# Rapid Structure Searches *via* Permuted Chemical Line Notations. II. A Key-Punch Procedure for the Generation of an Index for a Small File

CHARLES E. GRANITO, ALAN GELBERG,
JOHN E. SCHULTZ, GERALD W. GIBSON, and EDWARD A. METCALF
Industrial Liaison Office, Office of the Technical Director, U. S. Army Chemical Research
and Development Laboratories, Edgewood Arsenal, Maryland

Received June 26, 1964

In a previous communication,[1] the feasibility of using an index of permuted Wiswesser chemical line notations[2] for rapid structure searches was demonstrated. Only simple punch-card equipment was used in that study of a small number (120) of compounds which could be represented by line notations of less than thirty-five symbols.

If a large number of structures is involved, as in our file of about 50,000 structures, a computer is required to prepare the permutations and the index thereof. To this end, the Data Processing Division, Edgewood Arsenal, Maryland, has written a program for an Univac File Computer (Model II).[3]

Updating the computer-prepared list will, of necessity, be a semiannual or annual requirement depending on the number of new structures added to the files, the necessity for currency, and the availability of computer time. The purpose of this report is twofold: (1) to describe an inexpensive, rapid method for maintaining searchable intermediary records between up-dating periods; and (2) to describe a procedure, requiring only simple punch-card equipment, for obtaining and maintaining a permuted index of small holdings of chemical structures, *i.e.*, up to approximately 5000.

The procedures, hereinafter described, provide a method for the preparation of the various permutations of a chemical line notation by a key-punch operator. The latter works from an especially prepared program sheet on which the "encoder" has written the notation only once. This is an important factor in determining the cost of a system wherein salaries and levels of training must be considered. A chemist trained to prepare notations can prepare the input at an average rate of one structure every 30 sec.; about 500 notations can be prepared in one working day before the coding becomes tedious. As each notation requires an average of six entries, it is estimated that about 12 hr. of key-punching time and 12 hr. of verification time will be required for the 3000 entries (500 structures). The key-punch operator, once adjusted to this type of a program, can punch about 250 cards in an hour. Using a 650 card per minute sorter, the 3000 entries can be alphanumerically arranged, assuming a twenty-column (columns 70 through 51), double sort (for rows 0–9 and then the zones), in about 4 hr.

The above operations produce a deck of cards arranged in alphanumeric order. The sorted cards are stored in this sequence so that new entries can be blended in by hand or machine. The meshing of new cards into the master deck can be done with an 082 single column sorter. However, a collator is more desirable since it can accomplish this blending operation in a fraction of the time.

In this program, an IBM 026 printing key punch is used to prepare the input cards. This permits the preparation of a list of the entries, for immediate look-up, in any number of ways depending on the ancillary equipment available: (a) direct tabulation with an IBM tabulator or an off-line printer; (b) typewriter output from the IBM document writer; (c) overlapping the cards, as "shingles" with the upper edge containing the print-out exposed, and Xeroxing the cards with the Xerox 914 office copier.

Wiswesser[4] has indicated that approximately 60% of the notations of a random selection of compounds will require no more than ten symbols. In the files of this office, 80–90% of the notations require fewer than twenty symbols. Therefore, most notations can be permuted by the first and simpler of the two procedures described below; a more complicated procedure is required for notations containing from thirty to sixty symbols. A specially designed program sheet (Figure 1) is used by the encoder to provide the key-punch operator with the necessary information. The encoder uses the following format:

(a) *Columns 1 through 6* are used for the accession number or other "address" of the compound. The last digit of the accession number is placed in column 6, the next to last in column 5, etc. An alphanumeric address may be used if desirable. For files containing less then 10,000 structures (only columns 3, 4, 5, and 6 are required for the accession number), columns 1 and 2 may contain a two-letter prefix for the identification of various decks.

(b) *Column 7* is not used.

