structures that are available in MARPAT may be displayed on both of the terminal types available for searching on STN, i.e., the graphics terminal (the type 2), as well as the ASCII-text terminal (the type 3). The graphics output for Markush structures is used for offline prints. In addition to the Markush structures, a searcher can display online or print offline any of the CA File data (bibliographic, abstract, or indexing information) for any of the answers retrieved in MARPAT without having to leave the file. A future system enhancement will permit text searching of this CA File data directly in MARPAT.

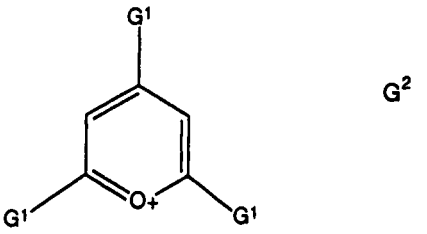
An online display of a MARPAT record might include some bibliographic data, as well as the hit Markush structure (Figure 8). The SAMPLE predefined display format includes the title of the patent document, as well as the HIT Markush structure. HIT displays the one Markush structure (of possibly several) for a patent citation that was matched with the structure query. The display also indicates whether the Markush structure was found in the claims or in the disclosure of the patent.

SEARCH SUPPORT

The support available through STN for MARPAT includes a learning file, LMARPAT. Containing about 500 documents,

```
TI   Preparation of arylbenzenes, useful as intermediate

  ***MSTR 2***
```




```
VAR G1     = aryl (SO) / (EX Ph (SO G3))
VAR G2     = R<TX "anion", CH (1) -> / (EX perchlorate)
VAR G3     = Me / OMe / Cl / Br / NO2

MPL:         claim 1
```

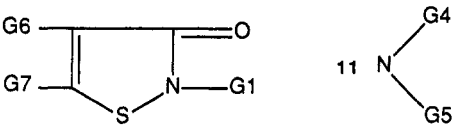
Figure 8. SAMPLE display format in MARPAT.

Chart II

```
=> d 4 hit              (Display the hit Markush structure for answer 4.)

L4  ANSWER 4 OF 6
COPYRIGHT (C) 1990 AMERICAN CHEMICAL SOCIETY

  ***MSTR 1***
```




```
VAR G1 = alkyl (SO G2) / aryl (SO G3) / cycloalkyl (SO) /
         aryl (SR alkyl) / (SC octyl / cyclohexyl)
VAR G2 = OH / X / CN / alkylamino / dialkylamino / arylamino /
         CO2H / alkoxycarbonyl / alkoxy (SO X) / aryloxy /
         alkylthio / arylthio / isothiazolyl / OCONH2 /
         morpholino / piperidino / pyrrolidino
VAR G3 = X / CN / NO2 / alkyl<(1-4)> / alkoxy<(1-4)> / 11 /
         alkoxycarbonyl<(1-4)> / SO2NH2
VAR G4 = alkyl<(1-4)>
VAR G5 = acyl
VAR G6 = H / X / alkyl / (SC Cl)
VAR G7 = H / X / (SC Cl)

MPL:     claim 1
```

LMARPAT can be searched at a low-connect hour rate, with no additional fees for search or display. All the MARPAT capabilities may be learned and practiced in LMARPAT. A User Guide containing more than 100 pages includes many search examples and practice problems. Both the Macintosh and the DOS versions of STN Express fully support all of the features needed to construct structure queries offline to be searched in MARPAT.

ILLUSTRATIVE SEARCH EXAMPLE

The following search example (see Chart I) illustrates a structure query built and searched in the REGISTRY File that retrieves six specific substances. The REGISTRY File answer set is then transferred to and searched in the CA File to retrieve two citations that reference one or more of these specific substances. The structure query is then transferred to the MARPAT File and is directly search, unchanged, to retrieve six additional patent document citations containing Markush structures that match the structure query. These six patents are not retrieved by the REGISTRY–CA searches because the MARPAT matches are with the generic or prophetic substances represented by the Markush structure, none of which are covered in the REGISTRY and CA Files.

The two Gk fragments that were matched by the system with the structure query are highlighted in the answer display in Chart II. Highlighting of hit Gk fragments is being developed for a future system enhancement. The required isothiazolone ring is found in the base structure. The alkyl alternative in G1 matches the query AK. Because G1 is optionally substituted (SO) by G2, the piperidino alternative

for G2 matches the query HY.

