was in the searching of unique terms. As mentioned earlier, the percentage of duplication varies according to the data base and the number of profiles being searched. Thus, using unique terms only can significantly decrease the number of matches performed in searching. The second area of improvement was the use of the two-dimensional matrix approach in searching instead of the vector string approach used in the CAS search system.

CONCLUSIONS

The emphasis of this study has been two-fold: to develop statistical models for estimating computer search time for CA Condensates using the UGA text search system, and to make comparative timings between the CAS search program and the UGA search program using the CA Condensates data base.

The two statistical models accounted for over 99% of the variation in search time. The models are currently being used for estimating CPU run times in the University of Georgia Information Science Center. Similar models are being constructed for other data bases run in the information center.

The UGA search program showed a significant improvement in over-all performance over the CAS programs in searching CA Condensates. Other preliminary studies at the University of Georgia Information Science Center have shown even a greater improvement for other data bases searched. However, this is due to the fact that the CAS programs were not specifically designed for searching these data bases. Therefore, the comparative timings were not presented in this study.

ACKNOWLEDGMENT

This work was partially supported by the National Science Foundation under Grants GN-851 and GN-32214.

LITERATURE CITED

(1) Grunstra, Neale S., and Johnson, K. Jeffrey, "Implementation and Evaluation of Two Computerized Information Retrieval Systems at the University of Pittsburgh," *J. Chem. Doc.* **10**, 272-7 (1970).
(2) Bourne, C. P., and Ford, D. F., "Cost Analysis and Simulation Procedures for the Evaluation of Large Information Systems," *Amer. Doc.* **15**, 142-9 (1964).
(3) Blount, C. R., Duquet, R. T., and Luckie, P. T., "A General Model for Simulating Information Storage and Retrieval Systems," HRB-Singer, Inc., Science Park, State College, Pa., April 1966.
(4) Park, M. K., Carmon, J. L., and Stearns, R. E., "The Development of a General Model for Estimating Computer Search Time for CA Condensates," *J. Chem. Doc.* **10**, 282-4 (1970).
(5) "Standard Distribution Format Technical Specifications," American Chemical Society, Washington, D.C., Std. Book 8412-0106-4.

Development and Use of a Termatrex Chemical File System*

THEODORE LEGATT,** ELIZABETH A. BELLAMY, SAMUEL X. deLORENZO
Department of Technical Documentation, Corporate Laboratories,
Research Division, Schering Corp., 60 Orange St., Bloomfield, N.J. 07003

Received June 26, 1972

A storage and retrieval system for chemical structures, using optical coincidence, is in use at Schering Corp. The system rapidly provides the research laboratories with structural information at any generic level. The design, maintenance, costs, and applications are discussed.

One of the responsibilities of a Technical Documentation center in a pharmaceutical organization is to provide information on the chemical structure of compounds. Basically, each compound being evaluated must be recorded and systematized so that at any time, information on its whole or partial structure can be supplied on request.

The internal file of chemical structures at Schering Corp. was reorganized about seven years ago. At that time, a Beilstein-type classification code was used; the structures were incorporated on McBee Keysort cards. Several factors necessitated a reorganization of the chemical coding system. The number of compounds had grown to about 12,000, making the needle-sorting of McBee cards a slow and cumbersome task. As the file size grew, compounds were lost and false drops increased owing to the inherent inadequacies of a Beilstein classification code. In addition, a system was needed to provide the department of analytical chemistry with correlations between structural features and data from nuclear magnetic resonance or infrared analysis. To meet these requirements, a new chemical code and new search equipment were needed. The code had to be relatively easy to learn and capable of searching for generic and specific structures. The search equipment had to be compact, provide rapid output, browsability, and personal control.

DESCRIPTION OF SYSTEM

A modified Ringdoc code and a Termatrex optical-coincidence system[3] were chosen, since they seemed to fulfill our requirements. Optical coincidence was desirable

*Presented before the Seventh Middle-Atlantic Regional Meeting, ACS, Philadelphia, Pa., February 1972.
**To whom questions should be addressed.