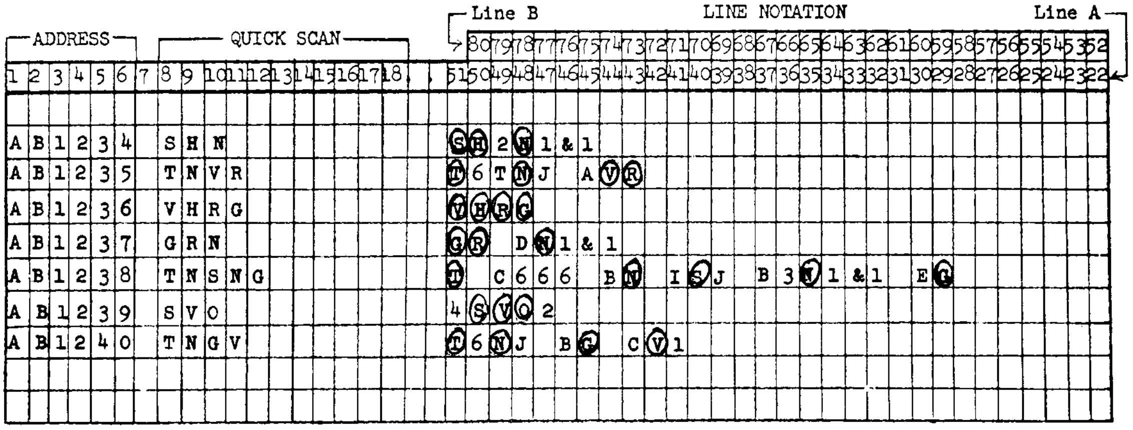
(c) *Columns 8 through 18* contain (beginning in column 8) the symbols, in order of the appearance in the line notation, that will be indexed and are of importance to the searcher. This QUICK-SCAN area provides a "screen" for the user of the index. Consider, for example, the search for all

[1] P. F. Sorter, C. E. Granito, J. C. Gilmer, A. Gelberg, and E. A. Metcalf, *J. Chem. Doc.*, 4, 56 (1964).

[2] Some notations, used as examples in this paper, may not be consistent with the revision of Wiswesser line notation rules currently being prepared for publication by Dr. E. G. Smith, Mills College, Oakland, Calif.

[3] The subject of a future publication.

[4] W. J. Wiswesser, private communication.


ADDRESS
QUICK SCAN
Line B
LINE NOTATION
Line A
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
51 50 49 48 47 46 45 44 43 42 41 40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22
80 79 78 77 76 75 74 73 72 71 70 69 68 67 66 65 64 63 62 61 60 59 58 57 56 55 54 53 52

| ADDRESS | QUICK SCAN | LINE NOTATION |
|---|---|---|
| AB1234 | SHN | SH2N1&1 |
| AB1235 | TNVR | T6TNJ AVR |
| AB1236 | VHRG | VHRG |
| AB1237 | GRN | GR DN1&1 |
| AB1238 | TNSNG | T C666 BN ISJ B3N1&1 EG |
| AB1239 | SVO | 4SVO2 |
| AB1240 | TNGV | T6NJ BG CV1 |

Figure 1.—Permuted chemical line notation program sheet.

chlorinated pyridines. All pyridines could be located immediately by turning to the "T6NJ" section. However, to find those which contain chlorine, the searcher (if the QUICK-SCAN area were omitted) would be required to scan all the notations in that section for those containing a "G" (chlorine). Inasmuch as the number of symbols on both sides of the indexed symbol could approach thirty, the scanning of these numerous symbols would be much more laborious than looking at the eleven-column field of the QUICK-SCAN area. For example, compare the time required to pick out the symbol "F" (fluorine) in the QUICK-SCAN area and in the line notations given in Figure 3.

For the QUICK-SCAN area only it has been shown to be advantageous to include the "A" symbol (contrary to a previously published[1] exclusion list). Including this symbol for the QUICK-SCAN area assists in the rapid pinpointing of sodium (-NA-), calcium (-CA-), etc.

(d) *Columns 19 and 20* are not used.

(e) *Columns 21 through 80* of the punch card are used for the chemical line notation. However, on the program sheet, the encoder places the first symbol in column 51 (line A of column numbers, Figure 1), the second in 50, etc. If more than 30 columns are required for the notation, fill the spaces 51 through 22 on the program sheet (as in line 1, Figure 3) and continue on the next line again starting in column 51 (as in line 2, Figure 3). After writing the notation, the first and all other pertinent symbols are encircled. With the exception of numerals which initiate notations, the pertinent symbols are rewritten in columns 8 through 18, the QUICK-SCAN area.