CONCLUSION

For many years, searchers have been asking CAS to offer access to the generic and prophetic chemistry represented by Markush structures in patents, i.e., to those chemical substances that were not specifically prepared or claimed and therefore are not in files of specific substances such as the REGISTRY File. MARPAT now offers convenient and powerful structure-based access to the prophetic and the generic substances presented in the chemical patent literature. Only one structure query need be constructed to search MARPAT or any of the other structure-searchable files on STN. MARPAT automatically matches the generic nodes and specific atoms in the query with the file structures and offers the user an option to control the level of specificity. All of the Markush structures available in MARPAT may be displayed online or printed offline with complete or selected portions of the CA File data for those answers. Significant enhancements are now under development in the areas of search response time, hit fragment highlighting, search precision, and query formulation. The file is updated biweekly and automatic current awareness searches are available to keep up with the current patent literature. Quality STN technical support is available to help searchers as needed. Searchers interested in comprehensive access to chemical patent information should consider performing MARPAT searches in the course of their normal structure searching efforts.

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) Fisanick, W. Storage and Retrieval of Generic Chemical Structure Representations. U.S. Patent 4,642,762, Feb 10, 1987.
(4) Fisanick, W. The Chemical Abstracts Service Generic Chemical (Markush) Structure Storage and Retrieval Capability. 1. Basic Concepts. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 145-154.

Searching for Simple Generic Structures†

ROBERT N. WILKE

Amoco Research Center, Amoco Corporation, P.O. Box 3011, Naperville, Illinois 60566

Received November 14, 1990

Markush structures representing millions of possible compounds have been indexed by using the Markush DARC system; the new MARPAT system offers similar capabilities. However, these files and their simpler versions (the CAS Registry system and the Generic DARC system) cannot easily provide the searcher with a searchable list of the many possible compounds represented by a simple generic structure. Examples of generic searches and how they are handled in the various structure files will be discussed. Comparisons between these files and the bibliographic files containing generic structure information will be shown, and some recommendations for the future will be examined.

INTRODUCTION

A lot of discussion has centered around the various methods of searching for chemical structures in patents and in the chemical literature. Papers and talks have been given on how the new substructure searching systems that employ connection tables are far superior to those that use codes to represent chemical fragments.[1-9] These substructure systems are shown to give very precise and mostly complete retrieval of complex chemical structures. With the recent introduction of MARPAT and the introduction of the Markush DARC system in 1989, we now have two substructure systems capable of retrieving the complex generic structures that are often found in patents. However, little or nothing has been said about how these search systems work in retrieving information about simple generic structures. One may wonder why anyone would want to search for simple generic chemical structures; but these searches are very common in companies that produce commodity chemicals and monomers, companies that are process oriented.

It is very difficult in connection table based systems to search for simple generic structures, which may contain only a single functional group. The problems that arise in searching these simple generic compounds are due to the large number of possible answers. For example, in searching for a new preparation of aliphatic diamines, any aliphatic diamine that is the product of a reaction could potentially have been prepared by this new method. So to obtain as complete recall as possible all aliphatic diamine containing products will have to be retrieved and then further qualified with the new reaction conditions and or starting materials.

PROCEDURES

In order to investigate how simple generic structures can be searched, the capabilities of six different structure searching systems were studied. These systems were the API Chemical Aspects system, the Chemical Abstracts Registry system, the MARPAT File, Derwent's CPI Chemical Fragmentation, the IFI Chemical Fragmentation, and the Markush DARC system. The Generic DARC system and the Beilstein File on STN were not covered in these examples, simply to cut down on the amount of data. These files consist of three fragmentation-based systems and three substructure-based systems. The first topic will discuss why they can still be searched in fragmentation-based databases. The second topic to be covered will deal with the problems associated with searching for these simple generic chemical substances in the substructure databases; the final topic will be on what the future may bring. To make things easier, databases where the structure is defined

† Presented at the 200th National Meeting of the American Chemical Society, Washington, DC, August 1990, Symposium on Markush Structure Files and Searching, CINF 32.

0095-2338/91/1631-0036$02.50/0