The Parke-Davis Code for Chemical Structures*

By HARRIET A. GEER, ALEXANDER M. MOORE, C. CECILY HOWARD, and CAROLYN E. EADY
Parke, Davis & Company, Ann Arbor, Michigan
Received September 22, 1961

When a central chemical file was established at Parke-Davis in 1946, an adaptation of the Wiselogle Classification System [1] to edge-notched cards was set up for coding chemical structures. This proved to be a workable system until our file grew to over 20,000 coded compounds. The time required to sort this number of cards made it necessary for us to convert to mechanical devices. Although it was possible to convert directly from the edge-notched cards to IBM without recoding, maximum use of more versatile equipment would not have been realized. We decided to revise our old code or develop a new one.

To assist us in developing a code, three chemists** from our research staff were appointed to act in an advisory capacity. We found their presentation of the research man's viewpoint extremely valuable. We were also fortunate in being able to call on Dr. Karl Heumann of Chemical Abstracts Service for suggestions and advice.

Although it would have been feasible to develop the structure code before selecting the equipment, we preferred to design the code for the equipment. In making our choice, we were influenced by these considerations: (1) we wanted the searching equipment easily accessible to our office; (2) we wanted the answers to a search delivered as structures without recourse to another file.

When economic considerations were taken into account, we decided to use a simple IBM sorter with the structure reproduced on the card. If the simple sorter proved inadequate, we could progress to the IBM-101 without repunching.

We were aware that a specific notation punched on the card could serve in place of a printed structure. However, direct examination of the cards would be meaningful only to those familiar with the notation. For this reason, we decided that a notation would not be a satisfactory substitute for a structure. Since the structures of our compounds are on ozalid masters, ozalid coated IBM cards, obtained from the Ozalid Company, were selected. After initial difficulties, this type of card proved to be a satisfactory solution. However, the Ozalid Company plans to discontinue this line and we are presently considering other ways of reproducing the structure on the card.

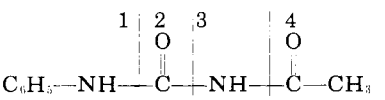
For the most efficient use of accounting-type equipment, we selected direct punching rather than field coding. Both Dow [2] and Merrell [3] had found a direct punch code satisfactory for structure searching of relatively large numbers of compounds. More recently, Lederle [4] has reported on their use of this type of structure code for some 60,000 compounds. A member of the IBM research staff assured us that a direct punch code could be transferred to magnetic tape should we eventually outgrow accounting-type equipment.

*Presented before the Division of Chemical Literature, ACS National Meeting, Chicago, Illinois, September 7, 1961.
**John R. Dice, Franklin W. Short, and Leslie M. Werbel.

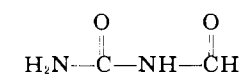
After choosing our equipment, we established these criteria to guide us in the development of the code: (1) it should be possible to code any compound; (2) rules for coding should not be overly complex; (3) reproducibility of coding is essential; (4) when a search is made, it should produce all of the compounds of the specified type in the files; (5) the number of unwanted compounds should be at a minimum in a search; (6) the possibility of locating analogs from more than one approach should be inherent.

The primary use of our edge-notched card file had been to search for analogs of a given compound. A code which describes structural fragments but not the exact structure was still adequate for our use and is easily manipulated on relatively simple machines.

In the past, people have used two general approaches to describe the functional groups present. In one approach, simple fragments are used from which the more complex groups are built. In the other, known functional groups are listed with new ones added as it becomes necessary. The Wiselogle code uses the first approach and the CBCC (Chemical Biological Coordination Center) Code [5] the second. The Wiselogle system hydrolyzes complex functional groups into their component parts and lists end products of hydrolysis. An example of the two treatments is shown in Fig. 1.



Wiselogle Code Designations



CBCC Code Designation

Fig. 1.

However, when groups of varying complexity are all listed as end products of hydrolysis, the false drops become increasingly numerous as the size of the file increases. At Warner-Lambert, [6] this situation has been alleviated by assigning additional code designations to separate unhydrolyzed groups from those obtained by hydrolysis.

The advantage of a single code designation for a complex group is obvious when that unit is desired in a search. However, it is impractical to list all possibilities. The CBCC Code permits the coding of newly encountered groups by using a combination of those already listed. If this is not possible, a new group is added to the code.

