Rapid Structure Searches *via* Permuted Chemical Line-Notations*

By PETER F. SORTER,** CHARLES E. GRANITO, JOHN C. GILMER, ALAN GELBERG, and EDWARD A. METCALF
Industrial Liaison Office, Office of the Scientific Director, U. S. Army Chemical Research and Development Laboratories, Edgewood Arsenal, Maryland
Received August 5, 1963

The Wiswesser chemical line-notation is an unique and unambiguous method of representing chemical structures by a linear series of letters, numbers, ampersands, and hyphens.[1] These symbols are meaningful to chemists familiar with the notation and can be processed by automatic data processing (ADP) equipment.

The uniqueness of the line-notation permits the use of alphanumerically arranged lists of notations for dictionary-type searches. This ordered arrangement permits the rapid location of a specific compound or a specific class of ring compounds other than benzenoid.

Figure 1 represents a section of a tabulated list along with the structures for the notations given.

The six compounds, possessing the same ring structure, are grouped together and are easily located in such a list.

Compound no.	Notation
1	T6TM CUJ B2G
2	T6TMJ
3	T6TMJ B C2G
4	T6TMJ B2G
5	T6TMJ B2G CZ D3
6	T6TMJ B2MX CG D EZ
7	T6TMJ B2M2 CG D EZ

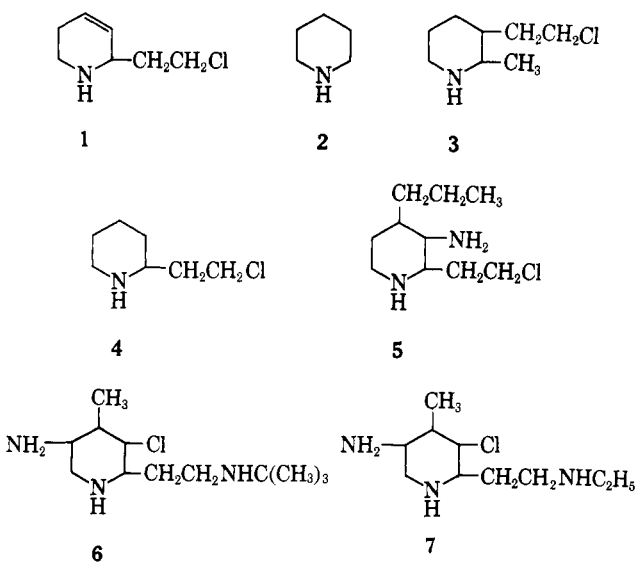


Fig. 1.—Section of a tabulated list indexed on initial symbol of notations.

In each instance, the notation begins with the symbols T6TM. The space after the fourth symbol in the notation, T6TM CUJ B2G (compound 1), places it before any notation in which the fifth position is occupied by a character. Thus, a slight change in structure could result in a wide separation of notations of similar compounds. A search for compounds possessing the same functional group or atom, *e.g.*, chlorine (represented by G), is obviously impractical. A functional group search would require the scanning of the line-notation for each compound in the entire list. To overcome this deficiency, it has been necessary to adopt functional group codes for use with IBM punch-card systems.[1,2] Naturally, if functional groups could be located by means of the line-notation, the need for such a special code would be obviated.

The problem of searching line-notations for functional groups is analogous to the selection of keywords in an ordinary list of the titles of scientific papers. The search for specific subjects in such a list has been shown to be facilitated by the preparation of alphabetic lists of permuted titles.[3] This method, first cited in 1856,[4] has recently come into wide use.[5]

A list of permutations of chemical line-notations alphabetized on individual symbols could be used to readily locate all compounds containing any specified functional group as well as specific compounds and specific classes of carbocyclic or heterocyclic structures. It is the intent of this paper to show that this procedure is applicable to line-notations and to their more efficient use.

