

The Cyanamid Organic Structure Code and Search System*

By LEE N. STARKER and JOSE A. CORDERO

Technical Information Services, Lederle Laboratories Division,
American Cyanamid Company, Pearl River, New York

Received August 24, 1961

About ten years ago, a new method for filing organic compound data by molecular formula was developed by J.H. Fletcher¹ of the Stamford Laboratories of the American Cyanamid Company. At that time, the Cyanamid Molecular Formula File contained but 20,000 compounds, and the generic search feature of the Fletcher System permitted searches to be carried out conveniently and in a reasonable length of time.

Since then, the number of compounds contained in the file has grown to 60,000 and in five years should approach the 100,000 mark. The rapid growth of the file and an increased demand for generic searches has severely limited our capacity to produce answers without undue delay.

In 1959 the decision was made to adopt or develop a more efficient search system. After a survey of several of the codes currently in use for mechanized search systems, it was decided to use the Merrell Code² as a first approximation. An analysis of the results of coding and searching 1000 compounds with this code showed that, while it was well suited to a small compound collection, it was not sufficiently precise or specific for our needs. Its major fault as applied to our own compound collection lay in the size of the false drop that resulted. Several modifications were made and, after another false start, the code now in use was developed. This was designed, with one exception, as a direct-punch code and uses 68 columns of the IBM card. Figure 1 indicates how the code is broken down and the number of groups, etc. that may be punched.

The CL number is our identification number and is assigned serially to all compounds synthesized in the laboratory, or purchased for research use. The Molecular Formula field permits the coding and counting of nitrogen, sulfur, oxygen, carbon, the halogens, phosphorus and boron. The presence of other elements is recorded but not counted.

The presence of 131 functional group codes has proved sufficient up to this time. Sufficient space exists on the card to add other groups, as required. Monocyclic rings (Fig. 2) are coded for size and (for five and six membered rings) degree of saturation; for the number and types of hetero atoms present, and for the relationship of hetero atoms to each other. Thus pyridine would be coded at 25-6, 26-1, and 28-1 while piperazine would be coded at 25-3, 26-2 and 28-4.

Substituents are counted and recorded for monocyclic rings, but not for fused ring systems. By indicating the total number of substituents on a ring, certain search problems can be simplified. Thus in a narrowly defined request—compounds in which a nitro group and an amino group are substituted onto a benzene ring—only those compounds which are disubstituted need be searched.

(1) J.H. Fletcher and D.S. Dubbs, *Chem. Eng. News*, 5888 (1956).

(2) K.W. Wheeler, E.R. Andrews, F. Fallon, G.R. Krueger, F.P. Palopoli and E.L. Schumann, *Am. Doc.*, 9, 198 (1958).

* Presented before the Division of Chemical Literature, American Chemical Society, St. Louis, Mo., March, 1961.

ITEM	COLUMNS
SERIAL NUMBER	6
MOL. FORMULA, MISC. ELEMENTS	11
FUNCTIONAL GROUPS	33 (131 CODES)
MONOCYCLIC RINGS	5
RING SUBSTITUTION	2
ALKYLENE and ALKYL GROUPS	2
FUSED RINGS	6
MISCELLANEOUS PROPERTIES	3
UNUSED COLUMNS	12

FIG. 1 CYANAMID GENERIC CODE










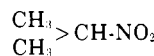
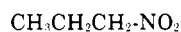
	24 CARBOCYCLIC RINGS	25 HETERO RINGS	26 HETERO ATOMS	28 HETERO ORIENT'N	
1			N	MONO	
2			2N	1, 2	
3			3N	1, 3	25-6, 26-1, 28-1
4			4N	1, 4	
5	 or 		S	1, 2, 3	
6			2S	1, 2, 4	25-3, 26-2, 28-4

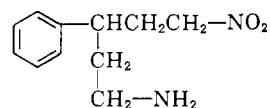
FIG. 2 MONOCYCLIC RINGS

Where a ring only bears two substituents the orientation of these groups also is recorded, whereas for more highly substituted rings only the orientation of identical groups is recorded.