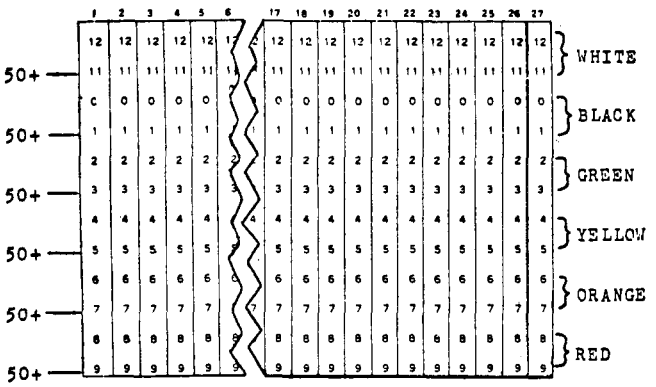
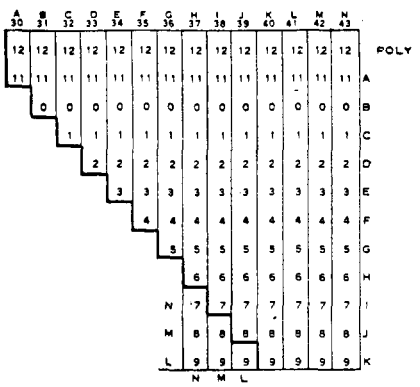


Figure 1. IBM Card.



A - Col. 11 • 12 (CHAIN)
B - Col. 23 (POLYVALENT HETEROFUNCTIONS)
C - Col. 6 (HETEROCYCLIC RINGS)
D - Col. 3 (AROMATIC RINGS)
E - Col. 17 (HALOGENS)
F - Col. 17/4-7 (NO_2, NITRILES, ETC.)
G - Col. 18 (HYDROXY)
H - Col. 18/3-6 (ETHER)
I - Col. 19 (AMINES)
J - Col. 20 (POLYVALENT S, P, Si, B)
K - Col. 22 (AZO + DIAZO)
L - Col. 4 5/4-8 (ALICYCLES)
M - Col. 24 (A-C-A')
N - Col. 26 + 27

Figure 2. Connectivity Matrix.

when file size, desk-top control, and the inaccessibility of computer equipment were considered. The Cancer Chemotherapy National Service Center (CCNSC)[1] and, more recently, Warner-Lambert[4] have reported the use of optical coincidence for the storage and retrieval of chemical structures.

The chemical code format now in use at Schering Corp. is condensed on 27 columns of an IBM card. Figure 1 shows the relationship of the IBM card with the subsequent transfer to the "peak-a-boo" term card. Each chemical descriptor has a corresponding color-tabbed term card. Two rows on the punched card comprise each of the six colors; the number of the term card corresponds with the number representing the column of the specific chemical descriptor. A factor of 50 is added to the column location in the second row of each colored set of digits so that the corresponding Trematrex card can be selected easily. For example, the descriptor bromine, found in column 17, row 2, is identified in Trematrex as green 17. Iodine, found in column 17, row 3, is 50 plus 17, or green 67, in Trematrex. Retrieval from the Trematrex system, therefore, involves relating the format of the chemical code on the punched card to the corresponding term card.

The chemical code was modified for our purposes. We redefined some of the chemical descriptors, adapting them further to our file of compounds. Several new descriptors were added, since a multiple negative search cannot be made in an optical-coincidence system. A major modification was the inclusion of a connectivity matrix in the code. This inclusion will minimize the number of false drops that might occur as the file size increases, which is inherent in fragmentation-type codes. False drops are not necessarily unwanted, as they can restimulate search strategy; however, the option to reduce them is desirable.