The remainder of the process is performed by the key-punch operator. All symbols appearing in columns 1 through 20 on the program sheet appear in the same columns of all of the punch cards concerned with a single structure or address. However, the notation is punched in a different series of columns on each card in order that each encircled symbol will appear once in column 51 (the index column). The key-punch operator prepares the first card (card 1, Figure 2) as it appears on the program sheet

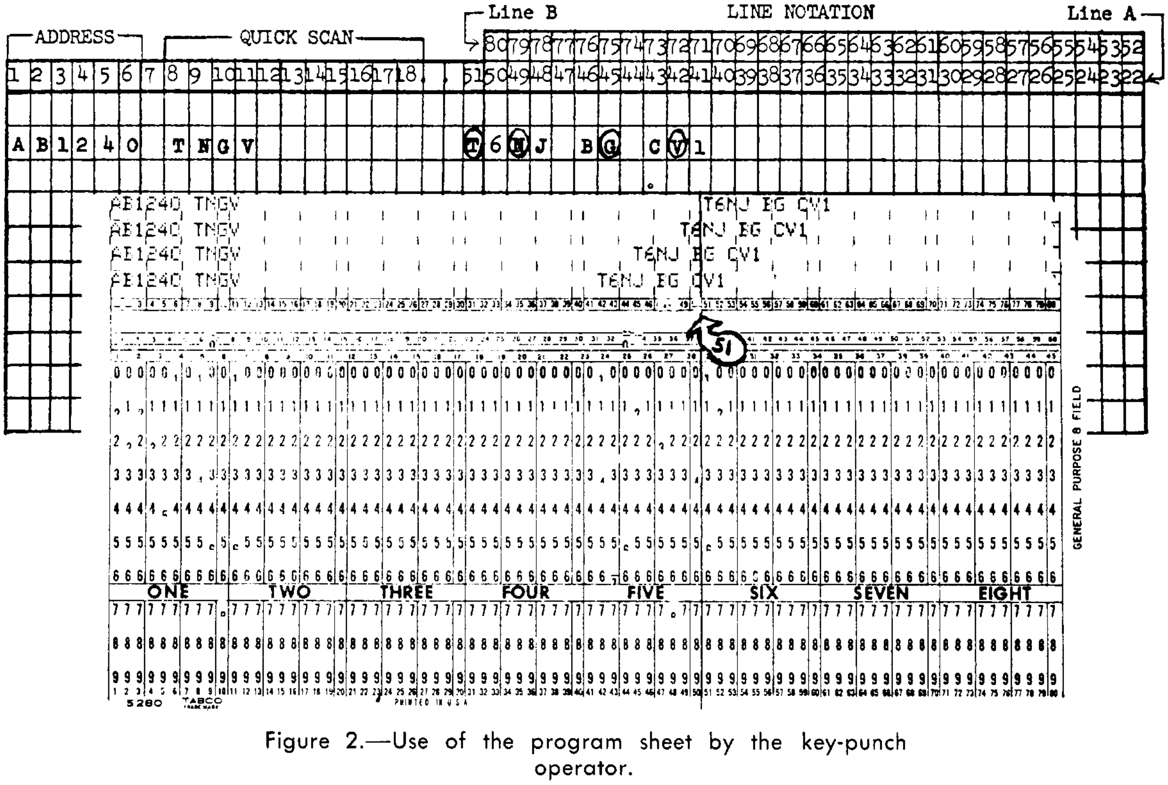


Figure 2.—Use of the program sheet by the key-punch operator.

(Figure 2). The first symbol of the notation is placed in column 51 of the card, the second in column 52, etc., paying no attention, at this time, to the numbers of the columns on the program sheet. For the second card, and all subsequent cards, the key-punch operator punches the address and the QUICK-SCAN area in the same columns as indicated for card 1. This is easily accomplished by duplicating the information in these two fields. For the second card, the number of the column in which the notation should be started is found by reading the number (line A, Figure 2) of the column in which the second, encircled symbol appears. For AB1240 (Figure 2) the notation for the second card would start in column 49, in column 45 for the third card, and in column 42 for the fourth. To avoid duplication or skipping a symbol, it is desirable to have the first encircled symbol check marked (√) on the program sheet after the first card has been prepared, the second encircled symbol check marked after the second card, etc.

The foregoing procedure is capable of handling chemical structure notations occupying thirty columns, or less, of an IBM card. This will include most organic structures. However, a modified procedure is necessary for the permutation of notations occupying from thirty-one through sixty columns. An example of a notation requiring forty-one columns is shown in Figure 3. The encoder, after filling the spaces in columns 51 through 22 on the program sheet, starts the "extension" (line 2, Figure 3) of the notation in column 51 and continues as needed.