In order to test the feasibility of this approach, an index was created for 120 unclassified compounds, selected from the Industrial Liaison Office punch-card files. Sixty of the compounds were selected at random and an additional sixty were selected on the basis of the symbols used in the notations in order to ensure the inclusion of symbols which might cause difficulty. The punch cards, standard 80-column IBM cards, were punched with an IBM 26 printing card punch. All of the sorting steps were carried out on an IBM 82 single column sorter.

For this study, the maximum length of line-notations used was 35 symbols. Those possessing more than 35 symbols were excluded only because our investigation required that any one notation when doubled in length should fit on one card (80 columns). However, the use of a computer or other procedures (to be reported at a later date) will eliminate this limitation.

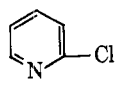
Columns 1–8 contained an identification number and columns 9–43 contained the notation. An IBM 514

* Presented before the Division of Chemical Literature, 145th National Meeting, New York, September 11, 1963.
** To whom inquiries should be addressed: Hoffman-LaRoche, Inc., Nutley, N. J.

reproducer was used to reproduce the original deck of 120 cards 35 times, with a one-column shift of the entire notation block (columns 9–43) for each individual reproduction. Thus, the second deck contained these same symbols in the column block 10–44, the third 11–45, etc. Columns 1–8 were reproduced in the same position on each card. When this reproduction was completed, each symbol of a given notation appeared in column 44 on one of the 35 cards bearing that notation. In this manner, 4200 cards were generated from the original 120 cards. However, as demonstrated later, only 806 of the 4200 cards generated proved to be of value for our purposes.

In Fig. 2, the permutation of the notation T6NJ BG (2-chloropyridine), is shown. Lines 1–7 represent separate punch cards. Actually, 35 cards were created from the successive shifts of columns 9–43. The 28 cards not shown did not have a notation that included column 44. It is anticipated that permutation performed by a computer would be programmed so that only characters comprising the notation would be permuted, with disregard for spaces following a notation.

It can be seen from Fig. 2 that column 44 could be used for setting up an index.



T6NJ BG

						44								
					T	6	N	J		B	G			1
				T	6	N	J		B	G				2
			T	6	N	J		B	G					3
		T	6	N	J		B	G						4
	T	6	N	J		B	G							5
T	6	N	J		B	G								6
T	6	N	J		B	G								7

Fig. 2.—Seven of the 35 cards for 2-chloropyridine.

If all of the columns from 44 to 79 (and perhaps 9–43 to give a "wrap around" feature as found in *Chemical Titles*) are placed in alphanumeric order, then this compound could be found in 35 locations in a permuted index. In this example, only six of the entries might be considered to be of value, since column 44 is blank on the remaining 29 cards. In permuted titles, only preselected key-words are indexed; all others are excluded in accordance with established rules. For example, words such as "and," "the," "but," etc. are not included in order to produce a list which is indexed only on words of interest and which can be searched rapidly. For permuted line-notations, exclusion rules are also required to avoid indexing symbols for which one would rarely search.

Our exclusion rules, other than the obvious deletion of spaces, are based on the requirements of the Wiswesser line-notation symbology. These are described hereinafter and summarized in Table I. In this notation, a letter preceded by a space represents the location of a substituent on a ring. These locants were not selected as index symbols inasmuch as the searches conducted by the Industrial Liaison Office require that more significance be given to functional groups than to their specific location. However, in some areas, *e.g.*, the pharmaceutical industry, locants would be of greater interest. Bonnett,[6] who has employed tabulated lists of notations for some time at G. D. Searle & Co., is interested in the approach of indexing *on* the locants. Such an index would, for example, permit the location of all steroids substituted at a specific position. Figure 3 shows how steroids substituted at C-17 (locant F) could be rapidly found by looking under the F's in column 44. It should be emphasized that the selection of symbols to be indexed is arbitrary and can be varied according to the needs of the user.