The connectivity matrix, designed on an IBM card, is shown in Figure 2. The matrix consists of a horizontal axis, lettered A through N, and a vertical axis, lettered A through K. Because of the size of the card, the vertical axis is continued along the lower central area of the card. Each letter has a chemical meaning which is identified more specifically by the first 27 columns of the code. Thus, a letter can signify either a whole column of chemical descriptors or only part of one. The letter A refers to chain descriptors listed in columns 11 and 12; B refers to polyvalent heterofunctions of carbon atom descriptors listed in column 23. Each matrix letter and its corresponding descriptors are shown in Figure 2. For instance, when an X is coded under column 34 (E), row 1 (C), it signifies the attachment of a halogen to a heterocyclic ring, previously identified in columns 17 and 6, respectively. Although the connectivity code itself is quite broad, it contributes a great deal of specificity when used in conjunction with the appropriate descriptors in the general Ringdoc code. In addition to defining specific group attachments, the connectivity code may be used to retrieve generic groups. For example, a search for all compounds with a halogen attached to an aromatic group may be requested. Since the Ringdoc code does not include the general halogen or aromatic descriptor, a number of individual halogen and aromatic term cards would have to be selected. Using the connectivity code, however, the retrieval would be broadened by inserting only the halogen/aromatic term card.

The connectivity code is used only when retrieval with the main code produces too many false drops. It is, in itself, a fragmentation code and will not necessarily eliminate all false drops. At best, the connectivity code will screen the obvious inconsistencies obtained with the main code. The effectiveness of the connectivity code was evaluated by comparing its ability to reduce the number of false drops with that of the Ringdoc code. Approximately 50 random structure searches were formulated. In each search, only the Ringdoc code was used at first; then the connectivity code was added. If a particular search resulted in ten false drops, and the addition of a connectivity descriptor reduced the number to two, the reduction in false drops was measured as 80% for that search. When the 50 random searches were completed, the average reduction in false drops after use of the connectivity code measured about 85%. Although this figure is applicable to our present file size and structural types, it may not be consistent for other files using the Ringdoc format.

The chemical code, which includes the organic, steroid, and connectivity sections, comprises a Trematrex deck of about 430 structure cards. Each card defines a chemical fragment and has reference numbers of compounds containing that fragment drilled at their specific coordinate locations. Sorter cards have been used to distinguish ste-

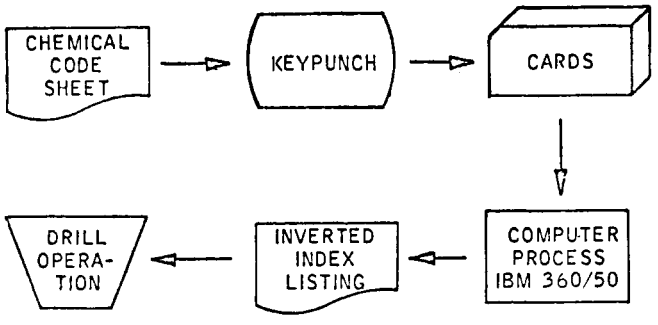


Figure 3. Flow Diagram of Termatrex Operation.

GREEN 67 (COL. 17 - ROW 3)

0001	CC90	0967	2293
05	91	69	2328
07	92	70	29
08	C1C6	74	40
09	09	76	44
10	20	80	48
11	21	101C	2513
12	32	11	14
13	35	27	86
14	40	55	2621
15	50	71	23
71	59	40	3551
72	63	1756	56
75	64	66	57
76	70	86	58
80	C871	87	59
81	C920	1896	75

Figure 4. Printout of Chemical Descriptors Index.

roidal from nonsteroidal compounds. As a result, a double chemical meaning has been assigned to each Termatrex card, thereby reducing the size of the deck. For example, a specific card can identify all compounds containing three aromatic rings in the nonsteroid code, or all compounds containing a 3-hydroxy group in the steroid code. By superimposing the correct sorter card, only those compounds with features in either steroidal or nonsteroidal meaning can be retrieved.

INPUT

Either of two input methods can be used with a Termatrex system. One involves a term-by-term approach, in which all compound accession numbers to be entered on one card are drilled in a continuous operation. The other involves the simultaneous entry of the accession number of one compound into all appropriate Termatrex cards. We developed our own approach, based on the term-by-term method, since we needed to drill duplicate decks, retain a punched by-product deck for future computer application, and avoid the use of more expensive automatic equipment. Figure 3 is a flow diagram of the operation.

