SRU's. The Central Patent Index and Dr. Alderson devised earlier systems for retrieving families of polymers through the features of their monomers or individual SRU's.

LITERATURE CITED

(1) Cahn, R. S., and Ingold, C. K., "Specification of Configuration about Quadricovalent Asymmetric Atoms," *J. Chem. Soc.*, 612 (1951).
(2) Committee on Nomenclature, American Chemical Society Division of Polymer Chemistry, "A Structure-Based Nomenclature for Linear Polymers," *Macromolecules*, **1**, 193 (1968).
(3) Montague, B. A., and Schirmer, R. F., "Du Pont Central Report Index: System Design, Operation, and Performance," *J. Chem. Doc.* **8**, 33 (1968).
(4) Schultz, J. L., "Handling Chemical Information in the Du Pont Central Report Index," *J. Chem. Doc.*, **14**, 171 (1974).

A Unique Chemical Fragmentation System for Indexing Patent Literature†

MARY Z. BALENT* and JANE M. EMBERGER

IFI/Plenum Data Company,** Wilmington, Delaware 19808

Received November 5, 1974

A new adaptation of a chemical fragmentation system provides a unique procedure for indexing and searching the specific chemicals, classes of compounds, and Markush structures found in patent literature. This computer-based system employs a POSSIBLE and MUST approach which allows generic structures to be searched with a minimum of false retrieval. The data base includes over 300,000 chemical and chemically related patents. Searches can be structured using the fragmentation system alone or in conjunction with general terms, compound terms, assignees, and U.S. Patent Office class codes.

INTRODUCTION

The IFI fragmentation system for organic chemicals is an adaptation of a scheme developed at Du Pont's Central Research Department in the late 1950's.[1] It was first used for patent literature by Du Pont's Central Patent Index in 1964.[2] In early 1972, the IFI/Plenum Data Division of Plenum Publishing Corporation purchased the rights to the Du Pont Company's system for machine retrieval of patent information, including the fragmentation system. During the remainder of that year, IFI successfully integrated the Du Pont system and data base into IFI's Comprehensive Index to U.S. Chemical Patents.[3]

This computer-based fragmentation system employs a POSSIBLE and MUST approach which, we believe, provides a unique method especially suited for indexing and searching the classes of compounds and Markush structures found in the patent literature. Also, since fragmentation is only a part of the total IFI index, the information scientist has the ability to pinpoint chemical information by utilizing fragments in conjunction with general terms, chemical terms, assignees, and/or United States Patent Office class codes.

FRAGMENTATION—DESCRIPTION

The IFI fragmentation system uses systematic rules and a controlled open-ended vocabulary. Chemical compounds are indexed in terms of substructural pieces that characterize them. The indexing terms or fragments are grouped into four categories:

1. atoms present terms
2. functional group terms
3. ring terms
4. configuration terms

Atoms present terms describe the number of carbon atoms and any specific halogen and metal atoms included in a compound. The 5-8 CARBONS and CHLORINE shown in Figure 1 are examples of atoms present terms.

Functional groups (FG's) are atoms or groups of atoms which characterize classes of compounds to which an indexed structure belongs. The system has a unique procedure for defining functional groups and allows for the introduction of new structures in a systematic and straightforward fashion. Examples of functional groups are the C to C DOUBLE BOND, HYDROXY, SULFONAMIDE, and HYDRAZIDE shown in Figure 1.

Functional groups are found in the fragment vocabulary list using atom counts. A name or linear structure is used to identify specific functional groups. For example, in Figure 1, the sulfonamide group is listed as NO2S SULFONAMIDE FG and the hydrazide as CN2O followed by a linear notation.

