Eleven Years of Structure Retrieval Using the SK&F Fragment Codes*

PAUL N. CRAIG and HELEN M. EBERT
Smith Kline & French Laboratories, Philadelphia, Pa. 19101

Received March 3, 1968

The SK&F version of the CBCC chemical fragment code has been in operation since 1957. In 1964, we added another 107 code terms, taking into consideration the strengths and weaknesses of the SK&F-CBCC code and the codes used by the "Dokumentationsring." The utility and limitations of the revised code are discussed, based on the experience gained from some 1200 structure searches since 1964. Sample searches are described.

Many research laboratories face the common problem of organizing a large number of research chemicals so that they may be assembled quickly into groups containing desirable structural features. The state of the art was studied in considerable depth in two reports issued by the National Academy of Sciences–National Research Council in 1964 and 1965.[1,2] Although these reports dwelled heavily on notation systems, most of the systems employed for structure retrieval use fragment codes, and we are now reporting our experiences in this area.

MODIFICATION OF THE CBCC CHEMICAL CODE

By 1954, more than 3000 research chemicals had accumulated in the files of Smith Kline & French Laboratories. An edge-notched punched card chemical code containing 121 fragment terms, which had been satisfactory when the file had less than 2000 compounds, now failed to provide the desired selectivity in its classification of chemical structures. George P. Hager surveyed the existing systems for chemical structure coding and chose to adapt the structure fragment code of the Chemical-Biological Coordination Center (CBCC) to the specific interests of Smith Kline & French Laboratories. His decision was based on the ease of use of fragments, since they are basically the same structure groups encountered in nomenclature and structural chemistry. Important background information for this task was provided by CBCC in the form of a study which was never published in the scientific literature.[3] It reported the frequency of assignments made for the 212 most commonly used fragment codes, based on the results of encoding more than 44,000 compounds into the CBCC system. The code itself was described in a publication issued in 1950.[4] Using these data, Hager expanded the coding provisions for those chemical categories which appeared to be inadequately classified. On an over-all basis, the number of code terms assigned by Hager was approximately 1½ times the number in the original CBCC code. This revised code is called the SAC code.

Even more important than the changes in coding provisions were the improvements made in the CBCC format of the 80-column punched card, which enabled the much greater search capabilities of the multicolumn card sorters of the IBM-101 type to be realized. One of the greatest drawbacks to the efficient use and general acceptance of the original CBCC code was its card format, which was designed expressly for use by a single column card sorter. This was a serious limitation. A revised card layout was developed by Eugene Garfield in cooperation with Robert L. Hayne.

Like the CBCC chemical code, the SK&F modified code consists of a series of 4-digit α-numeric descriptors. An example of such a descriptor is F541 (Figure 1). The first digit, F, indicates that the functional group being described contains only nitrogen atoms in addition to carbon and hydrogen. The second digit, 5, when used in conjunction with F, indicates that an amine function is being described. When the third digit is added to give F54, the amine function is described as a tertiary aliphatic amine. Finally, the addition of the fourth digit (giving the complete descriptor, F541) indicates that the group described occurs one time in the molecule being coded. Thus, the code becomes increasingly more specific as 1-digit, 2-digit, 3-digit, or 4-digit descriptors are assigned. Conversely, the generic nature of the code decreases as the descriptor is increased from 1 to 4 digits.

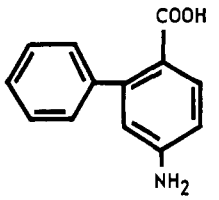
A sample compound, 4-amino-2-phenylbenzoic acid, is assigned the code terms described in Figure 2. No provision is made to indicate the particular positional isomer in this code. Since the code was designed to bring together compounds which are closely related so that they could be considered for testing in specific biological test systems, we did not feel it was necessary to distinguish between structural isomers. After some four years of experience in the use of the SAC code, during which about 1500

Descriptor "F541"

1. F = only nitrogen atoms in functional group
2. F5 = amine function
3. F54 = A III° – aliphatic amine
4. F541 = above group occurs once in molecule

Figure 1.

