A Program Retrieval of Organic Structure Information Via Punched-Cards*

ALAN GELBERG, WILLIAM NELSON,*[a] GEORGE S. YEE,*[b] and E. A. METCALF

Industrial Liaison Office, Directorate of Research, U. S. Army Chemical Research and Development Laboratories, Army Chemical Center, Maryland

Received August 24, 1961

The Industrial Liaison Office (ILO) was established by the U. S. Army Chemical Corps to solicit from industry data, know-how, and ideas to enhance the Corps' research and development efforts. Information obtained through this program has provided leads to the solution of some of the most urgent problems and has prevented duplication by the Chemical Corps of costly research already completed in industrial laboratories but which has not been publicly reported.

More than 90% of the information sent to the ILO consists of proprietary lists of recently prepared compounds with attached physical, chemical, and biological screening data. At the present time, lists containing approximately 1000 structures are received each month. These structures are reviewed by the key Chemical Corps scientists responsible for directing the research, development, and screening programs. Based on the needs of their programs, selected samples are requested from the cooperating companies. It is anticipated that more than 1000 samples, acquired through the ILO, will have been studied for possible application in the Chemical Corps programs during fiscal year 1961. Unclassified information, developed from the study of these compounds, is sent only to the donor of the sample. Naturally, the ILO will not release the data to other groups without the written approval of the company submitting the original sample.

Upon receipt of information or a chemical structure, the name of the source is removed from the material and an accession number [Commercial Source (CS) number] is assigned. This number has no connection with the source; it merely facilitates internal handling. The information is assigned a level of confidentiality in accordance with the directions of the submitter, which indicates the safeguards that must be applied during its processing and handling.

The material is prepared for dissemination and entrance into the office storage and retrieval system: (1) The structures of chemical compounds are drawn if they have not been supplied. (2) Submitted names are used; no attempt is made to rename or assign a missing name to an item. (3) An inverted molecular formula is prepared. (4) Pertinent data are extracted and organized. (5) A chemical line-notation is prepared. (6) A program sheet is prepared for IBM punch-cards. Except for typing and key-punching, it takes an average of seven minutes to prepare the above information for each compound. This average is considerably influenced by the fact that most of the complex molecules submitted are part of a series. About 30 seconds are required for a Wiswesser line-notation and two to three minutes for a structure analysis for the punch-card program.

THE NEED FOR AN INFORMATION RETRIEVAL SYSTEM. In 1957, it was recognized that a system would be required to rapidly retrieve information on the chemical structures submitted to the ILO. From discussions with information scientists and by carefully evaluating their ideas, systems, approaches, and suggestions, it appeared that no one proposal met the needs of this office. However, some degree of automation did seem necessary. In 1959, the U. S. Army Research and Development Command gave approval for a study to be performed for organizing a chemical compound file which could be applied throughout the U. S. Army Chemical Corps. The chemical compound file of the ILO was selected for this study.

In order to keep within a limited budget, a multi-thousand dollar program could not be considered. Readily available equipment, without modification, had to be used. Inexpensive accounting machines were being used successfully for programs of this type by a number of companies. A simple sorter and key-punch were the only machines basically required. The more elaborate operations, such as reproduction of decks, collation, and tabulation, could be performed when required, as a service, by the U. S. Army Chemical Corps Data Processing Center located at this installation.

An IBM 82 Sorter was selected for this program because it is inexpensive to rent, simple to operate, sorts approximately 650 cards per minute, and has the essential feature of being able to select any hole in any column. The second piece of equipment, the IBM 26 Printing Card Punch, was chosen because it prints both alphabetic as well as numeric characters at the top edge of the card, is inexpensive to rent, simple to operate, and allows for multiple punching within a column without loss of the number of available columns.

THE INVERTED MOLECULAR FORMULA FILE. The inverted molecular formula has been adopted because it permits either the immediate location of a specific compound or a group of compounds containing a rare element. For instance, most all organotin or organoboron compounds can be retrieved at will from this file. The frequency of occurrence of the elements in organic compounds is the basis for this arrangement (1). The most common elements, in increasing order of occurrence, are P, halogens, S, N, O, C, and H.