The key-punch operator follows the same directions as above for punching the address and the QUICK-SCAN areas of the cards. A different procedure must be used, however, for positioning the notations. To position correctly the notation on the first card, the notation "extension" (line 2, Figure 3) is started in column 21 with the "6", the T is placed in 22, and subsequent symbols and spaces follow in columns 23 through 31; *i.e.*, line 2 is completed. The notation in line 1 is started in column 51 and continued through column 80, following the same order of symbols and spaces as shown on the program sheet.

To position the notation on the second card, it must be placed so that the second encircled symbol (N) of line 1 falls in column 51. This is accomplished by:

(a) Reading the symbol in line 2 that appears immediately below the "N", *i.e.*, a "2"

(b) Punching the "2" in column 21 and the remainder of line 2 in columns 22 and 23

(c) Reading the column number immediately above the "N" in line A, *i.e.*, 43

(d) Punching the notation (line 1) starting with the first symbol, "T", in 43, a space in 44, "C" in 45, etc.

(e) After the symbols "DT" (at the end of line 1), continue punching the symbols in line 2 until the symbol "2" is reached (*i.e.*, through symbol "B"). The "2" and subsequent symbols in line 2 nave been punched previously.

The notation on the third card begins in column 40 (as noted by reading the column number, line A, directly above the third encircled symbol, "S") with the symbol T. The symbols and spaces in lines 1 and 2 are all punched in the order shown in lines 1 and 2 of the program sheet; the symbols in line 2 follow those in line 1. In this instance, no punching is required in columns 21, 22, etc., because there are no symbols in the space immediately below the third encircled symbol, "S", or following that position on line 2. Cards four, five, and six are prepared similarly.

To accomplish the correct positioning of the notation in order that encircled symbols in line 2 will fall in column 51, a slightly different procedure must be used. The notation for the seventh card ("N" is the seventh encircled symbol) is positioned as:

(a) Read the symbol in line 1 that appears immediately above the "N" in line 2, *i.e.*, a "C"