* Presented before the Division of Chemical Literature, 156th Meeting, ACS, Atlantic City, N.J., Sept. 13, 1968



SAC

F5L1 = 1-primary aromatic amine function
HC11 = 1-aromatic carboxylic acid function
RW71 = Potential zwitterion structure
NY82 = 2-isolated phenyl rings
0001 = 1-carbon atom isolated in a functional group

Figure 2.

searches were run, we added a few code terms which enabled us to distinguish between isoquinolines and quinolines, isoindoles and indoles, etc. Experience gained independently from using a different type of fragment code for several years helped us to make more fundamental changes and additions to the SAC code.

DEVELOPMENT OF THE EXTENSION CODE

In 1960, SK&F Laboratories became a member of the pharmaceutical Dokumentationsring. This group, comprised of one Swiss, one American, and four German pharmaceutical companies, agreed to use the same chemical and biological punched card codes on a cooperative basis. By the end of 1962, over 75,000 abstracts and 225,000 punched cards had been prepared and distributed to each Ring member. In 1963, SK&F changed its operation of the Dokumentationsring file to a computer basis because of the sheer bulk of the data involved. By then, we had run over 500 card searches of the Ring files, most of them involving the chemical structure codes, and were able to contrast the coding approaches used in the Ring chemical codes with those of the SAC chemical code. We then incorporated 69 of the Ring code terms directly into our SK&F system. In addition, our experiences with the efficiency of retrieval with both systems during hundreds of searches pointed out the need for new types of fragment codes, 38 of which were designed to allow us to run searches which could not be carried out by either the Dokumentationsring or SAC codes. The 107 code terms added were called the Extension Code, and were ready for use in 1965, after the backlog of some 25,000 compounds had been recoded (Table I).

Application of the Extension Code to 4-amino-2-phenylbenzoic acid adds the eight code terms shown in Figure 3.

COMPARISON OF SAC AND EXTENSION CHEMICAL CODES

The original CBCC code and the SAC version of it were strong in their provisions for functional groups, especially in terms of specific retrieval. For example, over 485 of the 3-digit descriptor terms had been assigned

Table I. Extension Code Sheet

SK&F No. _____

	69 General Rings	70 Aromatic Rings	71 Heterocyclic Rings: Isolated het.	72 Heterocyclic Rings: Poly	73 Heterocyclic Rings: Poly	74 Substituents: Poly	75 Ring Relationships: other polysub.	76 Ring Relationships	77 Chain Relationships	78 Chain Relationships
Y	1	Isolated arom.	Isolated het.	Poly	Poly	Poly	other polysub.	(X, H, R, Rg) Rg–C–C–Rg	R^1–CH_2–X,Rg	(H, R, Z, Rg)
X	2	Cond. arom.	Cond. het.	≥ 2 like het. atoms same rg.	het. atom at pt. of fusion	α to pt. of fusion / carbocyclic and monohetero rings	Geminal Subst.	(X, H, R, Rg) Rg–C–C–C–Rg	R–C–X,Rg	(H, R, Z, Rg)
0	3	Cond. alicycle		≥ 2 diff. het. atoms same rg.	het. atom α to pt. of fusion / monohetero rings	β to pt. of fusion / carbocyclic and monohetero rings	Ring=Ring	Rg–C–(H,R)	R–C–X,Rg	
1	4	Cond. het.		1N	het. atom β to pt. of fusion / monohetero rings	α to het. atom		Rg–C–X		Allylic–X
2	≥ 5	1		2N	1,2	β to het. atom		Rg–C=X		
3	Bridged Rgs.	2		≥ 3N	1,3	γ to het. atom		Rg–C=X		
4	Ang. sub.	3		1-0	1, ≥4	ortho (1,2)				
5	Central ring even	≥ 4		≥ 2-0	≥ 3 symm.	meta (1,3)				
6	Central ring odd	Alicycle isol.		5	≥ 3 asymm.	para (1,4)				
7	Peri. fus.	Cond. alicycle		only one het. atom in ring	≥ 3 vic.	vicinal (1,2,3)				
8	Ang.-Ang. fus.	Cond. het.	≥ 7	het. atoms other than NOS	het. atom subst.	1,2,4				
9		Rg size not 5 or 6	Δ, □			1,3,5				

LEGEND:
X = any element other than C or H.
Y = any element other than C or H, but not X.
Z = X or Y.
A = any element (C, H, X, Y) or Rg.
Rg = ring of any type.