Initially, a backlog of 12,000 compounds was converted to the new system. Each compound was recoded and entered into the Termatrex decks. Currently, the file contains over 18,000 structures. Input begins with the chemical coding of compounds on a form designed to facilitate the keypunch operation. The layout of this code sheet is similar to the Ringdoc format, with the exception of the connectivity code, which begins in column 30. Structural information, stored on punched cards, creates a computer-generated, inverted index, which is necessary for term-by-term processing in the final drill operation. As the file increases, the by-product deck of punched cards can also be used for future computer applications.

Every three to four months, approximately 400 new structures numbered consecutively, are coded and checked by our chemical information staff; this stage of processing requires about five minutes per structure. Keypunch and verification requires an additional four minutes per structure. Then the card deck is processed on a 360/50 computer, using a program written in Cobol. The generated listing contains an inverted index of chemical descriptors, as well as the reference number of each compound that contains that descriptor. A printout of one page of this index is shown in Figure 4. At the top of the page is the color code, which indicates the column and row punch of the chemical descriptor. This color code, which corresponds to the individual colored tabs of the Termatrex deck, aids the clerks in selecting the appropriate card for the drilling input. The print format is designed so that the clerk can relate the indented numbers to the X-axis of the Termatrex drill and see at which point the Y-axis should be changed. The manual Termatrex drill must be done by two clerks to ensure completeness and accuracy. This stage of the operation is relatively time consuming: 400 structures can be drilled into the system in about 33 hours per clerk, or a total of 66 hours.

The calculated cost of updating the chemical file on Termatrex, based on a periodic input of 400 compounds, is slightly more than $1.00 per compound. This cost includes only the keypunch, computer, and drill operations; it does not include the initial step of chemical coding, which will vary among organizations.

USE

Since the capacity of a Termatrex deck is 10,000 compounds, we use two decks for our file of 18,000 compounds. The biology, chemistry, or patent departments may request structural information. For the information chemist, responding to a search request is exceptionally simple and convenient with the Termatrex system. The system is compact and located near the searcher. The information chemist has personal control of the unit; therefore, he is not delayed or frustrated by problems that may arise from a more complex system involving other personnel. As a result, structure searches are very rapid and can be processed immediately, if necessary. The output of the system is a list of compound reference numbers. A secondary file of structures is then used to determine the pertinence of the output to the search request. The Termatrex deck of structures is also used in conjunction with our computerized file of biological data to retrieve information on select compounds.[2]

In addition to processing routine requests from laboratory investigators for structure searches, an optical-coincidence system of chemical descriptors readily lends itself to searches of compound types in which the limits of the search are not defined specifically. The limits of such a search are determined by the information chemist as he browses the deck for related structures. For example, a service called "New Drug Activity Correlations" was instituted in our Technical Documentation area. Its purpose was to correlate interesting structural leads reported in the literature with our internal file of compounds. Hopefully, the service will locate and recommend similar types of compounds in our file that had not been tested in the screen and that may possess biological activity, as reported in the literature. The sources used in these correlations include deHaen Drugs in Prospect, Unlisted Drugs,

Pharmascope, patents, or other abstracting services that provide a large group of compounds which can be rapidly scanned for structural leads. Termatrex is ideal for this application. When the limits of a search are not defined precisely, the information chemist can search for analogous structures in any number of ways. Initially, structures closely related to the compound listed in the literature are sought; if none are available in our internal file, the search strategy can be narrowed or broadened quite easily by adding or removing Termatrex cards. Although other systems can perform the same service, an optical coincidence system offers, for our purposes, a rapid, personally controlled, inexpensive approach. As the file increases to a point where optical coincidence becomes unwieldy, conversion to a computerized file will be necessary. Meanwhile, the IBM deck of structures, a by-product of the present system, is accumulating for use at the time of conversion.