It was, therefore, desirable in this file to arrange the compounds in the order of elemental compositions: Rarer elements (in alphabetical order), P, F, Cl, Br, I, X, S, N, O, C, H, R, Miscellaneous. R represents an unnamed group that may be attached as a fragment to a molecule (such as, an undetermined polysaccharide attached to an adenosine group) and miscellaneous includes completely unknown structures (such as antibiotic brews, natural products, or reaction mixtures that have not been analyzed completely). The hydrogen count is eliminated in any compound containing at least three elements, exclusive of hydrogen, because the hydrogen count only indicates the degree of unsaturation (which may or may not be important). If there are only three elements in a compound, including hydrogen, the hydrogen count is included to assist in the filing and look-up. Acid or base salts are either listed as the parent compound, or are doubly entered if the salt former is not apparent. A 3 × 5 inch card is prepared for each compound. It contains a number, an inverted molecular formula, submitted name, a drawn structure, and a line-notation. These cards are assembled in the inverted molecular formula order. This comprises the molecular formula file.

THE WISWESSER CHEMICAL LINE-NOTATION. The problem of placing a representative structure on an IBM card and how it could best be utilized was studied. Hand-drawn or typed structures appeared to be wasteful from the viewpoint of original input-time and the duplication of this waste when deck-replacement was required. Lack of a structure representation on a card meant continual file searching for the structure. The independent and almost identical approaches followed by Benson (2), Smith (3) and Bonnett (4) appeared to be the most logical solution.

These investigators used the Wiswesser chemical line-notation (5) as a language that proved to be intelligible to both a machine and a chemist. In addition, standard unmodified accounting equipment could be used and the possible misinterpretation of names would not have to be considered. Furthermore, line-notations require considerably less space than the chemical name would occupy on a punch-card. Any chemist, after a short training period, can convert a line-notation to the familiar two-dimensional representation that chemists are trained to recognize. Organization of a file of punch-cards on the basis of the Wiswesser chemical line-notation automatically arranges the carbocyclic and heterocyclic structures into well defined groups. This becomes readily apparent when tabulated listings are prepared. These lists supplement the molecular formula file for rapid specific compound look-up as well as grouping the cyclic structures. The value of tabulated indexes has been discussed by Bernier (6) and Bonnett (4).

More than 90% of the line-notations of chemical structures submitted to the ILO occupy fewer than 20 columns of an IBM card; however, a field of 35 columns (9-43) is retained for the line-notations in view of dyes and other complex structures.

PUNCH-CARD DECK ARRANGEMENT-CLASSIFICATION NUMBER. In order to arrange the punch-cards into a useful order based on structure, a two-symbol classification number is assigned for use in partitioning the file. This number appears in columns 44 and 45. Column 44 is a ring index (see Fig. 1). This classification symbol is based on the following: (1) A separate card is prepared for each classification number, (2) Absence of a symbol indicates either an aliphatic, inorganic, or unknown structure (3) Degrees of unsaturation are not considered, only the ring skeleton, (4) The hetero atoms, number, positions, etc. are not defined in this area. The only consideration is the presence or absence of a hetero atom in the ring. (5) Fused ring systems are assumed to contain one six-membered carbon ring which need not be classified, *e.g.*, a five-membered ring (an indane) is assigned one classification number (a 7 punch) while a four-membered carbon ring fused to a five-membered carbon ring is assigned two classification numbers (a 6 or 7 punch), one on each of two cards. (6) Spiranes are treated as fused rings. (7) Simple bridged cyclics are considered as monocyclics unless fused to another ring.

Fig. 1. Ring Index for Column 44

Unfused carbocyclics:

Punch-Position	Meaning
0	Cyclopropane (also used for unfused 3-membered heterocyclic compounds)
1	A cyclobutane
2	A cyclopentane
3	A cyclohexane
4	A cycloalkane containing 7 or more atoms

Fused carbocyclics:

Punch-Position	Meaning
5	A three membered ring fused to one or more rings
6	A four membered ring fused to one or more rings
7	A five membered ring fused to one or more rings (indanes, fluorenes, steroids)
8	A six membered ring fused to one or more rings (naphthalenes, anthracenes, phenanthrenes, etc.)
9	A seven or more membered ring fused to at least one other ring

Unfused heterocyclics:

Punch-Position	Meaning
0	A three membered ring (also used for cyclopropane)
A	A four membered ring
B	A five membered ring (furanes, pyrroles, thiophenes)
C	A six membered ring (dioxanes, piperidines, pyridines, etc.)
D	Seven or more atoms in a ring

Two fused rings, one of which is heterocyclic:

Punch-Position	Meaning
E	A three membered ring fused to another ring
F	A four membered ring fused to another ring
G	A five membered ring fused to another ring (indoles)
H	A six membered ring fused to another ring (quinolines)
I	A seven or more membered ring fused to another ring.