 = ring of less than maximum unsaturation.

= optional ring of less than maximum unsaturation which may encompass the parts of the structure shown. *It can be interpreted as double bond in 76/Y, X and 78/0.*

R = C-atom in chain or in ring of less than maximum unsaturation.
R^1 = Carbon chain of 3 or more C-atoms.

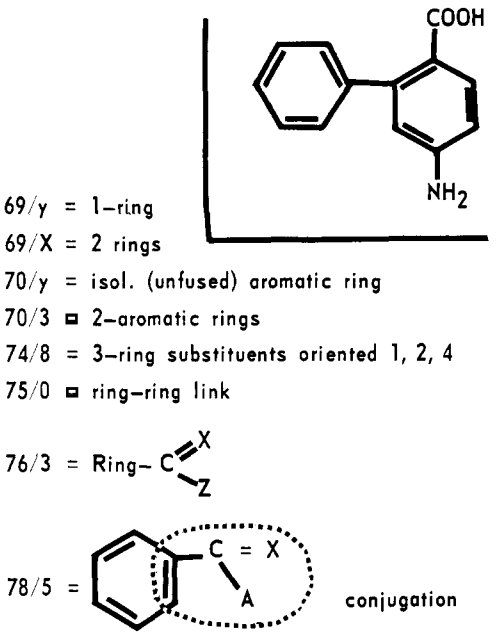


Figure 3.

to fewer than 30 compounds out of 25,000. However, the approach that was used did not lend itself to easy generic retrieval of functional groups; in addition, it was especially difficult to search for ring systems in a broad generic sense. The Dokumentationsring codes were much stronger in their provisions for generic retrieval of both ring systems and functional groups, but were weak in terms of their provisions for specific retrieval of complex functional groups. Thus, portions of the codes complemented each other very nicely. The combined code now enables us to search either specifically or broadly on graded levels of generic depth. Based on some 1500 searches with the combined system, we roughly estimate that we can obtain 80 to 90% relevance, compared with about 40 to 50% relevance when the SAC or Ring codes are used alone.