Approximately 2350 cards, of the original 4200, were eliminated by sorting out those which had a space in column 44. Removal of those cards having a locant symbol in column 44 eliminated another 250 cards. The remaining cards (*ca.* 1700) were sorted into order from columns 79 to 44. For this column-by-column arrangement, the order of priority was: space, letters (in alphabetical order), numerals (in numerical sequence), ampersands, and hyphens. These ordered cards were used to print a tabulated list of the notations by means of an IBM 407 accounting machine. Close scrutiny of this list revealed that each of the following symbols need not be indexed: A, J, X, Y, &, -, T when preceded by a number, and numerals (see Table I).

The J, &, and - do not represent functions and can be dropped without impairing the usefulness of the index. *Cis* and *trans* isomers can be readily found under U& and U-, respectively; the inclusion of & and - would only duplicate these entries. The symbol T, when preceded by a number, usually indicates a saturated ring and gives no information which could not be found just as easily under other entries. A search for an unknown alkyl group, represented by A, would not be made. The elimination of the symbols X (a carbon atom bonded to four nonhydrogen atoms), Y (a carbon atom bonded to three nonhydrogen atoms), and numerals (denoting the number of catenary carbon atoms, ring size, or multipliers) is warranted on the basis of high frequency of occurrence and little indexing value (see Table II).

Following our exclusion rules, lines 2, 4, 5, and 6 (Fig. 2) would be discarded to leave but three entries (lines 1, 3, and 7). This compound could then be found in a search for: (1) heterocyclic six-membered rings (T6), or more specifically, pyridine derivatives (T6NJ); (2) compounds containing nitrogen (N); or (3) compounds containing chlorine (G). In each case one would look up the appropriate identifying symbol(s) in the tabulated list to locate compounds containing a specific ring, functional group, or atom.

Application of these rules finally reduced our original deck to 806 cards, *i.e.*, 6–7 entries per compound in the tabulated list. Since more than 90% of the line notations for compounds submitted to the Industrial Liaison Office occupy fewer than 20 columns on an IBM card, it is estimated that the average number of entries per structure will be not greater than 10. This means that a file of 50,000 compounds would create an index between 400,000 and 500,000 printed lines. Figuring 60 lines to a page, this would result in a directory of approximately 7500 pages.

It has been estimated that with an IBM 1401, the permutation would require about four hours, at a cost of approximately $200. To sort the permutations, an IBM

Table I
Symbols Not Indexed

Symbol	Meaning	Example	Notation
A	Generic alkyl		AR
J	Indicates ring closure		T6NJ
X	Carbon atom attached to four atoms other than hydrogen.	$\begin{matrix}Cl\\NH_2\end{matrix}>C<\begin{matrix}CH_2CH_3\\CH_3\end{matrix}$	ZXG2
Y	Carbon atom attached to three atoms other than hydrogen or doubly bonded oxygen.	$\begin{array}{c}CH_3\\\vert\\CH_3CH\\\vert\\OH\end{array}$	QY
&	Punctuation mark showing end of side chain; or following U, indicates *cis* or *syn* configuration; or preceded by a space, sign of molecular salt or addition compound.	$\begin{array}{c}Cl\\\vert\\HOCH_2CH_2CH_2C-CH_2CH_2CH_3\\\vert\\CH_2\\\vert\\CH_3\end{array}$	Q3XG3&2
		$\begin{matrix}CH_3\\H\end{matrix}>C=C<\begin{matrix}CH_3\\H\end{matrix}$	2U&2
		$CH_3CH_2NH_2\cdot HCl$	Z2 &GH
(Hyphen)	Separator and connective. Following U, indicates *trans* or *anti* configuration.		QR-G 5
		$\begin{matrix}CH_3\\H\end{matrix}>C=C<\begin{matrix}H\\CH_3\end{matrix}$	2U-2
T	When immediately preceded by a number, denotes saturated ring.		T6TMJ
Numerals	*Preceded by a space* are multipliers of preceding notations; or within ring signs L. . . . J or T. . . . J show the number of multicyclic points in the ring structure.		QR-G 5
	Not preceded by a space show ring sizes if within ring signs; elsewhere, numerals show the length of internally saturated, unbranched alkyl chains and segments.		L666 B6 P 2ABJ
			T6NJ
		$CH_3(CH_2)_3NHCH_2CH_3$	4M2

7074 would take about five hours ($1750). The tabulated list could be printed on a 600 lines/minute printer, with 60 lines per page, in 12.5 hours ($625). Therefore, the total cost, without estimating the cost for writing the program, comes to approximately $2600 for a file of 50,000 structures. Assuming a maximum of ten entries per compound on the tabulated list, the average cost would be approximately five cents per compound. The average number of entries may prove to be closer to seven with a concomitant reduction of cost.