Three fused rings, one of which is heterocyclic:

Punch-Position	Meaning
N	A three membered ring fused to two other rings
O	A four membered ring fused to two other rings
P	A five membered ring fused to two other rings (naphthoxazoles and carbazoles)
Q	A six membered ring fused to two other rings (phenothiazines)
R	A seven or more membered ring fused to two other rings

Four or more fused rings, one of which is heterocyclic:

Punch-Position	Meaning
V	A three membered ring fused to three or more rings
W	A four membered ring fused to three or more rings
X	A five membered ring fused to three or more rings
Y	A six membered ring fused to three or more rings
Z	A seven membered ring fused to three or more rings

The symbol in column 44 permits the very rapid selection of all basically related ring skeletons. A sort of the seven row in column 44, for example, will separate all of the indanes, fluorenes, steroids, indoles, benzimidazoles, carbazoles, *etc.*, from the deck. All other hetero-atom-containing molecules having a five-membered fused ring will also be separated.

The second symbol of the classification number (column 45) Fig. 2 is related to the elemental composition of the total molecule.

Fig. 2. Classification Number for Column 45

Punch-Position	Meaning
0	CH or CHO
1	CHO and 1 N
2	CHO and 2 or more N's
3	CHN or CHS
4	CHOS
5	CHONS
6	CHX or CHOX (X = halogen)
7	CHXN or CHOXN or CHXS or CHXOS or CHXNS or CHXONS
8	Uncommon rare elements
9	Inorganic compound

This column complements the first symbol of the classification number by: (1) providing a means for a rapid selection of general type molecules; (2) simplifying the deck arrangement for filing. An additional column, number 46, has been left blank for possible expansion of the classification number.

STRUCTURE ANALYSIS. The last field of the card is the machine-search area for functional groups. The fragment of the molecule under consideration is determined by a symbol in column 47 which, following suggestions by Bonnett (4) and Wiswesser (5a) is called a "prefix" (Fig. 3). For use in this analysis area, the term "Ali-Mode" has been coined to describe an approach wherein a molecule is fragmented and each fragment is analyzed. Only cyclic moieties are fragmented. For each fragment a separate punch-card is prepared. In addition, a summary type or "Ali-Mode" card is prepared wherein the molecule is "stretched out" to yield a semblance of an aliphatic appearing structure. This procedure involves breaking selected bonds of the ring structures to show the spatial relationships of functions and atoms. On an average, two or three cards are prepared for each compound.

The same CS number and line-notation (columns 1 through 43) will appear on each card. As previously mentioned, the classification number may change if more than one ring system is present in the molecule. For very general type searches, the "ampersand" deck is searched

Fig. 3. Molecular Fragment Code for Column 47

Punch-Position	Meaning
&	Ali-Mode analysis and aliphatic compounds
-	Natural product or polymer
0	Benzene ring
1	Monocyclic other than benzene
2	Bicyclic
3	Three fused rings
4	Four fused rings
5	Five fused rings
6	Six or more fused rings
7	Organo-metallic, chelate, or metallic complex
8	Spirane
9	Bridged cyclic

with the sorter. If a query is for a specific nucleus, the classification guides the search-pattern for sorting. This "sort" yields not only the particular nucleus but will yield other similar type ring systems, as well.

In order to search for compounds containing the same functional groups, many approaches have been reviewed. The cross-reference approach described by Smith (3) was adopted, but the functional group field was expanded to seven columns. Multiple punching within a column is performed. This becomes a little troublesome when the card must be reproduced since the key-punch will only accurately reproduce three punches within a column.

Columns 48 and 49 relate to the composition of a ring and the position of the hetero atoms (Fig. 4). Column 50 (Fig. 5), describes the positions of functions directly attached to the segment of the molecule defined by the prefix. The next four columns (51-54) describe the functional groups and combinations of atoms which are either part of the fragment described or directly attached to it. For example: a function such as a carbonyl group can be attached to carbon atoms only on either side of the function, which by definition is a ketone indicated by 53/1.