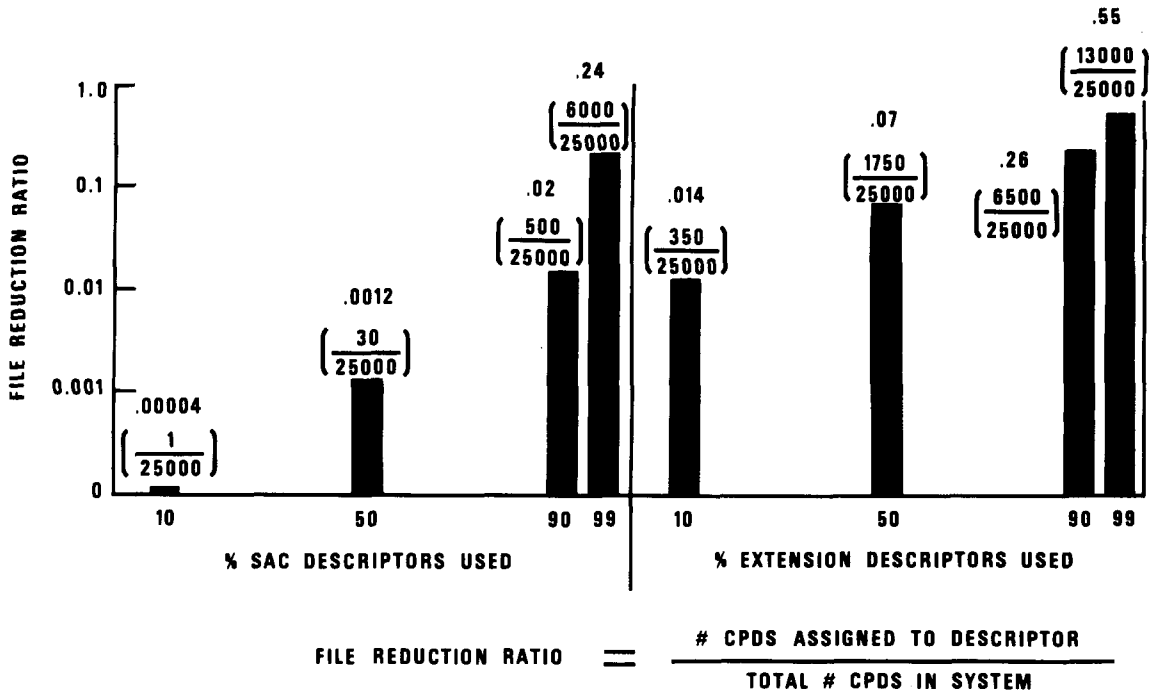
A comparison of the ability of the two portions of the code to reduce the total file of compounds in a search is shown in Table II. This file reduction ratio, which is defined as the ratio of the number of compounds assigned a descriptor to the total number of compounds in the system, is plotted against percentages of descriptors assigned from each portion of the code. The specific nature of the modified CBCC portion of the code is apparent when it is noted that 50% of the descriptors assigned reduce the file by a factor of 0.0012 or less. This point is further demonstrated by the fact that 90% of the descriptors assigned from the SAC code can reduce the file in searching by a factor of 0.02 or less. Similar observations for the Extention Code show that, in contrast, this code is more generic in nature, since small percentages of the descriptors have large file reduction ratios.

Comparing the 100 most frequently assigned descriptors with the 100 descriptors most frequently chosen in search requests shows that, in general, those most frequently assigned are also most frequently used in searching; 55 terms appear on both lists. Such a study also is helpful for pruning code terms, since terms frequently assigned and never used in searching are just dead wood.

PUNCHED CARD FORMAT

Prior to 1966, the information was searched by means of IBM punched cards, using an IBM 101 or 108 sorter. Our method for easy searching on punched cards is contained in columns 43 to 67 (Figure 4). To expedite generic searching, on the basis of frequency of assignment of 3-digit descriptors, 96 of the most frequently occurring 3-digit descriptors were arbitrarily assigned individual punches in this field (known as the 3-digit direct field).

Table II. Specific and Generic Nature of Codes



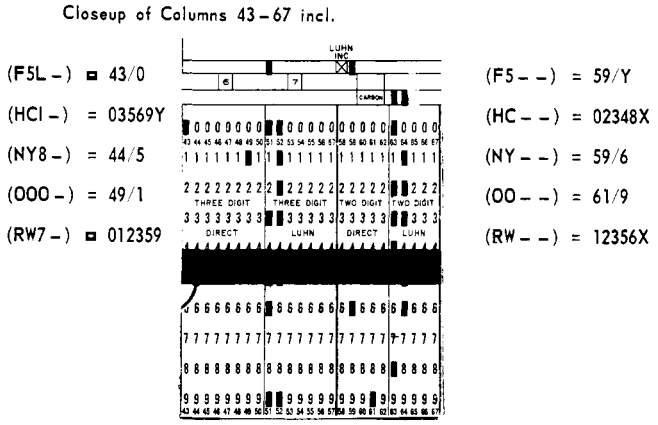


Figure 4.

For example, the 3-digit descriptor F51 occurs with high frequency and, whenever it was assigned, a punch was made in column 43 in the zero row. The same was true for NY8, a benzene ring, which was always punched in column 44, row 5. Similarly, the isolated carbon atom in the functional group was punched in column 49, row 1. HC1, which refers to the carboxylic acid group, did not occur among the 96 most frequently encoded 3-digit descriptors, nor did the zwitterion code, RW7. Therefore, these two terms which occurred less frequently were assigned random 6-digit numbers. These are referred to as Luhn codes, named for Hans Peter Luhn (1896–1964), who first proposed their use for an experimental photoptic sorter. Any molecule which was assigned an HC1 term was given a Luhn code 03569Y; the code for RW7 was 012359.

