A Modified Dot-Bond Structural Formula Font with Improved Stereochemical Notation Abilities*

RONALD GOTTARDI
Smith Kline & French Laboratories, Philadelphia, Pa. 19101

Received September 16, 1969

A computerized chemical information retrieval system is described, and details are given on a new set of graphic characters, called "octobliques," for input of structural diagrams using an automatic recording typewriter and for their output using a standard impact-type, high-speed computer printer. The new font, an extension of the existing dot-bond notation, is able to handle highly complex structures, including stereochemical forms ranging from simple bridged rings to the carborane icosahedron, that previous fonts were unable to accommodate. The "octobliques" and other symbols useful in nomenclature and in the preparation of laboratory reports have been placed on an interchangeable typing element for use in any Selectric typing mechanism. This element may be used to provide original documentation or to produce suitable computer input along with a typed copy of the structure. In addition, the same characters have been engineered for use on IBM high speed computer printers.

At Smith Kline & French Laboratories, we have had a computerized Chemical Information Retrieval System in operation for a number of years. The system consists of a file of fragment code descriptors of the molecular structures of approximately 40,000 compounds which SK&F has tested as potential new drug products over the years.[1] The fragment coding is based on a modified CBCC code and is done manually for all new compounds entering the file. Queries to the file are also manually coded, using these fragment descriptors. Boolean logic statements of these descriptors are matched serially against the file compound codes, and lists of compound identification numbers and compound names obtained from a secondary file are printed for each query. The compound identification numbers are then used to pull drawings of the structures, manually, on 3 × 5 cards; these are then sent to the questioner. To eliminate this manual card-pulling operation and thereby reduce the turnaround time on queries, we decided to implement a structural input/output capability.

Our first step was to get the structural information into digital form for storage in the computer; currently a chemical typewriter is most often used to accomplish this. Early development of chemical typewriters was done at American Cyanamid's Stamford Research Laboratories.[2] Feldman *et al.*[3] at Walter Reed Army Institute of Research added a recording medium to produce machine readable records of structural formulas. They used an automatic paper tape recording typewriter (later modified by the Mergenthaler Linotype Company to include a three shift keyboard) with an absolute coordinate generating device. This typewriter produced structures of the conventional closed ring or gapless notation. The triple shift keyboard was required to accomodate the many ring juncture characters needed to produce closed rings.

Mullen,[4] at Shell Development Co., designed a character set built around H. P. Luhn's suggestion of replacing the numerous ring juncture characters with a "fat dot." This open ring font, usually called the dot-bond notation, used nine basic graphics (Figure 1,*a*) to produce most structural formulas. Shell Development Co. built the prototype and engineered a "reverse index" feature into the device. This feature provides the keyboard control and functional code for the one missing direction (plus Y) needed to furnish a complete relative coordinate system instead of absolute coordinates, as in the Army-Mergenthaler typewriter. Depressing the reverse index key retracts the paper vertically or, in effect, moves the typing element up relative to the paper and records a code for this function. It is necessary to record all typing element movement (relative to a starting position) in order to reconstruct the structure upon playback of the recording medium for verification or in computer memory for analysis. Reverse indexing allows the typist to type structures with a completely random motion, following any desired character path—e.g., going around a ring as it is typed. Appendix 1 shows a code sequence as it would be generated on the recording medium during random motion typing of a structure. Display on a computer printer requires conversion from this random motion to a line-at-a-time configuration through a suitable algorithm.

Maxwell Gordon,[5] at SK&F, developed a new closed ring font that did not require ring juncture characters, and a Friden automatic typewriter was modified to include this font and the reverse index feature. Communications between Mullen and Gordon resulted in the addition of two elongated diagonals with a slope of 2.0 (Figure 1,*b*)

*Presented in part before the Division of Chemical Literature, 157th Meeting, ACS, Minneapolis, Minn., April 1969.