Ring systems are indexed by ring-type terms (ACYCLIC, CARBOCYCLIC (CARBO), HETEROCYCLIC (HETERO), FUSED OR BRIDGED), by degree of ring unsaturation terms (NO unsaturation, PARTIAL unsaturation, MAXIMUM unsaturation), by the number of ring units in a structure (ONE, TWO, THREE, FOUR or more) and by specific ring structure terms. The ring structure terms describe the skeletons of rings and follow the nomenclature established in "The Ring Index" by A. M. Patterson, L. T. Capell, and D. F. Walker. Their arrangement in the fragment vocabulary list is similar to that of "The Ring Index." A Roman numeral, used to denote the number of individual rings, is followed by a ring formula and a name. The benzothiophene ring in Figure 1 is described by HETEROCY-

† Presented before the Division of Chemical Literature, 168th National Meeting of the American Chemical Society, Atlantic City, N. J., Sept 10, 1974.

* Author to whom correspondence should be addressed.
** A division of Plenum Publishing Co., 227 W. 17th St., New York, N. Y. 10011.

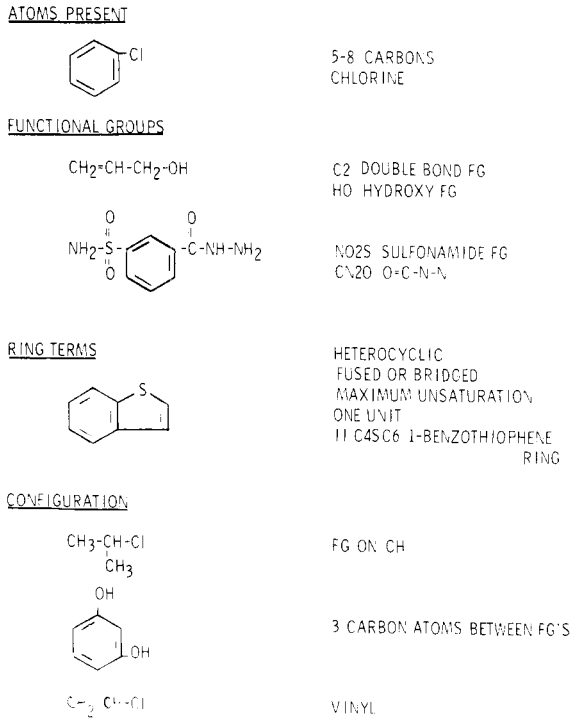


Figure 1. Fragment term categories.

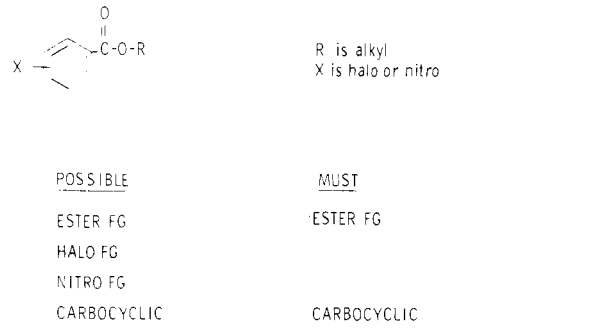


Figure 2. Sample Markush indexing.