Column 44

1									L	S	A	J		E		F	Q		O	Q									
2												L	S	T		F	V		M	U		O	V	J		A		E
3												L	S	A		F	V	J		E		O	Q					
4	L	S	T		M	U		O	V	J		A		E		F	V	1										

"ST" for E5T B6T6T6T
"SA" for E5T B6T6T6 } See ref. 6.

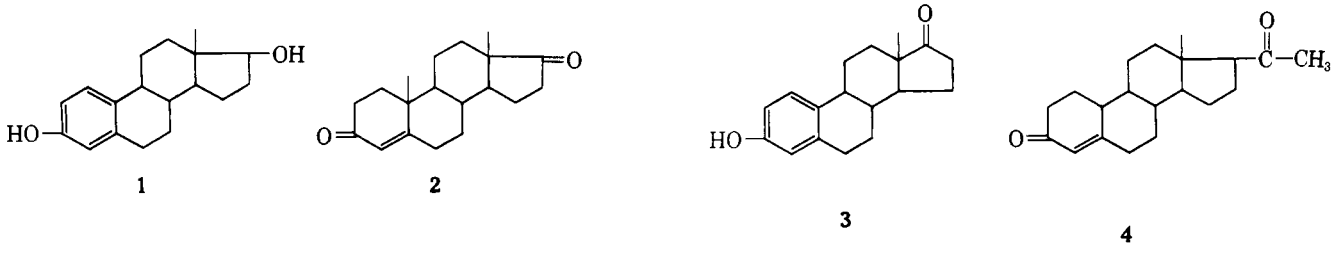


Figure 3.

Table II
Number of Cards Eliminated for Excluded Symbols[a]

Criteria symbols in column 43	44	Number of cards eliminated[b]
...	Space	*ca.* 2350
Space	Letter	*ca.* 250
...	X	*ca.* 80
...	Y	*ca.* 70
...	J	*ca.* 150
Numeral	T	*ca.* 50
...	Numeral	*ca.* 300
...	A	4
...	&	*ca.* 100
...	- (hyphen)	*ca.* 50
		3404

[a] Initially 4200 cards; finally 806 cards (6.7 cards per compound).
[b] Values calculated by measuring thickness of deck and using 140 cards/in.

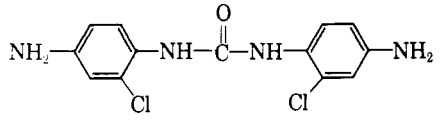
Updating procedures would involve the meshing of two tapes, the one used to produce the initial index and one containing notations for all newly added compounds. The resulting third tape would be used for preparing a new index. Should the cost of such a procedure become prohibitive, an alternative would be the use of supplements to the original index.

Utilization of the list of permuted notations would be facilitated by division, on the basis of symbology, into easily handled sections. Searches for specific compounds or general classes could be performed in the same manner in which a telephone directory is used. For example, to locate all aldehydes, one would select the V section of the index ($V = -\overset{|}{C}=O$) and read down to VH ($VH = H-\overset{|}{C}=O$). All aldehydes would be found grouped together (see Fig. 4).

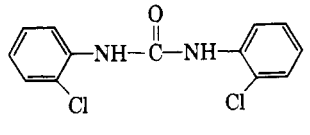
In this same V section, all acids could be located under VQ and QV (*e.g.*, compound **10**). All esters would be found under OV and VO (*e.g.*, compound **9**), etc.