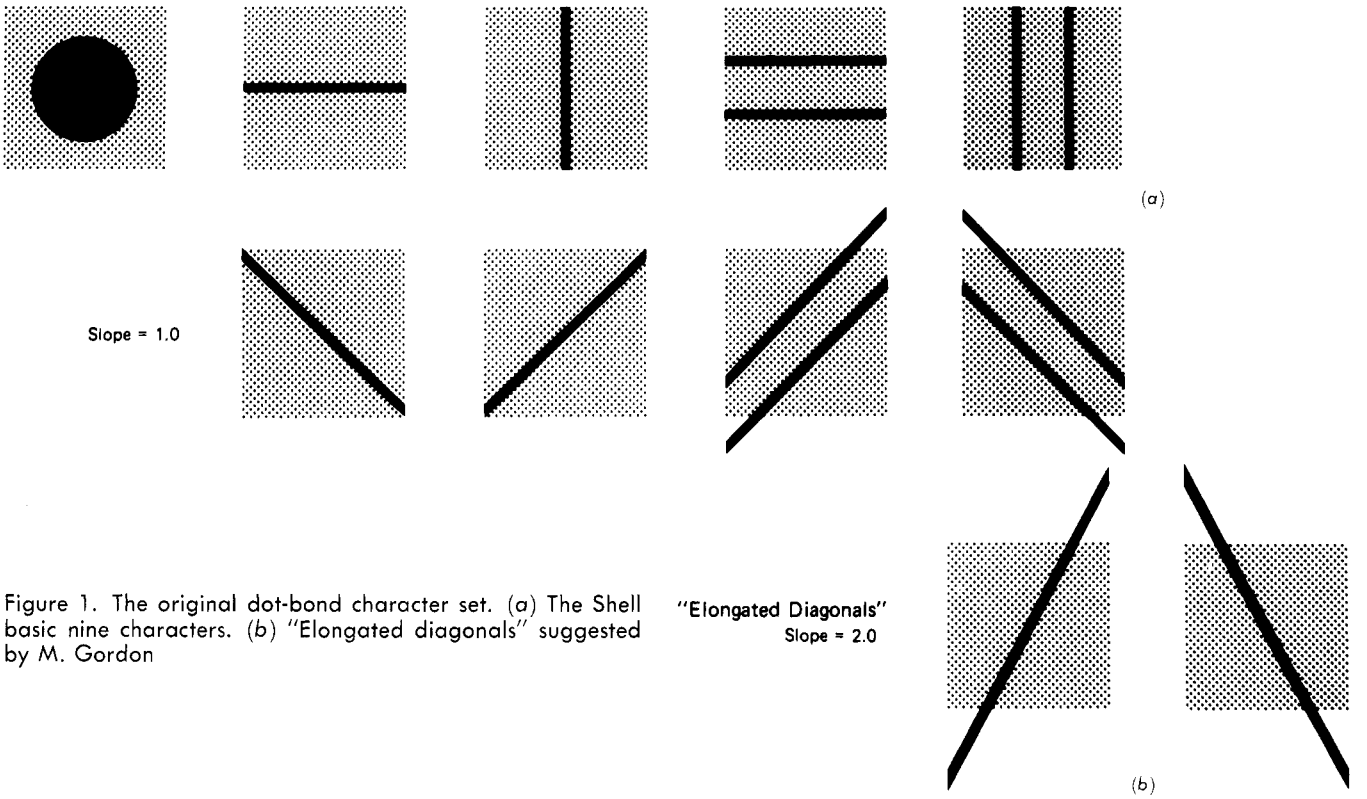


Figure 1. The original dot-bond character set. (*a*) The Shell basic nine characters. (*b*) "Elongated diagonals" suggested by M. Gordon

to the Shell set; these were suggested by Gordon for typing stereo configurations. The Shell font was implemented on a Selectric interchangeable typing element and has also been implemented with and without the elongated diagonals on several high speed computer printers, an early example of this being the work of Rice at Eli Lilly & Co.[6]

The Gordon closed-ring font had the advantage of producing structures of publishable quality in the conventional closed ring or gapless form, but the size of the characters employed prevented implementation on a Selectric typing sphere, and also presented difficulties when used on high speed computer printers of the impact type. Experimental work on this font is continuing on a photocomposition printer, where character size is not as strong a limiting factor.

To produce structures on our impact printer, we decided to use the dot-bond notation, but were somewhat discouraged by the Shell font's limited stereochemical abilities and the slow print speeds caused by the large number of characters in the sets used. In addition, the elongated diagonals used on the Shell sphere (shown in Figure 2,*a* on the typewriter) could not be engraved to the full height of two vertical line spaces required to have these