(b) Starting in column 21 with the symbol "C" continue

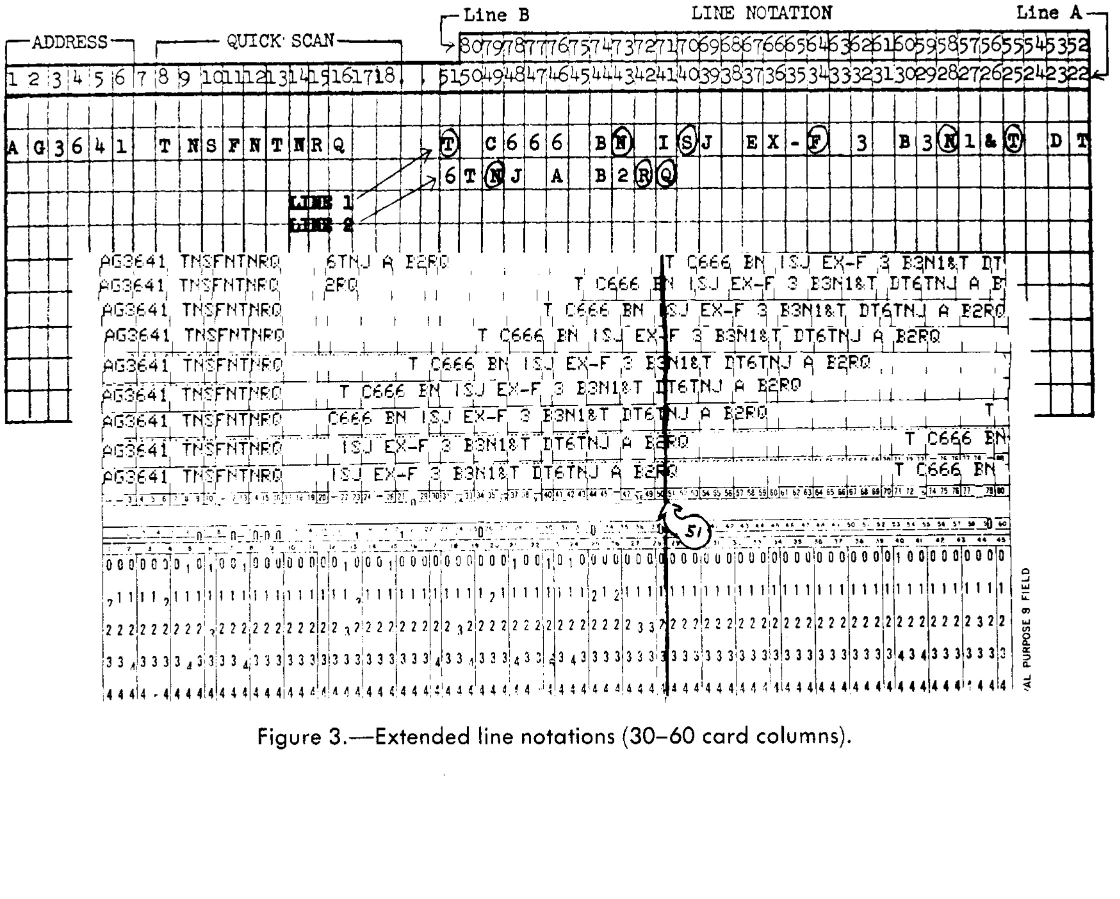


Figure 3.—Extended line notations (30–60 card columns).

punching in 22, 23, etc., the symbols and spaces following the "C" in line 1 and continue to the end of line 2

(c) Read the column number, line B, in which the "N" appears, *i.e.*, 79

(d) Starting in column 79, punch the remaining symbols of the notation that appear before "C", *i.e.*, a "T"

The next card requires that "R" in the notation be positioned in column 51. Following the above procedure:

(a) Column 21 is blank (space), and "I" in 22, etc., continuing with the remainder of line 1 and line 2

(b) The first part of the notation, that which occurs before the "space" and "I" in line 1, *i.e.*, T C666 BN, starts in column 72 as indicated by the number (line B) of the column containing the "R"

In this manner, permutations of notations requiring up to and including 60 columns can be indexed.

Although the procedure appears quite complex, it has been found that a key-punch operator can learn the mechanics of this operation in a very short time. Table I outlines the specific steps for the key-punch operator.

Table I
Permuting Instructions for the Key-Punch Operator

1. For any given number (structure), columns 1 through 18 are punched identically. After preparation of the first card, use duplicating key for this portion of card.
2. One-Line Notations
   a. Locate first encircled symbol
   b. Read number of column (line A) directly above symbol
   c. Start punching the notation in the card column having this number and complete notation in consecutive columns
   d. Check-mark symbol on program sheet
   e. Read number above second encircled symbol and repeat steps c and d on a second card
   f. Continue process until one card has been generated for each encircled symbol
3. Two-Line Notations
   a. For encircled symbols on line 1
      (1) Starting in column 21, punch the remainder of the information on line 2 starting with the symbol directly beneath the encircled symbol.
      (2) Read number of column (line A) in which encircled symbol appears and SKIP to this column of card. In that column, start punching notation from INITIAL symbol (line 1). Continue punching line 1 and then line 2 up to the symbol previously punched, *i.e.*, the one placed in column 21.
      (3) STOP. Check-mark encircled symbol on program sheet.
      (4) Prepare next card in the same manner working with the next consecutive, encircled symbol on line 1. Continue until all encircled symbols on line 1 have been operated on in this fashion.
   b. For encircled symbols on line 2
      (1) Starting in column 21, punch the information in line 1 beginning with symbol directly above encircled symbol.
      (2) At end of line 1, continue punching to the end of line 2.
      (3) SKIP to the column number, indicated in line B for the encircled symbol, and punch first portion of notation (line 1) up to the symbol immediately above encircled symbol.
      (4) STOP. Check-mark encircled symbol on program sheet.
      (5) Prepare next card in the same manner working with the next consecutive encircled symbol on line 2. Continue until all encircled symbols on line 2 have been operated on in this fashion.

## SUMMARY

An economical method is described for creating permuted listings of chemical line notations using simple punch-card equipment. The encoder writes the line notation for each structure only once on a program sheet. The key-punch operator generates the permutations by positioning the notations so that each pertinent symbol will fall in a predetermined column.

This methodology provides a means of creating a permuted index for small files (up to 5000 structures) without recourse to a computer. It may also be used as an interim procedure for maintaining searchability of structures acquired since the preparation of a computer-prepared index.

The suggested methodology may be modified to fit the needs of the users of other types of indexes: key words, molecular formulas, etc.

**Acknowledgment.**—The authors are indebted to Mrs. R. Bosely for the preparation of the manuscript.