Alkyl and alkylene groups are recorded without regard to branching. An alkyl group is defined as any chain of one or more carbons which is attached to no more than one functional group, ring, or alkylene chain. An alkylene chain is defined as any chain of one or more carbons which is attached to two functional groups, rings, or other alkylene chains.

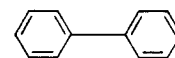
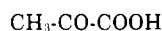


Alkyl of 3



Alkylens of 3 and 5

The alkylene of 0 is used to describe the fact that two functional groups or two rings are adjacent to each other.



Alkylens of Zero

Thus, by making appropriate use of the alkylene codes, differentiation can be made between such related compounds as α -keto esters (alkylene of 0), β -keto esters (alkylene of 1) and γ -keto esters (alkylene of 2).

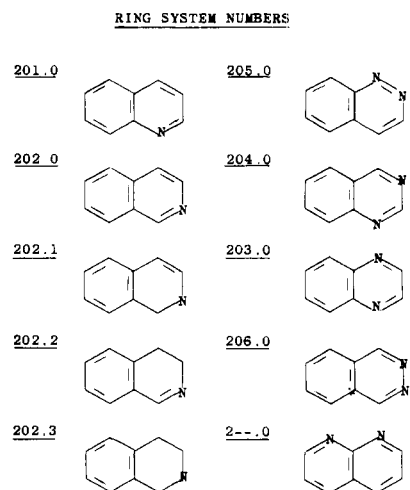


FIG. 3

The Ring System Number (RSN) is used to differentiate between various fused ring systems. This approach (Fig. 3) assigns to each fused ring system a four-digit number. The first three digits are characteristic of the system, while the fourth represents the degree or type of unsaturation present in the system.

Thus 201.0 represents the quinoline ring, while 202.0 represents the isoquinoline system. In general, a fourth digit of 0 represents the most highly unsaturated form, while a fourth digit of 9 represents the most highly saturated form. As of January 1, 1961, our system contained over 500 RSN's. In searching for various systems, only the first three digits of a number need be sorted if all members of the system, regardless of saturation are desired. In those cases where a specific type of unsaturation is wanted, searching for the fourth digit will eliminate all unwanted forms.

The number of rings in the system and the sizes of the rings also are recorded, as is the presence of certain specific heterocyclic atoms. Thus, while searching on the Ring System Number permits the retrieval of any specified fused ring type, the use of the more generic data allows the retrieval of broader classes of compounds. If we were then interested in all fused ring systems which contain two rings, one of which is six-membered and one of which is five membered, and which contained one nitrogen in either ring, we would answer this question by searching Column 22-12, 5, 6 and Column 23-1.

The functional group field occupies Columns 33-65. Specific codes have been included to cover the most common organic groupings. The presence of any of these groups in a molecule requires that the corresponding hole in the IBM card be punched—without immediate regard to the use of any of the other modifying codes that may be used. Among these modifying codes are the Ring and Poly Codes in Rows 12 and 11 and the linkage codes in Rows 6-9.

The Ring Code in Row 12 (Fig. 4) is common to all functional group columns and is used to denote the presence of a group within a ring. Thus a lactone is coded as a "Ring ester," 34-12,0. The Poly Code in Row 11 is also common to all functional group columns, and repre-

LINKAGE CODE



	33	34	47	48
12	RING	RING	RING	R ALK
11	POLY	POLY	POLY	ALICYC
0	-COOH	R-COO-R ₁	R-CON ^{R₁} _{R₂}	ARYL
1		R-O-R ₁		HET
2		R ALK		R ₁ ALK
3		ALICYC		ALICYC
4		ARYL		ARYL
5		HET		HET
6	ALK	R ₁ ALK		R ₂ ALK
7	ALICYC	ALICYC		ALICYC
8	ARYL	ARYL		ARYL
9	HET	HET		HET