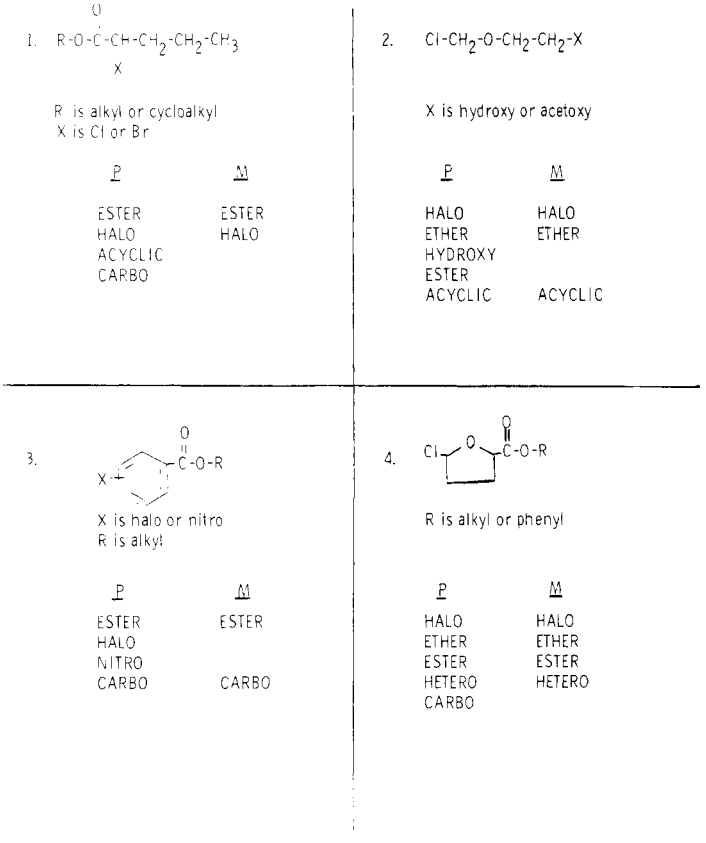


Figure 3. Demonstration file.

CLIC, FUSED OR BRIDGED, MAXIMUM unsaturation, ONE unit, and the ring structure term II C4SC6 1-BENZOTHIOPHENE RING.

Configuration terms are used to describe the types of carbon atoms to which functional groups are attached and to show relationships between two functional groups. For example, the term FG ON CH in Figure 1 indicates that a functional group (in this case a HALOGEN FG) is attached to an aliphatic carbon with one hydrogen. Resorcinol in Figure 1 is indexed by the configuration term THREE showing that there are three carbon atoms between the two HYDROXY FG's. The terms VINYL and ALPHA, BETA are used to show relationships between functional groups and C to C unsaturation.

All fragment terms, *i.e.*, atoms present terms, functional group terms, ring terms, and configuration terms, are located in alphabetical order in the fragment vocabulary list. Functional group terms and ring structure terms are preceded by the letters F and R, respectively, to list them together.

A detailed manual of instructions provides definitions and examples to explain the use of the fragmentation system.

MUST/POSSIBLE SCHEME

To improve the selectivity of the system, particularly with respect to the indexing of generic and Markush structures, fragment terms are further distinguished as POSSIBLE (P) fragments and MUST (M) fragments.

When indexing a Markush structure, all fragments applicable to any member of the family are indexed by the appropriate POSSIBLE terms.

Using only functional group and ring type terms, the Markush structure in Figure 2 would be indexed with the POSSIBLE terms for CARBOXYLIC ESTER FG, HALOGEN FG, NITRO FG, and CARBOCYCLIC RING.

In addition to the POSSIBLE indexing, each fragment which is present in *all* members of a Markush family is also coded with its corresponding MUST term. In Figure 2 ESTER FG and CARBOCYCLIC RING would also be coded as MUST terms.

When searching the IFI fragment file, POSSIBLE fragment terms are searched in a positive mode (*i.e.*, these are fragments which the retrieved compound should have) while MUST fragment terms are searched in a negative mode (*i.e.*, these are fragments which the retrieved compound should not have).

The effectiveness of the MUST and POSSIBLE scheme may be best shown by example. For demonstration purposes, let us establish a file which contains only the following fragment terms:

ESTER FG	ACYCLIC
ETHER FG	CARBOCYCLIC
HALO FG	HETEROCYCLIC
HYDROXY FG	
NITRO FG	

The compounds in Figure 3 are fragmented in the demonstration file using the terms shown beneath each compound.