In the Parke-Davis code, a middle course is followed in coding functional groups. The size of the coding unit is limited to a central atom plus its attached atoms. If a functional group is more complex than a single coding unit, it is described by more than one code designation.

Double coding is carried out only when all atom to atom attachments to elements other than to carbon or hydrogen have not been indicated. The following code designations for functional groups are listed: (1) C with two or more attached atoms other than C or H; (2) N, S, O and halogen with one or more attached atoms other than C or H; (3) N, S, O and halogen with no attached atoms other than C or H.

To avoid listing all possibilities, general designations are introduced for describing combinations of atoms rarely encountered. Commonly occurring groups, however, are specifically designated.

In deciding which groups occur frequently enough to be assigned a specific designation, we relied to a large extent upon a count made by Heumann and Dale.[7] Statistics on the number of compounds containing each group listed in the CBCC Code were compiled from some 50,000 compounds in the CBCC file. Since these compounds were a cross-section of those which had been tested for biological action, we believed the distribution of types should be similar to those in our file.

Certain aspects of the Wiselogle system were used without change. We retained the rules for aromatizing ring structures as well as those for tautomerism to allow reproducible coding of compounds which may exist in tautomeric form.

We use 60 columns of the IBM card including 6 for an identifying compound number. In Fig. 2 we have shown the general layout of the card.

COLUMN	
1-6	Compound accession number
7-9	Foreign elements (elements other than C, H, N, S, O, and X)
10-17	Carbon as central atom
18-32	N, S, O, X groups
33-44	Single heterocyclic rings
45-53	Total ring skeleton
54-55	Special designations
56-58	Spatial relationships
59	File order
60	Partial molecular formula

Fig. 2.—IBM card

Foreign Elements (Columns 7-9).—We have not coded foreign elements in detail. Indication of the element plus limited information on oxidation state and attached groups is adequate. Since the occurrence in a single compound of two different foreign elements is rare, we overpunch their codes. Figure 3 shows the method of coding.

COLUMN	
7 and 8	2-Digit number assigned to the element
9	Attaching groups or elements

Fig. 3.—Coding of foreign elements (elements other than C, H, X, O, S, N).

Functional Groups (Columns 10-32).—Code designations for the functional groups are shown on the bottom half of the IBM card and additional descriptive material is recorded at the top of the column. An example of the information indicated for functional groups is shown in Fig. 4.

IBM PUNCH	
12	>1 Group
11	Cyclic
0	Aliphatic or alicyclic group attached
1	Aromatic group attached
2	Substituted carbon attached
3	Element other than C or H attached

Fig. 4.—Descriptive material referring to functional groups

When carbon is a central atom with two or more other atoms attached, seven general designations cover the possible functional groups as shown in Fig. 5.

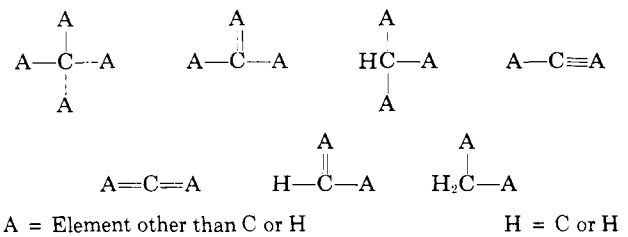


Fig. 5.—Carbon as central element

Commonly occurring groups are shown by specific designations, and less commonly occurring ones by general designations. When one of the general groups is used, the element represented by A is specified in another part of the code.

For the general structure HC(=A)—A, the specific groups in our code are given in Fig. 6.

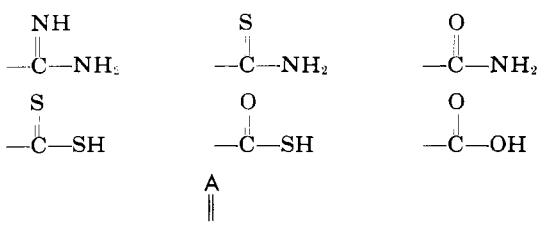


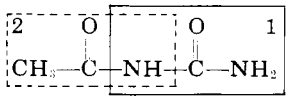
Fig. 6.— —C(=A)— groups listed in code.

Since the amido and carboxyl groups are so frequently present, each of these is assigned more than one code designation as shown in Fig. 7. Two amido groups are listed, namely, the amido group present in simple amides and a second one which is part of a more complex group. The latter unit is used in coding hydrazides, imides, *etc.* Three different types of carboxyl groups are listed, namely, the carboxyl group in acids, esters, and other structures such as anhydrides.