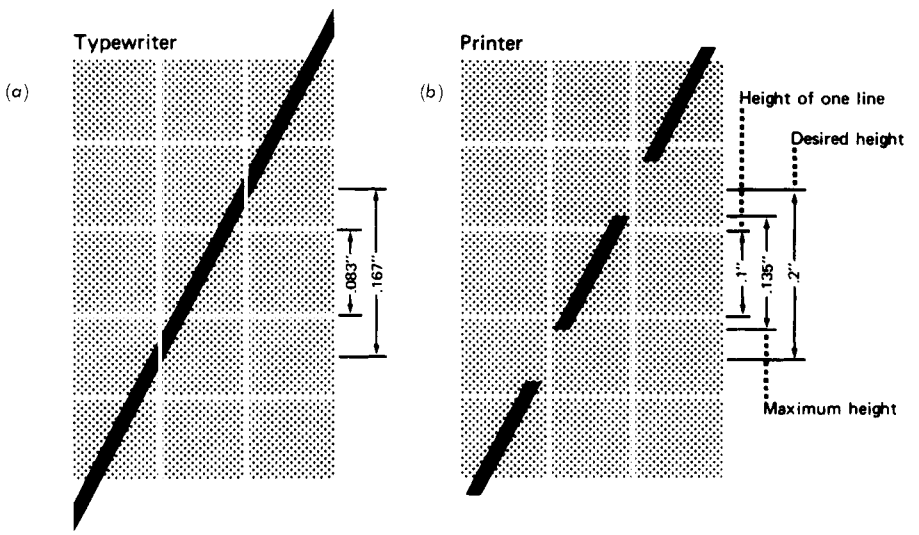


Figure 2. A 2.0 slope line produced by using the "elongated diagonals" of the original dot-bond character set on (*a*) the typewriter and (*b*) on the computer printer.

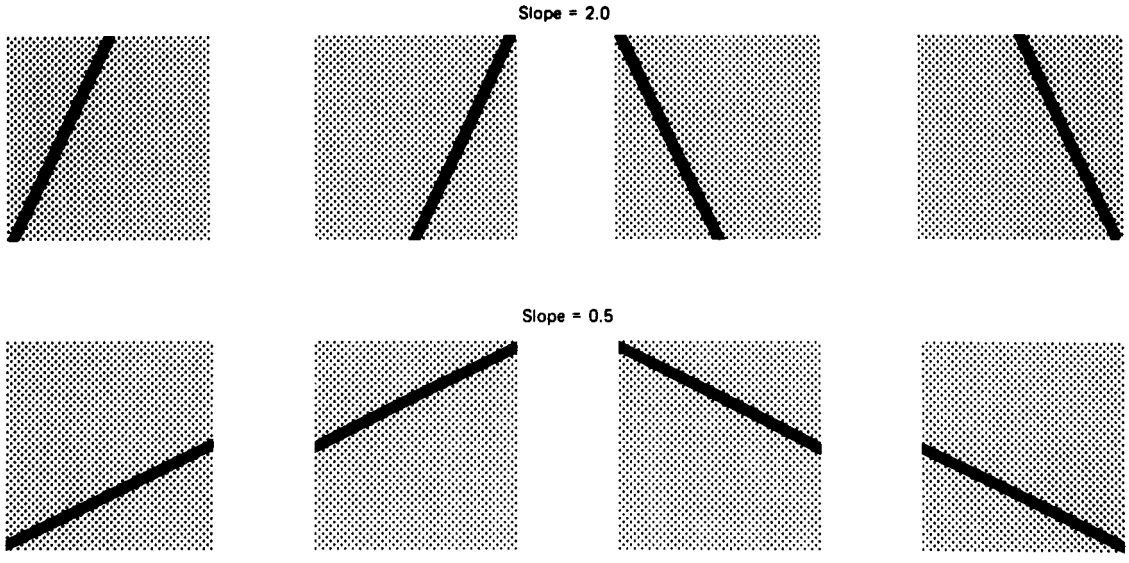


Figure 3. The "octobliques," characters added to the Shell dot-bond font to extend its graphic capabilities

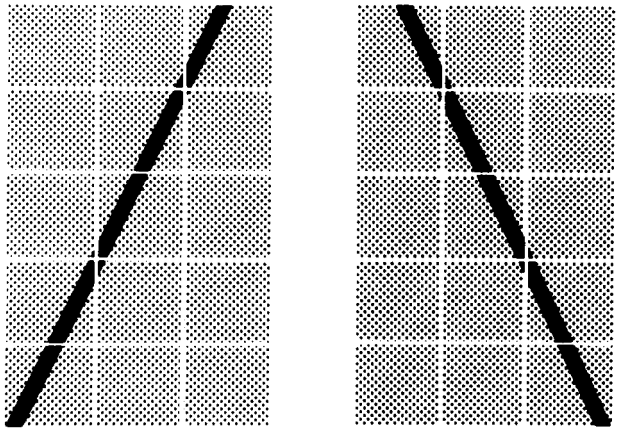


Figure 4. Continuous lines of slope ±2.0 produced using the "octobliques"

strokes meet and present a continuous straight line on the computer printer. The best that could be done is shown in Figure 2,*b*. Because of these limitations, and the growing stereochemical problem in our files, we modified the basic Shell dot-bond character set to include a set of eight additional oblique lines (Figure 3) in place of the two elongated 63° diagonals. These eight slants have been called "octobliques."