Now let us pose a series of questions to the demonstration file. First let us ask for any halo ester. Compounds 1 thru 4 are answers since they all have indexing to the possible fragments ESTER FG and HALO FG. If we restrict our

	TOTAL FILE SEARCH		BROAD SUB CLASS SEARCH		SPECIFIC SUB CLASS SEARCH	
	USPO	IFI	USPO	IFI	USPO	IFI
No. 1	141,000	29	1696	1	324	1
No. 2	141,000	249	1363	68	72	28
No. 3	141,000	11	1329	10	746	10
No. 4	141,000	5	1070	1	205	1
No. 5	141,000	336	542	52	61	21
No. 6	141,000	2	1625	2	165	2
No. 7	141,000	243	662	17	481	17
TOTAL		875	8287	151	2054	80
AVERAGE		125	1184	22	293	11

Figure 4. Specificity—USPO *vs.* IFI.

search to all acyclic and carbocyclic halo esters by including in the search logic "but not MUST HETERO," then compound 4 is eliminated as an answer. It has been indexed to the must fragment HETERO. If we go a step further and request all acyclic halo esters by including in the search logic "but not MUST HETERO and not MUST CARBO," then only compounds 1 and 2 answer our question. Compounds 3 and 4 have indexing to the must fragments CARBO or HETERO and are therefore eliminated as answers. If we would like to be more precise, we can request all acyclic halo esters with a further restriction that the answer compound should not have an ether group. Compound 1 then becomes the only answer to this question.

APPLICATION

In the practical application of the fragmentation system, the IFI Comprehensive Index has the ability to zero in on structural information still further. A unique system of links and roles is applied to fragmented structures. Links are used to prevent false correlation between compounds in the same patent. This prevents a searcher from retrieving, for example, a patent indexed by compounds 2 and 3 (Figure 3) if he were searching for nitro-substituted ethers. Roles, which are two-digit mode indicators, describe the function of a fragmented compound as either present, reactant, or product. If desired, the information chemist can request structural information in conjunction with general terms, with compound terms, with assignees, and/or with U.S. Patent Office class codes. The general term file includes references to natural materials, tradenames, reactions, uses, polymer class terms, and nonstructurable chemicals. The compound term file includes references to over 12,000 specific compounds. Each specific compound is fragmented in an integrated auxiliary file according to the procedures described above. Compounds can be searched either by using the compound term number when interested in one specific compound, *e.g.*, acetic acid, or by the fragmentation system when interested in classes of compounds, *e.g.*, all acyclic carboxylic acids. In other words, the fragmentation system allows the searcher to retrieve all acyclic carboxylic acids regardless of whether the patent mentioned specifically acetic acid, valeric acid, stearic acid, etc., or carboxylic acids in general. Roles are applied to all compound terms and to certain general terms.

Searches of the data base are processed using the IFI Patent Information Retrieval System. This system is based on a weighted-term search procedure developed by Gulf Research and Development for IFI in the early 1960's[4] and since modified to accommodate the fragment substructure data added in 1972. The data base is searchable from 1950 and includes over 300,000 chemical and chemically related patents.

In order to demonstrate the selectivity of the fragmentation system in a real situation, a series of test questions were recently run and the results compared with equivalent searches set up using the U.S. Patent Office Classification. The searches covered the period January 1965 to December 1971.

Seven questions were compared. The questions covered generic compounds or substructures, *e.g.*, carbodiimides, as well as specific compounds, *e.g.*, methyl γ-nitrosoperfluorobutyrate, and utilized the link-role capability. Each question was set up at three different levels of specifity. One series was conducted without U.S. Patent Office class code limitations. In other words, the entire collection of approximately 141,000 chemical patents for the period 1965–1971 was searched. Retrieval for the seven searches averaged 125 patents or a drop of 0.089% of the file.

The second series of searches was identical in setup with regard to fragmentation, but retrieval was restricted to certain USPO classes. In this case, the average retrieval was 22 patents for the computer searches against an average of 1184 patents for the comparable manual class code searches. This represents a retrieval of 0.016% of the total file. A searcher using the IFI system would screen less than 2% of the patents that his USPO counterpart would screen. Although the breadth of any search will depend on many factors—economics, timing, company policy, and the individual searcher—this particular series was intended to exemplify average searches with respect to completeness.