—C(=O)—NH2	—C(=O)—OH
1. H = H or R	1. H = H
2. H = A or substituted C	2. H = R
R is aryl, alkyl, *etc.*	3. H = A or substituted C
A is element other than C or H	

Fig. 7.—Designations for amido and carboxyl groups

An example of the coding of a complex group is shown in Fig. 8.



1. Ureido group attached to a substituted carbon
2. Amido group attached to a substituted carbon

Fig. 8.—Coding of a complex functional group

Acetylurea contains two central carbon atoms which are coded by the designations shown above. In addition, the code indicates that the carbon of this amido group is attached to an aliphatic or alicyclic group.

Following carbon as central atom, the various nitrogen, sulfur, oxygen and halogen groups are listed. For groups not designated specifically, general designations are used showing the number of atoms attached other than carbon or hydrogen. Figure 9 shows the treatment of nitrogen attached to two atoms other than carbon or hydrogen.

HO—NO	H_2N—NO
HNO_2 (Nitro)	HN_3
HNO_2 (Other than nitro)	HN=N—NH_2

A = Element other than carbon or hydrogen
H = Carbon or hydrogen

Fig. 9.—2A to N groups listed in code.

Once again, the attachments to these groups, number, *etc.*, are coded in the upper portion of the column. When the general designation 2A to N is used, the A-atom is specified elsewhere.

Ring Structures (Columns 33-53).—Ring structures are represented by coding the single heterocyclic rings as well as the total ring skeleton. Heterocyclic structures which occur frequently are indicated specifically, but less frequently occurring ones are given a general designation. Figure 10 illustrates the coding of 6-membered rings containing two nitrogens.

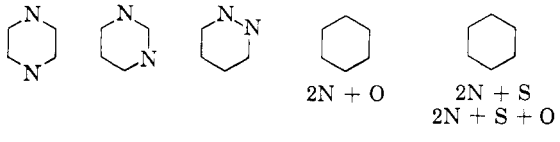


Fig. 10.—6-Membered rings containing 2 nitrogens.

The type of ring is indicated at the bottom of the column, and additional information relating to fusion, aromaticity and number present is noted at the top.

Frequently, rings of a definite configuration such as acridines or benzacridines are desired. For this reason, it should be possible to search for the over-all configuration of ring structures as well as the individual heterocyclic components. Norton and Opler [8] code fused rings as the parent cyclic hydrocarbons with special numbers assigned to the hetero-atoms. We specifically designate only those parent ring structures which occur most frequently and indicate the presence but not the position of the hetero-atom. Degree of saturation is indicated at the top of the column by one of the following designations: mixed aromatic and alicyclic rings, aromatic rings only, or alicyclic rings only.

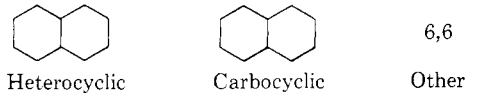


Fig. 11.—Fused rings containing two 6-membered rings

Spatial Relationships (Columns 56-58).—Spatial relationships, especially within relatively short atomic distances, exert important effects upon the properties of compounds. An example of this type of relationship which quickly comes to mind is the location of the amino group in the α-position in naturally occurring amino acids. In order to search for atomic distances between neighboring groups or elements, the number of carbons is recorded between carbon functional groups, nitrogen, sulfur or oxygen, and aromatic ring structures up to a maximum distance of four carbons. The spatial coding noted for carbon functional groups is shown in Fig. 12.

GROUP	NO. OF SEPARATING CARBONS
Aromatic ring	1, 2 and 3
S (or O)	1 and 2
N	1, 2 and 3
C=	0, 1, 2 and 3

C= is carbon attached by two or more bonds to an element or elements other than C or H.

Fig. 12.—Spatial coding of functional carbon(C=).

An application of this relationship is illustrated in a search for α-aminoacids. The nitrogen atom is separated by one carbon from a carbon attached by two or more bonds to another element. Phenylalanines could be further separated from a group of α-aminoacids by searching for an aromatic ring separated from the carbon functional group by two carbons, or an aromatic ring to nitrogen distance of two carbons.