Figure 4.
Composition of Ring and Hetero Atom Position

— bond; = double bond; ≡ triple bond; C cyclic; A any atom.

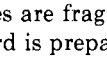
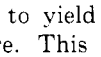
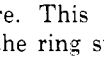
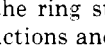
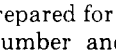
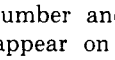
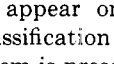
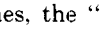
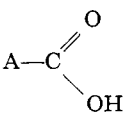
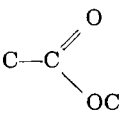
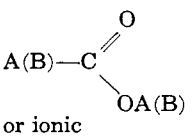
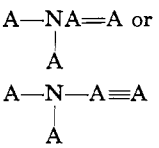
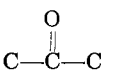
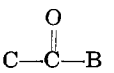
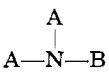
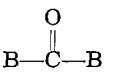
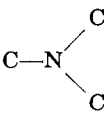
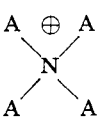
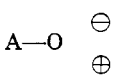
	Cyclic composition (48)	Arrangement (49)
12	Saturated ring	Hetero-ring fused to a hetero-ring
11	Phenyl ring	Hetero-ring fused to a non-hetero-ring
0	Atom other than C, S, N or O in a ring	Hetero-atom at a bridgehead
1		Atom in an arbitrary 1 position
2		Atom adjacent to 1 position
3		3
4		4
5		5
6		6
7	 Partial unsaturation	7
8	 Spirane	8
9	 Bridged cyclic	9

Figure 5. Functional Group Position

	(50)	(51)	(52)	(53)	(54)
	Ring bonded to another ring	Other valence state	A—NH—A=A or A—NH—A≡A		C—S—C
	Alkyl group on a ring	Radioactive atom	RN=A or C—NH—C		A—S—B
	1 or 2 atoms bridging two rings	Isomeric form	A—NH—B	 or ionic	A—S—A=A or A—S—A≡A
1		Cl, Br, I, X	 or A—N—A≡A		A=S
2		F	A—N=A or A—N≡A		A—H
3					C=C
4				A—OH	A=B
5		B		AR—OH	A≡A
6			A—NH_2	B=O	
7		Si	AR—NH_2	C—O—C	
8		Other atom	A—NO_2	B—O—A	
9		Inorganic	AR—NO_2		Multiple occurrence of a functional group

— bond; = double bond; ≡ triple bond; C cyclic; A any atom; B atom other than carbon; C carbon atom; AR aromatic ring.

A carbonyl function can be attached to two atoms, one of which is not carbon (53/2), and this can be either an amide (cross-referenced with 52/12, 52/1, or 52/6), an aldehyde (54/2), a thiolo-acid (54/1) or thioloester, or perhaps an acid halide (51/1 or 51/2). For this system, cystine is programmed for the punch-card as

MK 316 QVYZ1S 2
$(HOOC\text{-}CH(NH_2)\text{-}CH_2\text{-}S)_2$

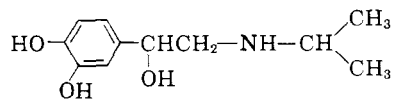
The MK number, in columns 1-8, relates to the page in the 7th Edition of Merck Index and is used rather than a CS number. Columns (9-43) are reserved for the line-notation.

Since the compound is aliphatic, column 44 is blank. The elemental composition of CHOSN is a 5 punch in column 45. Therefore, the classification number is (blank) 5.

Column 47 prefix is & for an aliphatic compound
48 blank (non-cyclic structure)
49 blank (non-cyclic structure)
50 punches 1, 2, 3, and 6 are functions related to each other by atom distances
51 blank
52 6, a primary amine group
53 12, a carboxylic acid group
54 11, a sulfur atom attached to at least one non-carbon atom
7, a tri-substituted carbon atom
9, multiple occurence of a function

A second example, Isoproterenol, MK 580, is programmed for two cards

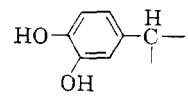
MK 580 QRQ DYQ1MY



The first card is for the benzenoid deck:

Column 1-8	MK 580
9-43	QRQ DYQ1MY
44-45	31
47	0

Columns 48-54 are analyzed for those groups directly attached to the phenyl ring



Column 48 11 indicates a phenyl ring
49 blank
50 11, alkyl substitution on the ring
1 through 4 describes either 3 or 4 groups substituted on the ring; 1 and 2 shows ortho substitutions; 1 and 3 shows meta; 1 and 4 shows para.
51 blank
52 blank
53 aromatic hydroxyl (5)
54 a trisubstituted carbon atom (7)
multiple occurrence of a function (9)

The second card prepared (MK 580) shows the Ali-Mode approach. Columns 1 through 45 are the same as the preceding card.