Class code coverage was further restricted in the third series. This series was intended to represent high spot or specific searches. The IFI retrieval averaged 11 patents *vs.* 293 patents for the class code searches. This shows a 96% reduction in screening effort required for the IFI searches.

In this set of search comparisons, no attempt was made to determine or compare recall. In order to avoid a comparison of the accuracy or validity of the IFI and USPO files, it was assumed that IFI indexing and computer records were 100% accurate, and that the USPO examiners were perfect and no patents were missing from the USPO search shoes. The object of the test was primarily to determine the selectivity of the IFI fragmentation system in handling the indexing and retrieval of the Markush, indefinite, and generic structures so prevalent in chemical patent literature.

Selectivity results for the seven searches are shown in Figure 4. Search questions and strategies are summarized in Appendix A.

CONCLUSION

In summary, the IFI fragmentation system for organic chemicals provides a unique procedure for indexing and searching the specific chemicals, classes of compounds, and Markush structures found in patent literature. The use of an open-ended vocabulary and systematic rules assures the searcher that the fragmentation system will keep pace with constantly changing chemical technology. During 1973, IFI generated approximately 370 new functional groups and 440 new ring structure terms to cover novel compounds found in the patent literature for that year. The MUST and POSSIBLE approach minimizes false retrieval at search and allows the information scientist to zero in on precise chemical structural information. The system is effective for all types of organic chemicals (pharmaceuticals, petroleum chemicals, dyes, etc.), and integration of the fragmentation system with the general terms, compound terms, assignees, and USPO class codes included in the IFI Data Base allow the searcher to design questions at different levels of specificity.

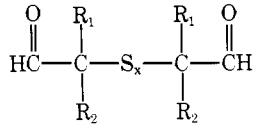
ACKNOWLEDGMENT

The authors are indebted to John W. Lotz, General

Manager, IFI/Plenum Data Company for his assistance and advice during the preparation of this paper.

APPENDIX A

TEST SEARCH 1



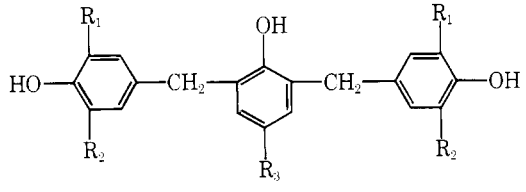
where R_1 is unsubstituted alkyl, R_2 is H or R_1, and x is 1 or 2. The compounds are useful as oil additives.

The computer search strategy included an intersection of the fragment terms F CHO ALDEHYDE FG (P − 2+) and F S SULFIDE, THIOETHER FG (P − 1) or F S2 DISULFIDE FG (P − 1). Negative weight logic (*i.e.*, must not be present) was utilized to eliminate unwanted fragments which had been indexed on a "must" basis. No roles were searched since any mention of the compounds was desired. Oil additive terms were used as "flagging" terms and were not required for retrieval.

Class codes searched were:

Class 44	FUEL AND IGNITING DEVICES
sub 76	Artificial fuel—sulfur, selenium, tellurium, silicon, phosphorus or boron containing
sub 77	Artificial fuel—oxo, or oxy compound containing
Class 252	COMPOSITIONS
sub 45	Lubricants—sulfur, selenium, or tellurium organic compounds containing
sub 48.2*	Lubricants—sulfur, selenium, or tellurium and oxygen organic compounds containing
sub 52	Lubricants—oxygen compound containing
Class 260	CHEMISTRY, CARBON COMPOUNDS
sub 601*	Acyclic aldehydes
sub 608	Carbocyclic or acyclic persulfides
sub 609	Carbocyclic or acyclic mercaptans, mercaptides or thioethers

TEST SEARCH 2



where R_1 and R_2 are C_{3-12} α-branched unsubstituted alkyl and R_3 is C_{1-8} alkyl. The compounds are antioxidants, especially for gasoline.