FIG. 4

LINKAGE CODING

	CODE
R-COO-R ₁ CH ₃ -COO-C ₂ H ₅	34-0, 2, 6
C ₆ H ₅ -COO-C ₂ H ₅	34-0, 4, 6
2-Py*-COO-C ₆ H ₅	34-0, 5, 8
R-CON(R ₁)(R ₂) CH ₃ -CON(CH ₃)(C ₆ H ₅)	47-0, 48-12, 2, 8
CH ₃ -CON(C ₆ H ₅)(2-Py*)	47-0, 48-12, 4, 9
2-Py*-CON(CH ₃)(C ₆ H ₅)	47-0, 48-1, 2, 8

*2-Py= 2-Pyridyl

FIG. 5

sents the presence of a given group more than once. A phthalic acid would be coded for two carboxy groups at 33-11,0.

The "linkage codes" (Fig. 4) were added to permit increased specificity when searching for functional groups. With this code, each functional group receives a punch for the group itself and another to show how it is linked to the remainder of the molecule. Polyfunctional groups such as the ester or amide groups receive two or three linkage punches. Thus, we can search generically for all carboxylic acids (33-0), or we can be more specific and separate alkyl carboxylic acids (33-0,6) from aryl carboxylic acids (33-0,8) from heterocyclic carboxylic acids (33-0,9).

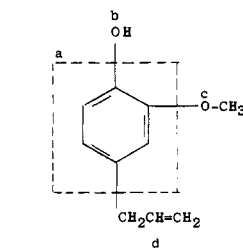
For esters, the group on carbon is always designated as R₁ while the group on oxygen is R₂. For amides, the group on carbon is again R₁ while the groups on nitrogen are R₂ and R₃. Since the two N-substituents are equivalent, a precedence rule—alkyl before alicyclic before aryl before hetero—defines which substituent is coded as R₂ and which one receives the R₃ linkage. Examples are shown in Fig. 5.

Another point of interest is our use of "buildup groups" (Fig. 6). Thus the code carries a "keto" carbonyl group and a "non-keto" carbonyl group. The keto carbonyl is the normal "oxo" carbonyl, while the non-keto carbonyl (Col. 35, Row 2) is used to build up complex groups which do not have their own code. For example, the urea group normally is coded at 47-2. An acyl urea, which contains a "non-keto" carbonyl is coded at 35-2 and 47-2. This permits the retrieval of all ureas, whether acylated or not by sorting on 47-2, or permits the retrieval of acylated ureas only by making another sort on 35-2. To further specify the answer, the alkylene of zero code, 31-0, is used to indicate the two adjacent functional groups. This combination of codes then makes up what we refer to as a "complex group code." The use of the "non-keto" carbonyl code also prevents compounds of the acylurea type from being retrieved when ketonic carbonyls are being searched.

Examples of completely coded compounds are shown in Figs. 7 and 8. Coding of compounds was begun in January of 1960 and, with the aid of four college students during

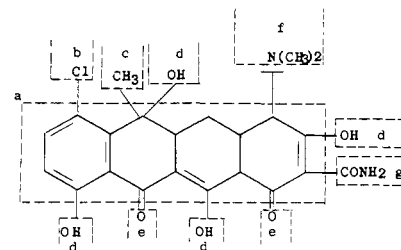
BUILDUP GROUPS			
	COL. 35	49	50
0	-CHO	-SH	-CHS
1	C=O (Both bonds to carbon)	-S- (Both bonds to carbon)	C=S (Both bonds to carbon)
2	C=O (No more than one bond to carbon)	-S- (No more than one bond to carbon)	C=S (No more than one bond to carbon)
3			

FIG. 6



- a - PHENYL RING 24-6
 TRISUBST RING 29-3
 b - ARYL OH 38-1
 c - ETHER, ALKYL and ARYL LINKS 34-1,2,8
 d - 3-CARBON ALKYL, DOUBLE-BONDED, LINKED TO ARYL RING 32-12,3,8