The set consists of two slants of slope +2.0, two slants of −2.0, and two each of slope plus and minus 0.5. Each pair of like slopes is situated in opposite halves of the character matrix.

ADVANTAGES OF THE OCTOBLIQUES

1. They are adaptable to almost all printing devices, because no slants exceed the printing "window" or matrix on any side; this feature eliminates the problem with the elongated diagonals on impact printers.

2. They complete the set of eight secondary vectors required to bring the Shell font to the next higher level of graphic capability, thus allowing an impact printer to approach the vector capabilities of a plotter. These vectors permit the generation of lines of slope ±2.0, as shown in Figure 4, and ±0.5 as shown in Figure 5.

3. Their position within the print matrix and their relationship to the fat dot permit the simulation of some additional vectors with slopes of intermediate values. In Figure 6, a 2.0 slope octoblique is used to connect dots separated by a center-to-center line of slope 1.5.

4. They reduce the amount of distortion of structural parts sometimes required to fit the structure to the available graphics. In Figure 7, the solid dot in the center of this 15 × 15 matrix can be joined easily and unambiguously to any open dot in the matrix.

5. They reduce the enlargement of structures sometimes required to accommodate intracycle conformations, such as bridges (Figure 8) and rings with an odd number of sides. Figure 9 shows rings of three to eight sides. The three-, five-, and seven-sided rings could not have been done with the Shell font without considerable enlargement.

6. They eliminate some of the reorientation of structures or substructural parts sometimes required to

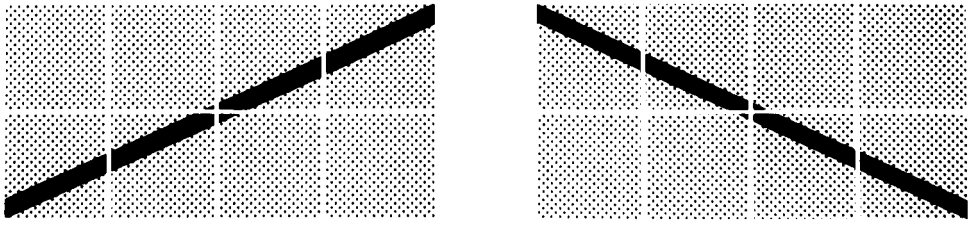


Figure 5. Continuous lines of slope ±0.5 produced using the "octobliques"

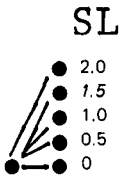


Figure 6. A single quadrant illustration of the dot joining ability of the SK&F dot-bond font. Here a 2.0 slope octoblique is used to simulate a line of 1.5 slope

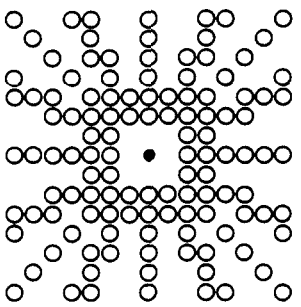


Figure 7. Dot joining ability of the SK&F dot-bond font means less distortion of structural parts. The solid dot in the center of this 15 × 15 matrix can be easily and unambiguously joined to any open dot in the matrix

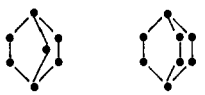


Figure 8. Ring bridges executed with the "octobliques"

fit the available character set—e.g., a ring with a horizontal bridge. With the Shell font, these rings must be rotated 90° before typing.

7. Because of their position within the print matrix and their intramatrix separation, they can be combined through overstriking to produce other useful graphics, such as double bonds at slopes of ±2.0 and ±0.5, and greater than and less than signs.

8. Their vector capabilities permit an approximation of curves and circles, useful in difficult conformations, in very large cycles (more than eight sides, for example), and in a convention where a circle with a subscripted group is used to represent rings of more than eight sides.

9. They permit the generation of arrows in 16 different directions and of any desired length, very useful characters in many aspects of chemical documentation—e.g., in structures such as metallic complexes, equations, diagrams, etc.

10. Finally, they can accomplish the listed graphics without extending beyond the print matrix, thus avoiding unaesthetic collisions or overlaps with other characters.

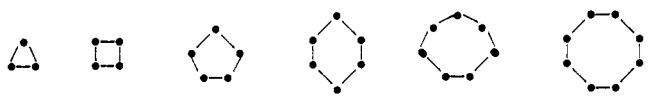


Figure 9. Rings of 3 to 8 sides using the SK&F dot-bond font

Figure 10 illustrates the use of the octobliques in some typical stereo configurations. These were typed with the stereochemical typing element in a standard Selectric typewriter.