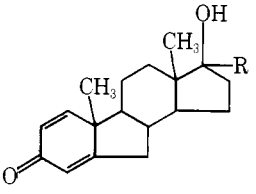
The computer search strategy included an intersection of antioxidant terms with F HO HYDROXY FG (P − 3) and R I C6 BENZENE RING (P). Negative weight logic (*i.e.*, must not be present) was utilized to eliminate unwanted fragments which had been indexed on a "must" basis. The fragment link set was searched in roles 10 (present) and 30 (product). Gasoline terms were used as "flagging" terms and were not required for retrieval.

Class codes searched were:

Class 44	FUELS AND IGNITING DEVICES
sub 77	Artificial fuel—oxo, or oxy compound containing
sub 78	Artificial fuel—carbocyclic oxo, or oxy compound containing
Class 252	COMPOSITIONS
sub 52	Lubricants—oxygen compound containing
sub 396	Anticorrosion agents—oxygen organic compound containing
sub 404*	Antioxidants—phenol or quinone radical containing
Class 260	CHEMISTRY, CARBON COMPOUNDS
sub 619	Phenols
sub 621	Phenols having less than 12 nuclear carbon atoms
sub 624	Phenols having a polycarbon alkyl or alkylene group
sub 625	Polyhydric phenols
sub 626	Phenols having an isopropyl or isopropylene group

TEST SEARCH 3

Methods of preparing



where R is H, CH_3, C_2H_5, or C≡CH.

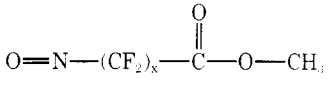
The computer search strategy required F CO KETONE FG (P − 1), F C2 DOUBLE BOND FG (P − 2+), F HO HYDROXY FG (P − 1), and R IV C5C5C6C6 CYCLOPENTA(A)FLUORENE RING. Negative weight logic (*i.e.*, must not be present) was utilized to eliminate unwanted fragments which had been indexed on a "must" basis. Role 30 (product) was searched. The term STEROIDS was used as a "flagging" term and was not required for retrieval.

Class codes searched were:

Class 260	CHEMISTRY, CARBON COMPOUNDS
sub 586*	Carbocyclic or acyclic ketones
sub 617	Carbocyclic or acyclic hydroxy compounds
Class 424	DRUG, BIO-AFFECTING AND BODY TREATING COMPOSITIONS
sub 331	Drug compositions containing ketones as the organic active ingredient

TEST SEARCH 4

Methods of preparing



where x is 2 or 3.

The computer search strategy required F CO2 CARBOXYLIC ESTER FG (P − 1), F NO NITROSO FG (P − 1), F X HALOGEN FG (P − 4+), and FLUORINE, ORGANIC (P). Negative weight logic (*i.e.*, must not be present) was utilized to eliminate unwanted fragments which had been indexed on a "must" basis. Role 30 (product) was searched.

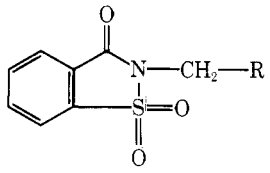
Class codes searched were:

Class 260	CHEMISTRY, CARBON COMPOUNDS
sub 468	Carbocyclic or acyclic carboxylic acid esters
sub 478	Acyclic carboxylic acid esters
sub 487*	Acyclic halogenated carboxylic acid esters
sub 647	Carbocyclic or acyclic nitroso compounds

* Indicates classes searched in the third series of searches.

TEST SEARCH 5

Any reference to compounds having the structure:



where R is any substitutent.

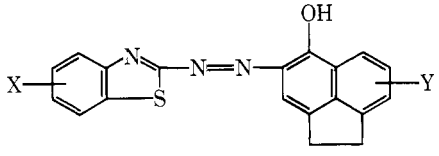
The computer search strategy required F CNO3S O=C-N-S(=O)2 (P − 1) or F CNO3S O=C-N-S(=O)2 (P − 2+) and R II C3NSC6 1,2-BENZISOTHIAZOLE RING. Neither roles nor negative weight logic was utilized.