These are subsets of 6 out of 12 possible digits in any single column. There are 924 such subsets, and the provision of seven columns allows us to record any seven of these Luhn codes; these, plus the direct punch provision for 96 of the most frequently occurring 3-digit codes, allowed us to request any desired Boolean logic in searching by using only a few selector switches. Exactly the same principal was used to handle the 2-digit direct punches (the F5, the NY, and the 00). Since there are fewer 2-digit descriptors than 3-digit descriptors, only 60 fixed punches were assigned for the most frequently occurring 2-digit descriptors. Again, the HC and the RW codes received Luhn codes, and these were punched in columns 63 through 67. For explicit details of how to use these codes in searching, contact R. L. Hayne or P. R. Ackley at the authors' address.

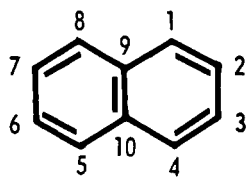
PROCESSING TIMES

The time required for coding the compounds and processing the punched cards according to the procedures just described are as follows: Assigning the 4-digit descriptors and extension codes—including checking these assignments—requires about 10 hours for 100 compounds, or 6 minutes per compound. These assignments are made by chemists. Assigning direct-punch and Luhn codes is done by clerical help. In this step, approximately 3 minutes per compound is required for assigning and checking. Punching and verifying of the IBM cards requires 4 to 6 minutes per compound. The time for total processing, then, is about 15 minutes per compound. Our recent switch to computer storage and retrieval of structural information has eliminated the assignment of Luhn and direct punch codes and a corresponding portion of the card punching operation. As a result, processing time has been cut to about 10 minutes per compound.

TYPICAL SEARCHES

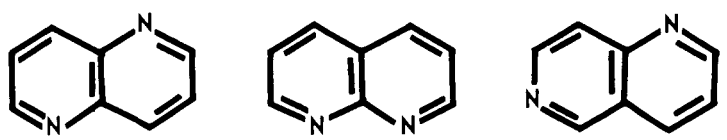
A search for naphthyridines illustrates the problems of generic retrieval and also describes the increased versatility which our combined codes allow over the original SAC code.

NAPHTHYRIDINES (1)



1 – Hetero – nitrogen atom in each ring

e.g.



etc.

Figure 5.

The SAC descriptors used (Figure 5) required that there should be two or more different types of 6-membered fused heterocyclic rings, each containing one nitrogen atom, or that there should be two or more identical types of 6-membered fused heterocyclic rings, each containing one nitrogen atom. These codes retrieved more than 600 compounds, most of which were not pertinent. Typical "false drops" included compounds of the type shown in Figure 6. In each case, the codes were assigned correctly, but the structures were not those of the naphthyridine ring system. Including certain extension codes in the search helps eliminate false drops. Using the codes 69/X, indicating a 2-ring system, and 71/X, indicating that a hetero ring must be condensed with another hetero ring, and excluding the 73/X punch, which indicates that a hetero atom is at a fusion point (see Table I for details), the file was reduced to 83 compounds, 61 of which were relevant.

The term 'relevancy' or 'pertinency' as used to describe the results of a search is a most subjective parameter and should be taken with a grain of salt. From time to time, the same person will consider something to be either relevant or not relevant, depending on his particular interest at that moment.

By adding the provision that the 69/X punch for 2-ring systems must be present, we automatically eliminated any benzonaphthyridines. The information scientists designing the search must be aware of the results of such restrictions, and must convey these limitations to the person making the request. A number of bridged ring alkaloids, such as strychnine and the lupinane alkaloids, were still answering to these descriptors. We negated the term "bridged ring systems," and received a return of 66 compounds, with all 61 from the last version of the search present, so the relevancy was increased to 92%.