KEYBOARD DEVELOPMENT

In the process of incorporating the octobliques into the typewriter keyboard, we were faced with the problem of finding key positions for them. We started with the Shell keyboard to introduce as few incompatabilities with other systems as we could. (For further information on compatability of chemical information systems, see Jacobus.[7]) Our decision to sacrifice part of the lower case alphabet in order to place the octobliques on the keyboard was based primarily on a prior decision not to include any lower case letters on the print train that would generate output from our system on the high speed computer printer. We felt that including the full lower case alphabet would depress the line speed of the printer by an unacceptable amount. However, we stopped short of eliminating the lower case alphabet entirely, because some letters are useful for the more frequently occurring element symbols in organic chemistry and for nomenclature.

The lower case characters retained are a, d, e, g, i, l, m, o, p, r, and u. This set of 11 will allow the conventional typing (a lower case second letter where needed) of 78 element symbols from the table of elements and all of the more commonly occurring elements in organic compounds. Some of the lower case letters serve double duty, being used for nomenclature as well as for element symbols. Lower case d, l, m, o, and p can be used as usual for dextro, levo, meta, ortho, and para. After we gave up a few of the lower case letters, it was easy to add other characters and symbols that are useful in chemistry.

Some of these additions were:

Additional characters for structures: a triple bond and an assortment of dotted bonds. The dotted bonds are useful in steroids.

Additional characters useful in nomenclature: Greek letters and some punctuation marks.

Additional characters for laboratory work: degree mark and per cent sign.

A sample of nomenclature and equation typing using the SK&F stereochemical typing element is shown in Figure 11.

The basic Shell element design was changed in other ways to improve the quality and reproducibility of typed structures. These changes include:

A smaller fat dot.

Slightly larger numbers with slight style changes to eliminate some ambiguities.

Smaller upper case letters and style changes in the alphabet to further eliminate ambiguities.

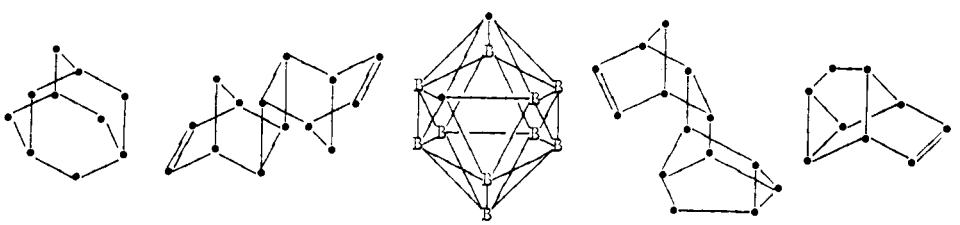


Figure 10. Some typical stereochemical configurations typed with the SK&F dot-bond font

$ENZYME\text{-}Cu^{2+} + AH_2 \longrightarrow AH\cdot + ENZYME\text{-}Cu^{+} + H^{+}$

$RCOOH + SOCl_2 \longrightarrow RCOCl + SO_2\uparrow + HCl\uparrow$

3-METHOXY-4-AZATRICYCLO [3.3.2.$0^{2,8}$]DECA-3,6,9-TRIENE

1-α AMINOISOVALERIC ACID

$CuSO_4$,AQ.PYRIDINE,95° 86%

Figure 11. Equation and nomenclature typing using the SK&F stereochemical typing element

Figure 12. Keyboard layout of the SK&F stereochemical typing element

The ambiguities were confusion between upper case I, lower case l, and the vertical single bond, and also between zero and alpha 0.

The final keyboard layout is shown in Figure 12. The spherical typing element (Figure 13) was custom-made for us by Camwil, Inc. (Part No. 678-M, Camwil, Inc., 835 Keeaumoku St., Honolulu, Hawaii 96814), current supplier of the Shell sphere, and has been added to Camwil's catalog of stock elements. The full stereochemical element character set is typed in Figure 14.

INPUT DEVICE

For the initial input device, we selected an automatic input typewriter functionally similar to the machine first built by Shell. Our units included punched paper tape and edge punched card input and output, reverse index, and BCD code with parity checking. We plan to convert to a magnetic tape automatic typewriter in the near future.

The automatic reading feature on the input device will help in the keyboarding of the 40,000 structures in our current files, a process that will take some time. Master records are prepared for the basic structural features, such as ring nuclei, that occur in many structures. The records are then read, typed, and recorded automatically; the operator has to add only the parts that vary from structure to structure within a group of similar ring nuclei.