This study has also revealed that multipliers should be used wherever possible. A multiplier (a number preceded by a space) indicates that certain preceding symbols in the notation are repeated. In effect, it contracts a notation. According to present usage multipliers should only be employed if they result in a saving of four symbols or more in the notation (see Table III). In example 1, the use of a multiplier saves four symbols and should be contracted according to the present rule. In example 2, only two symbols would be conserved and, hence, no multiplier is used. However, use of the multiplier in the latter case would result in a saving of three lines of print (three entries in a permuted index). Any modification of coding rules which will reduce the number of entries per compound should certainly be considered. It is here suggested that multipliers be used as frequently as possible when writing notations.

Table III
Uses of Multipliers



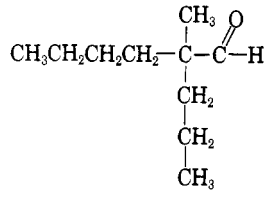
1



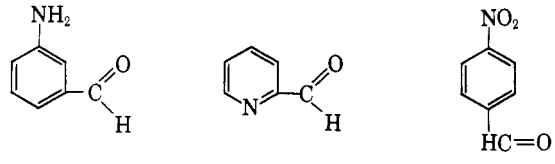
2

	Notation without contraction	No. of symbols	With contraction	No. of symbols
(1)	ZR CG DMVMRG DZ	12	ZR CG DM 2V	8
(2)	GRMVMRG	7	GRM 2V	5

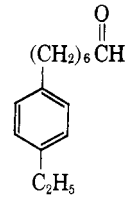
Compound No.							Column 44							
1			Z	R		C	V	H						
2	T	6	N	J		B	V	H						
3		W	N	R		D	V	H						
4							V	H	R	G				
5							V	H	R	2	N	1	&	1
6							V	H	X	4	&	3		
7							V	H	6	R		D	2	
8							V	H	7	V	2	M	2	G
9						2	V	0	1					
10						Q	V	1						



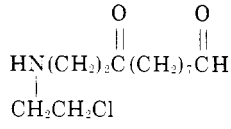
6



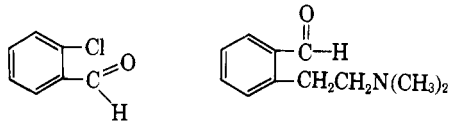
1 2 3



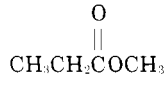
7



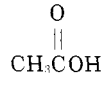
8



4 5



9



10

Fig. 4.—Part of V section of permuted index.

SUMMARY

(1) This study of permuted line notations has shown that tabulated lists can be used to rapidly locate: (a) specific compounds; (b) classes of compounds having similar ring systems, including benzenoids; (c) all compounds having the same functional group.

(2) The need for a functional group field for a chemical structure retrieval program can be eliminated.

(3) This approach appears to be economically feasible for medium (and possibly large) files of chemical structures.

(4) Multipliers not only save punch-card space but also considerably reduce the number of tabulated entries of redundant symbols.

ACKNOWLEDGMENT

The authors are indebted to the Data Processing Systems Division at the National Bureau of Standards for the suggestion that the permutation study should first be carried out with punch cards; Dr. Howard Bonnett for his helpful suggestions; the Data Processing Center at Edgewood Arsenal for card reproduction and tabulated lists; and Mrs. R. Bosely for the preparation of this manuscript.

REFERENCES

(1) A. Gelberg, W. Nelson, G. S. Yee, and E. A. Metcalf, *J. Chem. Doc.*, **2**, 7 (1962).
(2) E. G. Smith, *Science*, **131**, 142 (1960).
(3) See *e.g.*, *Chemical Titles*.
(4) A. Crestadoro, "Art of Making Catalogs of Libraries," London, 1856 (Science, Government, and Information, A Report of the President's Science Advisory Committee, The White House, Jan. 10, 1963, p. 20).
(5) "Current Research and Development in Scientific Documentation, No. 11," National Science Foundation, November, 1962.
(6) H. T. Bonnett and D. W. Calhoun, *J. Chem. Doc.*, **2**, 2 (1962).