Searches for the aminoalkyl esters of diphenylacetic acid, known as Trasentin analogs (Figure 7), illustrate basic points. In the SAC system, we searched for tertiary aliphatic amines, alkyl esters of an aliphatic acid, and two or more unfused benzene rings. Of the 148 compounds returned, many had a hetero atom substituent 'X' in place of the encircled hydrogen in Figure 7. These are benzilic acid esters, or chloro or amino analogs, etc. If these are considered as "false drops," then the relevancy was 25%. The addition of three extension codes, again relating ring systems to each other and to other functional groups, reduced the number of compounds to 57, with the same 37 compounds being relevant; this increased the relevance to 65% (Figure 8). Again, there were 20

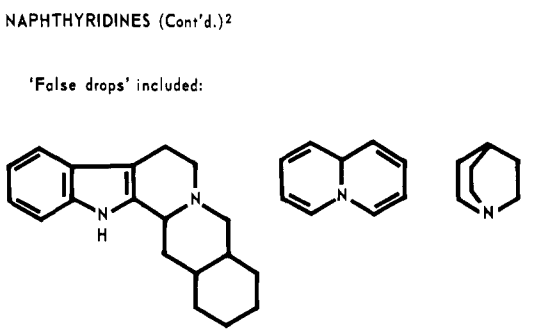


Figure 6.

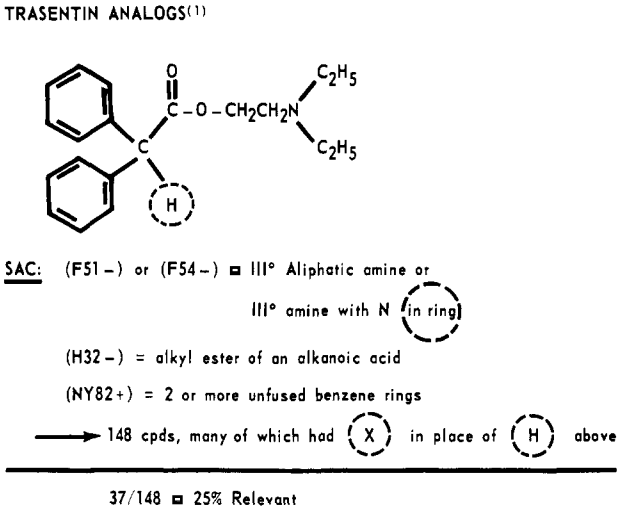


Figure 7.

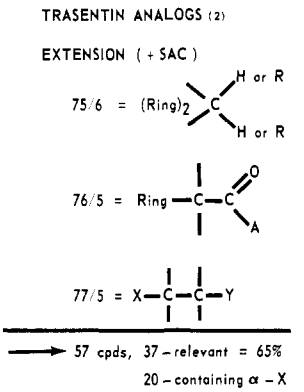


Figure 8.

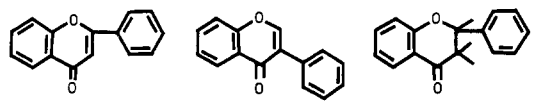
"false drops" containing an α-heteroatom, yet these still met the described code restrictions, since the structural requirement (75/6), which does not allow for α-halogen, was fulfilled elsewhere in the molecule.

A search for flavone analogs is illustrated in Figure 9. Using only the SAC 4-digit descriptors resulted in a 65% relevant return. Adding the two extension terms shown increased relevance to 93% (Figure 10).

OPERATION OF SYSTEMS

To conduct a search of this type, the original request is converted into the language of the SAC and Extension Codes by an information scientist who is familiar with these codes. Sorting the punched card file for 25,000 compounds requires about 45 minutes per search. Since January 1966, our chemical searches have been run on a computer, and the file now has approximately 40,000 compounds. Writing the desired search terms in compiler language (done by an information scientist) requires 1 to 2 minutes per search, key punching takes less than a minute, and the cards serve as input to the computer. A typical computer run of 15 to 25 searches requires about one hour on the IBM 360/40, operating in 1410 mode. This will be shortened when programs are rewritten for the 360/40.