COMPUTER PROCESSING OF STRUCTURES

Once the structural information has been keyed and recorded in paper tape on the input device, it is converted

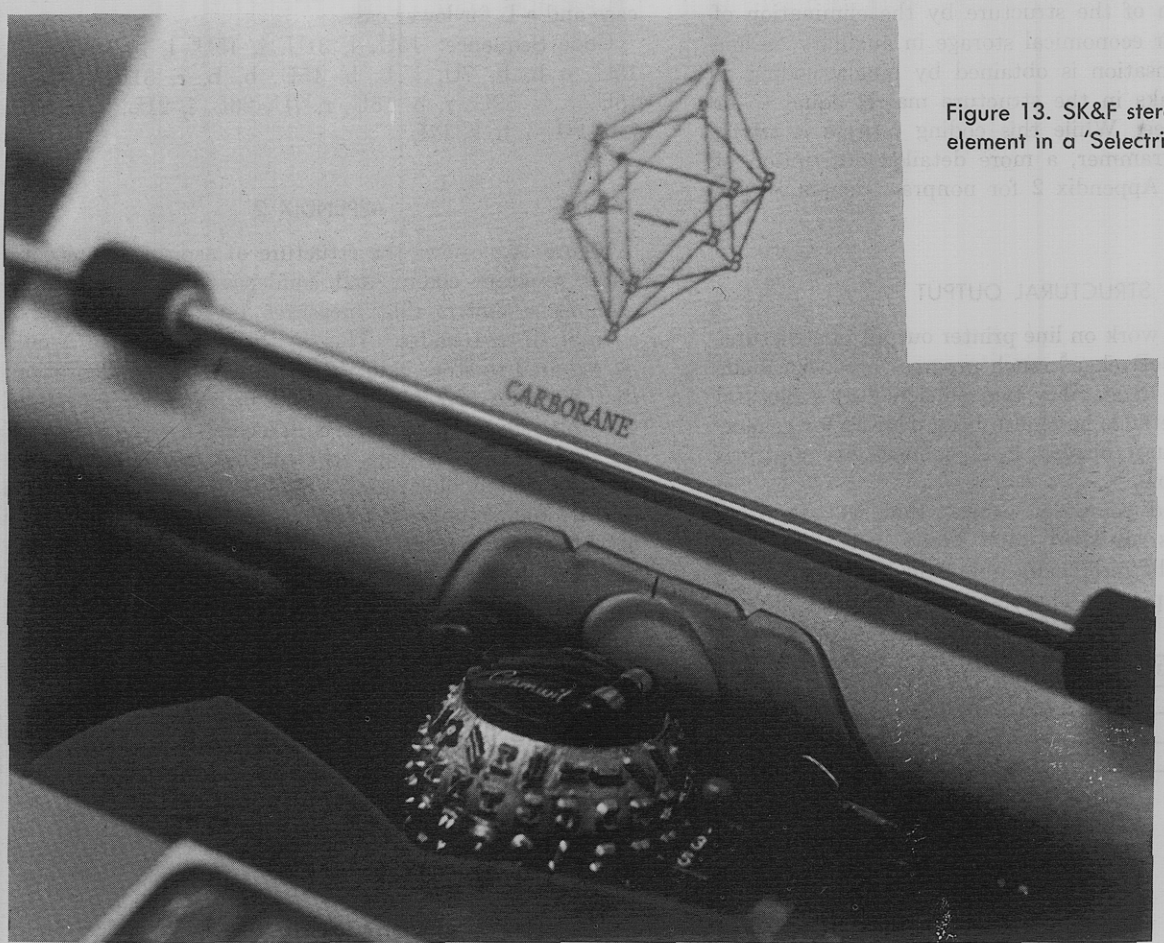


Figure 13. SK&F stereochemical typing element in a 'Selectric' Typewriter

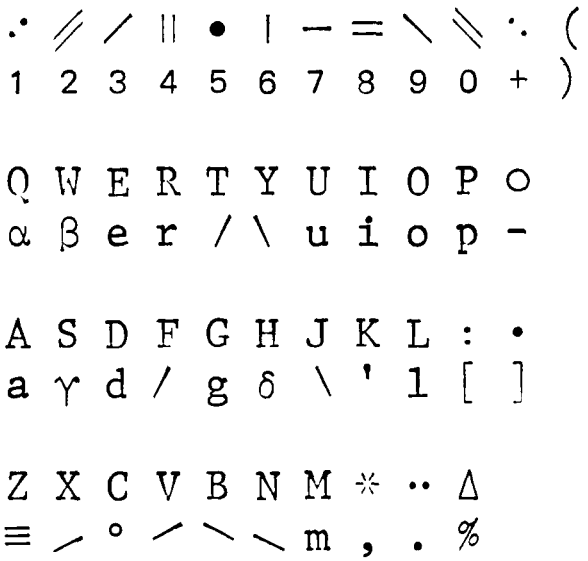


Figure 14. Type-off of SK&F stereochemical typing element

to magnetic tape for the IBM 360 computer. The 360 programs reconstruct the structure in computer memory in a four level array to accomplish three objectives: editing of the typed structure through element counts and bond analysis for derivation of implied hydrogen counts and comparison with the molecular formula; translation of the structure from the essentially random motion which the typist uses in keying the structure to a line-at-a-time construction required for output on the line printer; and condensation of the structure by the elimination of "white space" for economical storage in auxiliary on-line memory. Condensation is obtained by binary coding of all series of blanks in the structure matrix equal to or greater than three. While this coding process is trivial to a skilled programmer, a more detailed description of it is included in Appendix 2 for nonpractitioners of this modern art.

STRUCTURAL OUTPUT

Since the early work on line printer output of structures by Waldo and DeBacker,[8] much progress has been made in improving the readability and efficiency of structural output. Some methods have also been devised for generation of a structural display from connectivity matrices and linear notations.[9]

The high speed computer printer that we are using in the system is an IBM 1403 Model N1 printer on a System 360. The interchangeable train cartridge allows us to use special purpose character sets. For this application we added special features to the printer, such as wide hammers and 10 lines per inch vertical spacing. These features provide a square matrix and make the bonds meet to present a gapless line.

This printer is nominally rated at 1100 lines per minute when used with a standard commercial printing train of five arrays of 48 characters each. The chemical train, custom built for us by IBM, contains 69 different characters arranged in four partially identical arrays (Figure 15). This so-called "preferred arrangement" has the frequently used characters on all four arrays and the "nonpreferred" characters represented in only two of the four arrays. This arrangement will permit the train to operate at between 440 and 880 lines per minute, with a nominal speed of approximately 800. The quality of computer printed structures is as good as the typed structures.