Spatial relationships of substituent groups on aromatic rings have been indicated for benzene only. In a single column, different designations are assigned to singly, doubly, and triply substituted benzene rings with the substituents in the ortho, meta, para, vicinal, unsymmetrical or symmetrical positions, respectively. Benzene rings with four or more substituents are placed in a single category.

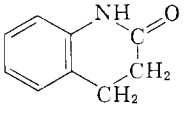
Other structural characteristics included are the presence of conjugated unsaturation, a spiro atom, an atomic bridge, and a quaternary carbon. A single column contains a limited molecular formula for halogen, oxygen, sulfur, and nitrogen. We have allowed for future expansion of the code by leaving spaces for more specific delineation of general groups which are over-loaded.

The value of prefiling of a punched card file has been demonstrated frequently. In our edge-notched card file, the cards were filed according to the main divisions of the Wiselogle Classification System. Frequently, it was possible to eliminate a large portion of the file before starting a search. Because of our satisfaction with this method, we based our filing of IBM cards on this classification system. Minor changes were required as we no longer hydrolyze the compounds when coding, but the over-all arrangement is essentially the same. The main divisions are based upon an assigned seniority of elements in the order of "foreign" elements, nitrogen, sulfur,

oxygen, halogen and carbon, respectively. Within these main divisions, the presence of the elements in aromatic rings and the groups attached to the elements serve to divide the compounds into smaller groups. Our file is divided into twelve sections by the punches in a single column. As the size of the file increases, further subdivision by means of the molecular formula may prove useful.

We spent over a year in the development of the code, which was not a full-time project but was carried out in addition to other duties. When the code was completed, 1,000 compounds were coded for a pilot run. As a result of this trial, certain changes were made in June, 1959. About 30,000 compounds now have been coded without introducing additional changes. Reproducibility of coding by the four chemists in the office is obtained routinely. Coding time, including checking, averages from about 1.5 to 2.5 minutes per compound. Conversion of the edge-notched card file to IBM at a rate of 10,000 compounds per year should be completed within the next two years.

An example of the use of the code is illustrated in Fig. 13 where coded characteristics of hydrocarbostyril are listed. Only five of the twelve divisions in the IBM File need be searched to locate analogs.



1. Mixed alicyclic-aromatic heterocyclic 6,6-ring fused on one face
2. Non-aromatic pyridine fused on B side
3. Aliphatic, cyclic amide
4. Central carbon atom separated by 2 carbons from aromatic ring

Fig. 13.—Search for analogs of hydrocarbostyril

When a search of this type is performed, the chemist may examine the structures on the IBM card directly, or the structures may be reproduced on the Xerox-914 to form a permanent record.

From our experience with the code, we believe that up to 100,000 compounds can be handled on accounting-type equipment without becoming unwieldy. Before developing more detailed methods of coding chemical structures, we plan to study how we can best use the code in conjunction with machine retrieval of biological information.

REFERENCES

(1) E. L. Buhle, E. D. Hartnell, A. M. Moore, L. R. Wiselogle, and F. Y. Wiselogle, *J. Chem. Educ.*, **23**, 375-391 (1946).
(2) H. S. Nutting and S. P. Klesney, "The Selection of Organic Compounds According to Structural Characteristics," Presented at the American Chemical Society Division of Chemical Literature, Pittsburgh, Pennsylvania, Jan. 20, 1958.
(3) K. W. Wheeler, E.R. Andrews, F. Fallon, G. L. Krueger, F. P. Palopoli, and E. L. Schumann, *Am. Document.*, **9**, 198-207 (1958).
(4) L. N. Starker and J. A. Cordero, *J. Chem. Doc.*, **2**, 12 (1962).
(5) "A Method of Coding Chemicals for Correlation and Classification," Chemical-Biological Coordination Center, National Research Council, Washington, D. C., 1950.
(6) F. H. Arendell, *J. Chem. Doc.*, **1**, 47 (1961).
(7) "Advances in Documentation and Library Science, Vol. 1, Progress Report in Chemical Literature Retrieval," edited by G. L. Peaks, A. Kent, J. W. Perry, Interscience Publishers, New York, 1957, pp. 201-214, "Survey of Chemical Structures," E. Dale and K. F. Heumann.
(8) T. R. Norton and A. Opler, "A Manual for Coding Organic Compounds for use with a Mechanized Searching System." Dow Chemical Company, Midland, Michigan (March 15, 1956).