Class codes searched were:

Class 99	FOODS AND BEVERAGES
sub 141	Saccharous
Class 260	CHEMISTRY, CARBON COMPOUNDS
sub 301*	Sulfoneazoles
Class 424	DRUG, BIO-AFFECTING AND BODY TREATING COMPOSITIONS
sub 269	Drug compositions containing a five-membered ring of at least two heteroatoms, one of which is nitrogen as the active organic ingredient
sub 270	Drug compositions containing a thiazole compound as the active organic ingredient
sub 321	Drug compositions containing a sulfonamide compound as the active organic ingredient

TEST SEARCH 6

Monoazo dyes of the formula:



where X and Y are any substituent.

The computer search strategy required F CN3S S-C(=N)-N=N (P − 1), R II C3NSC6 BENZOTHIAZOLE RING (P), R III C5C6C6 ACENAPHTHYLENE, ACENAPHTHENE RING, and F HO HYDROXY FG (P − 1) or F HO HYDROXY FG (P − 2) or F HO HYDROXY FG (P − 3) or F HO HYDROXY FG (P − 4+). Neither roles nor negative weight logic was utilized. Appropriate dye terms (*e.g.*, MONOAZO DYES) were used as "flagging" terms and were not required for retrieval.

Class codes searched were

Class 8	BLEACHING AND DYEING; FLUID TREATMENT AND CHEMICAL MODIFICATION OF TEXTILES AND FIBERS
sub 26	Dyeing processes or compositions containing several dyes one of which is an azo dye
sub 27	Dyeing processes or compositions containing an azo dye and vat or sulfur dyes
sub 41	Dyeing processes or compositions containing azo dyes
Class 96	PHOTOGRAPHIC CHEMISTRY, PROCESSES AND MATERIALS
sub 56.3	Developing composition containing an azo compound with a heterocyclic sulfur atom as a color developer
sub 75	Light sensitive element containing a diazo chromium compound or iron compound sensitizer
sub 91	Light sensitive composition containing a diazo compound
Class 260	CHEMISTRY, CARBON COMPOUNDS
sub 158*	Thiazole-azo compounds
sub 192	Monoazo compounds

TEST SEARCH 7

A composition containing compounds of the formula:

where R is unsubstituted or substituted alkyl, cycloalkyl, or carbaryl. The composition may be useful as a fungicide.

The computer search strategy required F CN2 N=C=N (P − 1) or F CN2 N=C=N (P − 2+). The fragment term HETEROCYCLIC RING (M) was negated and no roles were applied. Fungicide terms were "flagged" and were not required for retrieval.

Class codes searched were:

Class 71	CHEMISTRY, FERTILIZERS
sub 121*	Plant-growth regulating composition containing an amine as the active organic ingredient
Class 260	CHEMISTRY, CARBON COMPOUNDS
sub 551*	Carbocyclic or acyclic amides
Class 424	DRUG, BIO-AFFECTING AND BODY TREATING COMPOSITIONS
sub 325	Drug compositions containing an amine as the active organic ingredient

LITERATURE CITED

(1) Edge, E. B., Fisher, H. G., and Bannister, L. A., "System for Indexing Research Reports Using a Punch Card Machine," *Amer. Doc.*, **8** (4), 275 (1957).

(2) Rasmussen, L. E., and Van Oot, J. G., "Operation of Du Pont's Central Patent Index," *J. Chem. Doc.*, **9**, 201 (1969).

(3) Lotz, J. W., "Fragmentation: A Practical Solution to the Retrieval of Generic Structures Found in the Patent Literature," presented before the Chemistry Division Meeting at the Special Libraries Association 1973 Annual Conference.

(4) Cattley, J. M., *et al.*, "Weighted-Term Searching in Patent Information Retrieval," presented at the Symposium on Information Retrieval-Searching Techniques, American Institute of Chemical Engineers, 58th National Meeting, Feb 6-9, 1966.