FLAVONE ANALOGS



(1N4–) or (1N6–) = unsat. fused 6-membered ring with 1-oxygen in ring

(NY6–) = at least one fused benzene ring

(NY8–) = at least one unfused benzene ring

→65 99 compounds = 65% relevant

Figure 9.

FLAVONES, ETC.

69/X or (1F6–) = either 2-rings or a methylenedioxy ring

75/O = ring-ring

→ 64/69 relevant or 93%

1 of the 65 previously retrieved was lost.

Figure 10.

Since the retrieval system is run entirely by computer, we no longer assign Luhn punches and direct digit punches; the computer program needs only the 4-digit descriptors to let us search for 1-, 2-, 3-, or 4-digit descriptors at will. The details of the codes are identical, however, with the system which was designed originally for punched cards. The computer can serve as a superfast card sorter for a fragmentation code system, and when used in this simple manner, computer searching is extremely rapid and very inexpensive for a file of this size. We realize that this would not be true if our file were 100 times larger, which is the problem faced by Chemical Abstracts. About two years ago we made available the details concerning our codes, frequency of assignment, etc., to Chemical Abstracts for their consideration in planning screens to reduce the amount of atom-by-atom searching necessary in their computerized structure retrieval system.

Originally the results of our punched card searches were received in the form of tabulations of SK&F code numbers. A clerk then pulled a corresponding 3 × 5 inch structure card for each compound, and the information chemist would edit this group of cards before sending the results to the laboratory scientist who made the original request. Undoubtedly, this editing procedure conditioned the scientists favorably toward this service.

A chemical name file has been added for each compound so that the high-speed printer prints the SK&F number and the chemical name as direct search output. In many cases, this eliminates the need to have a structure card manually selected for each compound. In the future, we will have the capability of printing the structural formulas as direct computer output.[5] From our experience, this will be desired in almost every case, since a structural formula is so much easier to read and conveys structural information more rapidly and more accurately than a chemical name.

ERRORS

The compounds are encoded by hand by experienced chemists, and checked at least briefly by another organic chemist. In our experience, encoding with the fragment code of the CBCC type is readily learned, and the assigning of codes is quite free from error. Conversely, since the codes themselves consist of four digits each, there is a greater possibility for error in key punching than in coding. To help minimize the key punching errors, everything is verified. However, key punch errors do occasionally slip by. A code like our Extension offers the reverse situation. Here the code is more intricate, and requires a lot of subjective thinking on the part of the chemist doing the coding. Conversely, each code term is represented by a single punch, reducing the potential errors in key punching.

Our best estimate is that the error rate is approximately one miscoding per 1000 codes assigned. We estimate the relative amount of error caused by key punching is about 1/10 the coding error rate. To reduce this error rate, we are working towards computerized coding from a chemical structure to be input by a typewriter. Here the errors should be limited to those of structure proofreading and, although the errors can never be reduced to zero, various computer checks should lower the rate much below the present rate which, however, is quite acceptable. We cannot pretend to have a perfect system as long as we rely on human effort to encode the compounds and to operate the system.

ACKNOWLEDGMENT

The authors gratefully acknowledge the important contributions made by George P. Hager, who laid the foundation for the structure-retrieval system at SK&F Laboratories. We also appreciate the help and encouragement given by Maxwell Gordon.

LITERATURE CITED

(1) National Academy of Sciences–National Research Council, Washington, D. C., "Survey of Chemical Notation Systems," Publication **1150**, 1964.

(2) National Academy of Sciences–National Research Council, Washington, D. C., "Survey of European Non-Conventional Chemical Notation Systems," Publication **1278**, 1965.

(3) Dale, Esteleta, and Karl F. Heumann, "Statistical Information on Component Parts of Chemical Compounds," National Academy of Sciences–National Research Council, Chemical-Biological Coordination Center, Washington, D. C., 1955.

(4) National Academy of Sciences–National Research Council, Chemical-Biological Coordination Center, Washington, D. C., "A Method of Coding Chemicals for Correlation and Classification," 1950.

(5) Gottardi, R., Division of Chemical Literature, 157th Meeting, ACS, Minneapolis, Minn., April 1969.