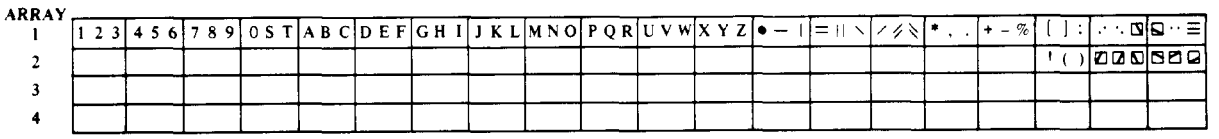
APPENDIX 1

The norbornane shown on the left in Figure 8 has been typed starting at the top and going around the outside of the ring in a clockwise manner, then down the bridge. This is not a particularly efficient path among several possible methods, but will suffice to illustrate the flexibility of motion and the code sequence generated. In the following code sequence resulting from typing the structure as indicated, functions are represented by a lower case letter as follows: i = index, r = reverse index, s = space, and b = backspace; characters are represented by their key number (see Figure 12) with a U for upper case and a L for lower case.

Code Sequence: 15U, i, 31U, i, 15U, i, b, 19U, i, b, 15U, i, b, b, 7U, i, b, b, 15U, b, b, r, 31U, r, b, b, 15U, r, b, 19U, r, b, 15U, r, 7U, 26L, i, 21L, i, b, 15U, i, b, 14L, i, b, b, 17L.

APPENDIX 2

Figure 16,*a* shows the structure of aspirin before computer program coding and condensation. As shown, it occupies a matrix that measures 11 × 14 positions, or a total of 154 codes. This is the number that would be required to store it in this form without condensation; that is, codes for blanks would be required for all of the white spaces in the matrix. To condense the structure, we established a coding scheme that has a signal, the asterisk in the illustration (Figure 16,*b*), that indicates that a binary number is to follow, and records the number of blank positions in binary from the left hand margin on each line to the first nonblank character. We also code any sequence of blanks of three or more that follow



SECOND ARRAY IDENTICAL TO FIRST EXCEPT FOR LAST THREE SLUGS SHOWN.
THIRD ARRAY IDENTICAL TO FIRST ARRAY.
FOURTH ARRAY IDENTICAL TO SECOND ARRAY.

Figure 15. Print train (IBM 1416) layout of SK&F stereochemical structural formula font

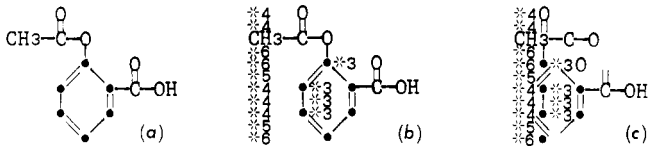


Figure 16. Computer program condensation of structure. Aspirin as typed (*a*), coded (*b*), and coded and condensed (*c*)

the character on the line except where there are no more significant characters until the right hand margin. With all of the left side blanks coded, we can, in effect, push the structure to the left, squeezing out the left side blanks, and then drop the right side blanks. By performing this process, we can reduce the number of positions required to store this structure from the 154 we started with to 61 without record marks and 72 with—a condensation of more than 50%. Figure 16,*c* shows aspirin coded and condensed. Actual storage of the structure is in one continuous record of the lines concatenated, with demarcation characters between the lines.