Column 47 &, Ali-Mode analysis
48 11, phenyl group in the molecule
49 blank
50 11, alkyl attachment of a ring
1 through 7 shows all combinations of atoms between which functions are attached
51 blank
52 11, a secondary amine attached only to carbon atoms
53 4, a hydroxyl group attached to an atom
5, a hydroxyl group attached to a ring
54 7, a tri-substituted carbon atom
9, multiple occurrence of a function

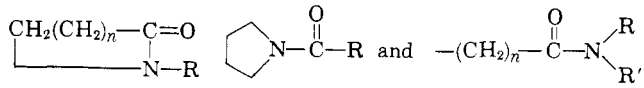
SEARCH PATTERN. The search of the chemical compound file depends upon the nature of the inquiry.

A. Specific compounds are located directly in either the inverted molecular formula file or from the tabulated listings in line-notation order. At present, only the former file is up to date, since only 13,000 compounds are represented on punch-cards.

B. For groups of compounds containing rarely-occurring elements, the inverted molecular formula file is scanned.

C. For groups of compounds containing alicyclic or heterocyclic systems, the tabulated lists are scanned (as one would use a telephone directory).

D. For compounds containing either specific functional groups or functional groups in specific (or general) relation to each other, the "Ali-Mode" deck is machine sorted. For example



would all be sorted from the selection of punches 52/1 and 53/2.

A requestor of chemical compound information will receive an answer either as: (a) a list of CS numbers which the requestor can check from his own file; or (b) drawn structures and accompanying CS numbers. Recently, a Xerox 914 Office Copier was installed in these laboratories. It has been found to be very convenient to pull the appropriate drawn structures from the Molecular Formula File, reproduce three of these selected 3 × 5 cards on an 8 × 11 page, and submit these to the inquirer.

GROSS ACTIVITY FIELD. At the present time, columns 55 through 79 are being left blank. These columns will be programmed to record gross activity of the compound and will be reported at a later date.

SUMMARY

The storage and retrieval of structural information on organic compounds in the Chemical Corps Industrial Liaison Office, requires only: (1) low-rental, easily available equipment; (2) a small number of technically trained personnel; (3) very short search periods to retrieve from storage: (a) all information concerning a specific compound if it is identified by name, structure, or accession (CS) number; (b) all specific compounds containing the same and/or related (generic) structures; (c) all specific compounds containing the same functional group(s); (d) all compounds having the same or specific functional groups in similar or identical special relationships to each other.

ACKNOWLEDGMENT. The authors are indebted to Mrs. R. Bosely for her careful editing and preparation of this manuscript; Specialist G. Johnson and Pfc. W. Lloyd, who assisted in the manuscript preparation; Lt. P.F. Sorter and Specialist L. Miller, who are programming the compounds for the punch-card file.

REFERENCES

(1) W.J. Wiswesser, "Advances in Chemistry Series," ACS No. 16 (1956).
(2) F. Benson, paper presented at the ACS National Meeting, September, 1953, Chicago, Illinois.
(3) E.G. Smith, *Science*, **131**, No. 3394, 142 (1960).
(4) H.T. Bonnett and D.W. Calhoun, paper presented at the 139th ACS National Meeting, Division of Chemical Literature, March 27, 1961, St. Louis, Missouri.
(5) W.J. Wiswesser, "A Line-Formula Chemical Notation," T.Y. Crowell, Co., New York, N. Y., 1954, (5a) p. 120.
(6) C.L. Bernier, *J. Chem. Doc.*, 1, 67 (1961).

*Presented at the 139th ACS National Meeting, Division of Chemical Literature, March 27, 1961, St. Louis, Missouri.

[a] CBR School, Dugway Proving Ground, Dugway, Utah.
[b] U. S. Army Chemical Corps Bio. Labs., Fort Detrick, Md.