FIG. 7



- a - TETRACYCLINE RING SYSTEM 708.1
 b - HALOGEN ON ARYL 59-1
 c - 1-CARBON ALKYL ON ALICYCLIC 32-1,7
 d - POLY OH ON ARYL AND ALICYCLIC 38-11,0,1
 e - POLY RING C=O 35-12,11,1
 f - TERT. AMINE, LINKED TO ALKYL, ALKYL, ALICYCLIC 41-0,42-12,2,7
 g - AMIDE, C-ALICYCLIC LINK 47-0,48-11

FIG. 8

the summer, was completed in September of the same year. The verification of the coding was done by our regular staff and was completed in February of this year. We estimate that the average time required to code and check a given compound is about 2-3 minutes. Further checks on the accuracy of the coding and the punching were made by comparing certain selected searches run by machine against the same search run manually by visual inspection of the Molecular Formula Index. While these comparative searches generally turn up a few compounds (usually less than 1%) which have been miscoded and missed, we also found that a larger number of appropriate compounds were missed on the corresponding hand search.

The steps through which a search request is processed are shown in Fig. 9.

The search request is received on a standard form which indicates the generic structure, with any limitations and alternates which may be desired. The request is checked over by an information chemist who then selects the appropriate subdeck and programs the search.

We use subdecks to eliminate end-to-end searching of the complete deck of 60,000 cards. These subdecks are based on either functional groups or card columns, the only criterion being that the subdeck be within specific size limits. Our present plans call for the use of subdecks containing 10-15,000 cards, thus limiting machine time on the 083 sorter to a maximum of about 20 minutes per search.

Despite the fact that the average number of sorts per search is 4, we decided that the 083 Sorter would be a more efficient machine for our purposes than would the IBM 101 Electronic Statistical Machine. To search a deck of 15,000 cards with the IBM 101 at 450 cards/minute requires a minimum of 33 minutes (no allowances being made for board wiring or card-handling time). The same search made on the IBM 083 operating at 1000 cards/minute requires 15 minutes for the first sort. If we assume that the second pass is made on a drop of 3000 cards (20%) and the second and third drops are also 20% of the previous drop then we arrive at a total time of 18.7 minutes which is substantially less than the 33 minutes required for the IBM 101.

Additional time occasionally can be saved since a sort involving a zone punch and a numerical punch in the same column (cyclic anhydrides, 33-12,1) can be sorted in one pass, by making use of the alphabetic sort device on the 083 sorter.

Once the deck has been sorted, the cards which comprise the answer are put into numerical order. The CL numbers are printed out by either an IBM 407 or the IBM Document Writer. The list of numbers goes to a clerk who reproduces the structures of the indicated compounds on a 3M Printer-Reader (Filmac-100) from a microfilm copy of the CL Compound Card File. These structures are then passed on to a chemist, who eliminates the false drop, and then transmits the remaining structures to the requestor as the answer to his search request.

An alternative reporting procedure permits a clerk to pull the 3 x 5 structure cards from the serial file and reproduce them on the Xerox 914 duplicator. On completion, the originals are refiled and the copies are analyzed as above. In this case, the valid Xerox copies become the search report. Either procedure permits a capability of at least three complete searches a day. Figure 10 gives some data on the searches that were run by machine during 1960. A total of 41 such searches was made. The deck size ranged from 5000-32,000 cards with an average of 15,000. Since the file had not been broken down into subdecks at this time, all searches were end to end and for this reason the time required is given in minutes per 1000 cards of the deck. This figure is 1.38 min./1000 cards which means that a 10,000 card deck required an average of 13.8 minutes to search. The figure also includes the time required to place the final answer in numerical order (six sorts).

The average number of sorts required was four per search and varied from one to nine sorts. The average false drop was 13% but half of these were 3% or less. Several runs were made where false drops as high as 100% were obtained. These resulted ordinarily from requests for rigidly oriented rings. Thus a request for *o*-chlorobenzoic acid derivatives might yield 10 compounds, all of which were chlorobenzoic acids, but only one of which had the chlorine and the carboxy group ortho to each other. This would result in a false drop of 90%. False drops can also occur due to ambiguous coding in certain areas, but as long as these areas are not those most commonly searched, the situation is tolerable. Should the character of our compound file change so that there is greater emphasis on these ambiguously coded areas, we feel that it will be a relatively easy matter to recode those particular compounds.