ACKNOWLEDGMENT

We wish to thank Al Ruffner and Helen Anderson for their programming efforts.

LITERATURE CITED

(1) Ihndris, R. W., "Structure Fragmentation for Use in a Coordinate Index Retrieval System," *J. Chem. Doc.* 4, 274–7 (1964).
(2) Legatt, T., Grandy, R. P., and deLorenzo, S. X., "A Biologically Oriented Data Retrieval System," *J. Chem. Doc.* 9, 177–83 (1969).
(3) Remac Corp., Gaithersburg, Md.
(4) Starker, L. N., Kish, J. A., and Arendell, F. H., "Multi-Level Retrieval Systems. III. A Generic Chemical Search System Using Optical Coincidence Cards," *J. Chem. Doc.* 10, 206–11 (1970).

CORA—A Semiautomatic Coding System Application to the Coding of Markush Formulas

HUGUETTE DEFOREIT,* ANNE CARIC, HENRIETTE COMBE, SYLVIANE LEVEQUE, ARMAND MALKA, and JACQUES VALLS
Centre de Recherches Roussel-Uclaf, Romainville, France

Received June 14, 1972

A computer system, named CORA, has been devised for coding chemical structures by fragmentation elements. It has been used to encode Markush formulas in patents according to the Ring codes used in the Ringdoc and Pestdoc services and results in an easy, speedy, reliable, and inexpensive method.

This system was devised to simplify the manual coding according to fragmentation codes by using computer facilities.

The need for such simplification is specially felt in the case of encoding Markush formulas in patents which often correspond to a very large number of possible combinations among the various fragments included in the general formula.

That is why we thought of applying the CORA system to the problem of encoding patents published by Derwent in CPI sections B and C (Farmdoc and Agdoc) according to the Ring codes (used in Derwent's Ringdoc, Pestdoc, and Vetdoc).[1,2] This work of coding patents in Ring code is a joint venture undertaken by the 13 Pharma-Dokumentationsring firms.[1,3] We believe, however, that the system can be used for any other fragmentation code.

CODING OF MARKUSH FORMULAS

The semiautomatic system is based upon the decomposition of a Markush formula into a number of elementary fragments which are coded separately. Then our program CORA combines these fragments, taking into account the overcoding rules, calculates the final number of punched cards, and actually produces these punched cards.

For instance, considering the following Markush formula:

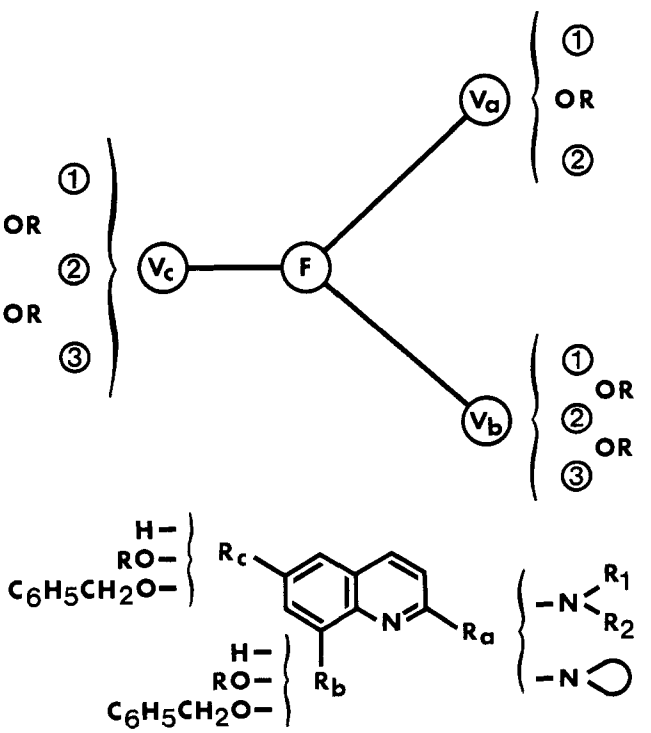


Figure 1

*To whom correspondence should be addressed.