ACKNOWLEDGMENT

The author gratefully acknowledges the important contributions to this work made by Maxwell Gordon of SK&F and the late J. M. Mullen of Shell Development Co. Thanks are also due Paul Craig, Helen Ebert, and Marianna White of SK&F for their technical and editorial assistance and encouragement.

LITERATURE CITED

(1) Craig, P. N., and H. M. Ebert, "Eleven Years of Structure Retrieval Using the SK&F Fragment Codes," J. CHEM. DOC. **9**, 141–6 (1969).
(2) "Chemical Typewriter Prepares Ring Structures, Complex Formulas," *Chemical and Engineering News* **30**, 2622 (1952).
(3) Feldman, A. P., D. B. Holland, and D. P. Jacobus, "The Automatic Encoding of Chemical Structures," J. CHEM. DOC. **3**, 187–9 (1963); "Survey of Chemical Notation Systems," NAS-NRC Publ. **1150**, p. 424, Washington, D.C., 1964.
(4) Mullen, J. M., "Atom-by-Atom Typewriter Input for Computerized Storage and Retrieval of Chemical Structures," J. CHEM. DOC. **7**, 88–9 (1967).
(5) Gordon, M., "The Potential Impact of Chemical Typewriters on Documentation," *Pharm. Ind.* **28**, 893–7 (1966).
(6) Rice, C. N., K. D. Ofer, R. B. Bourne, and S. W. Logan, "A Pilot Study for the Input to a Chemical-Structure Retrieval System," *Abstracts of Papers*, B14, 151st Meeting, ACS, Pittsburgh, Pa., March 1966.
(7) Jacobus, D. P., K. H. Zabriskie, and M. Gordon, "Compatability in Chemical Information Systems," J. CHEM. DOC. **9**, 118–25 (1969).
(8) Waldo, W. H., and M. DeBaker, "Printing Chemical Structures Electronically: Encoded Compounds Searched Generically with IBM-702," *Proc. Int. Conf. Scientific Inform.*, Washington, D. C., November 16–21, 1958, NAS-NRC, Washington, D. C., 1959; J. CHEM. DOC. **2**, 1–2 (1962).
(9) Hyde, E., and L. Thomson, "Structure Display," J. CHEM. DOC. **8**, 138–46 (1968).

The Multiterm Index: A New Concept in Information Storage and Retrieval*†

HERMAN SKOLNIK
Hercules Incorporated, Hercules Research Center, Wilmington, Del. 19899

Received September 2, 1969

An index not only *can* be a creative communication medium, it *needs* to be in a research and development environment. A creative index is achievable if the relationship and association of things and actions, one to another, can be communicated as a continuous function vis-a-vis the real world of science and technology.

A chemist does not think of a chemical, for example, ethyl alcohol, in isolation. Ethyl alcohol is not merely a word or a term without dimensions to a chemist. It is a concept that he associates with or relates to a product, a reactant, a solvent in a reaction, a use, a property, etc. It is within the semantics of his conceptual needs that he would like to use an index to retrieve those documents he needs. He wants more than documents, however, from the index. He wants the index to direct him to only those documents which are pertinent to his problem. He wants the index to help him to generate thoughts and to suggest new combinations. He wants the index to help him in terms of his language, logic, and semantics and through a generic or specific approach, whichever occurs to him first. He wants the ability to browse among the terms to discover the term that is on the tip of his tongue or recessed in his memory. These are the criteria an index must satisfy if it is to be a creative medium of communication.

Indexing via a strictly dictionary logic is the most prevalent noncreative system. For example, in this type of index, LIGHTNING and LIGHTNING BUG must be placed in a strictly alphabetical order within the L's, and there is no other alternative. The uniterm index, probably the most popular of the dictionary types, in its simplest form would post a document concerned with lightning bug under the separate terms LIGHTNING and BUG (or INSECT, if so directed by a thesaurus). More sophisticated uniterm indexes would employ roles and links to differentiate and to relate some terms. Semantic control in most uniterm indexes, if exercised at all, is through the use of roles and a thesaurus.

An index which has a purpose and which relates with

*Presented before the Division of Chemical Literature, 158th Meeting, ACS, New York, September 1969.

†Contribution No. 1487 from the Research Department of Hercules Incorporated.