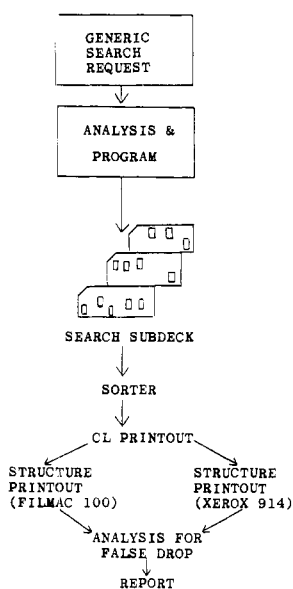


FIG. 9

REASONS FOR SEARCH REQUESTS

COMPOUNDS RELATED TO AN ACTIVE COMPOUND (COMMERCIAL).
 MODEL COMPOUNDS FOR IR AND UV COMPARISONS.
 MODEL COMPOUNDS FOR NMR COMPARISONS.
 CHECK NEW ANALYTIC PROCEDURE.
 FOLLOW-UP TESTING LEAD.
 LOCATE SYNTHETIC INTERMEDIATES.
 PRE-PROJECT SURVEY.

FIG. 11

1960 GENERIC MACHINE SEARCHES

	41 SEARCHES		
	RANGE	AVERAGE	MEDIAN
DECK SIZE	5,000-32,000	15,000	14,000
NO. OF SORTS	1-9	4	4
SEARCH TIME	0.2 min/1000- 2.8 min/1000	1.38 min/1000	1.60
DROP	0 - 484	91	32
FALSE DROP	0 - 100%	13%	3%

FIG. 10

The need for an improved and more rapid search has been graphically demonstrated by the increased number of search requests which have been received. These requests are now being made at the rate of about 200 per year. Prior to the installation of the system, an average of 15-20 searches per year was carried out at Pearl River.

Another area in which we hope to make extended use of the system is in the study of structure-activity correlations. This activity has not yet begun, but it is hoped that preliminary studies can begin in the near future.

Generic Mechanized Search System*

By JULIUS FROME

Office of Research and Development, Patent Office, U. S. Department of Commerce, Washington, D. C.

Received August 24, 1961

During the last three and one half years, in answering actual questions requested, the U. S. Patent Office has made well over ten thousand mechanized searches. This includes searches on the steroids (2, 3), resins (4, 9, 10), and phosphorus compounds (9). Mechanized searches were made on various machines such as computers, *i.e.*, SEAC, Bendix G-15, RAMAC 305, and punched card machines, *i.e.*, single column sorters, ILAS, and IBM-101. Although most of these mechanized searches were for patent examiners, a great many were conducted on an experimental basis for research workers in industry.

From an analysis of the questions submitted, it was quite apparent that a majority of the searches desired were generic rather than specific. In further studying mechanization from the viewpoint of the questioner, the following facts became apparent: (1) generic as well as specific searches are a necessity; (2) system should be compatible

to various machines; (3) system should be segmentable in accordance with need. The user must be able to make generic as well as specific searches. The system must be compatible with several machines since some users have access to punched card machines and some to computers. Finally, for some users a great depth of information and detail is necessary, whereas others are satisfied with less. Therefore, from the questioner's viewpoint, the system should be able to be segmented so as to give him as much or as little information as he desires.

From the viewpoint of an information scientist, a system also should have these capabilities: (4) should not have limited subject application; (5) should be capable of storing and retrieving compounds, processes, compositions, biological data; (6) should have an open-end dictionary; (7) Information extracted should be useful in systems organized for: (a) random access searching, (b) serial searching. A system has been developed which accomplishes many of the above objectives within practical limits.

*Presented before Division of Chemical Literature, American Chemical Society, St. Louis, Missouri, March, 1961.