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:

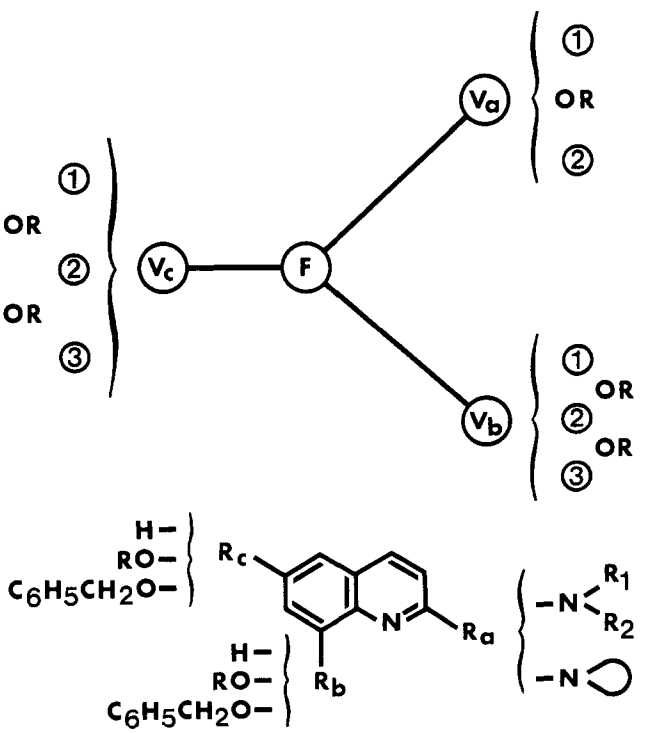


Figure 1

*To whom correspondence should be addressed.

the decomposition leads to fragments linked to one another according to the equation:

F and (Va1 or Va2) and (Vb1 or Vb2 or Vb3)
and (Vc1 or Vc2 or Vc3).

There are 18 combinations theoretically possible for these fragments:

1) F and Va1 and Vb1 and Vc1
2) F and Va1 and Vb1 and Vc2
3) F and Va1 and Vb1 and Vc3
4) F and Va1 and Vb2 and Vc1
5) F and Va1 and Vb2 and Vc2
6) F and Va1 and Vb2 and Vc3
7) F and Va1 and Vb3 and Vc1
8) F and Va1 and Vb3 and Vc2
9) F and Va1 and Vb3 and Vc3
10) F and Va2 and Vb1 and Vc1
11) F and Va2 and Vb1 and Vc2
12) F and Va2 and Vb1 and Vc3
13) F and Va2 and Vb2 and Vc1
14) F and Va2 and Vb2 and Vc2
15) F and Va2 and Vb2 and Vc3
16) F and Va2 and Vb3 and Vc1
17) F and Va2 and Vb3 and Vc2
18) F and Va2 and Vb3 and Vc3

The program applies then the overcoding rules of the Ring codes to all these combinations and calculates the number of final punched cards (here 6 cards out of the 18 combinations).

Finally the program generates these punched cards adding the biocode punches.

The manual encoding necessary for the input is done thanks to four different types of cards (starting, structure, conditions, and biological).

1. The *starting card* with just the patent reference number.

2. The *structure cards* on which the various structural fragments are coded (one fragment per card).

The chemical coding is done in columns 1 to 27 and 31 to 34 (according to the Ring code).

To eliminate the ambiguities of the definitions of certain punch positions (which in the Ring code can have two different meanings), we have made slight modifications for some coding rules and have created four new columns (31-34)—but only for input purposes.

The coded fragment is numbered in columns 36 to 39. The number of the fragment to which the fragment being coded is linked is indicated in columns 40 to 43. The nature of the link between these two fragments is given in column 35 = (and = 35/12, or 35/11).

To make the coding easier, we have created a thesaurus (see Figure 2). For structures existing in this thesaurus, the coder simply writes the word in clear text in columns 44 to 79.

This thesaurus is permanently memorized in the computer, and to have access to it one has to punch position 35/00.

Another possibility during coding of a patent is to memorize temporarily a fragment which occurs several times in the Markush formula; that prevents coding the same fragment several times and so saves time (this temporary memorization needs punch position 35/01).

3. The *conditions cards* enable the computer to put certain punch positions while reducing the number of fragments created. They simplify the work of the coder and can be applied to both the chemical and biological parts. A condition card has the general meaning:

"If there is (or not) a given fragment (and, or) another given fragment, add such and such punch position(s)."

The condition cards enable us also to link certain structures to certain biological properties.

4. The *biological cards* are coded in the classical way according to the biocodes of Ringdoc and Pestdoc.

THESAURUS

ALKOXYALKYL
12/12 12/00 12/01 12/04 12/06 12/07 13/12 13/11 16/07 18/03 18/04
ALKOXYCARBONYL
12/12 12/00 12/01 12/02 12/04 12/06 12/07 13/12 13/11 16/07 23/04 23/07
ALKYL-SUBS
12/12 12/00 12/01 12/03 12/04 12/06 12/07 13/12 13/11 16/07 17/09
ALKYLAMINO-ALKOXY
12/12 12/00 12/01 12/02 12/03 12/04 12/06 12/07 12/08 13/12 13/11 13/01 16/07 18/04 19/02
ALKYLMERCAPTO
12/12 12/00 12/01 12/04 12/06 12/07 13/12 13/11 16/07 18/03 18/04 18/07
ALKYLPYRIDYL
02/12 06/12 06/03 07/01 10/11 10/03 10/04 10/05 10/06 10/07 12/11 12/00 12/01 13/12 13/11 16/07
31/12
ALKYLSULFENYL
12/12 12/00 12/01 12/02 12/04 12/06 12/07 13/12 13/11 16/07 20/00 20/07 21/00
ALKYLSULFONYL
12/12 12/00 12/01 12/02 12/04 12/06 12/07 13/12 13/11 16/07 20/01 20/07 21/00
ALKYNYL
12/12 12/00 12/01 13/12 13/11 14/05 14/06 16/07
ALKYNYLOXY
12/12 12/00 12/01 12/04 12/06 12/07 13/12 13/11 13/04 14/05 14/06 14/07 16/08 18/11 18/04
AMINO ALKYL LOW
11/11 11/00 11/01 11/05 11/06 11/07 12/00 12/03 12/06 12/07 19/00
AMINOALKOXY
12/12 12/00 12/01 12/03 12/04 12/06 12/07 12/08 13/12 13/11 13/01 16/07 18/04 19/00
AMINOALKYL
12/12 12/00 12/01 12/03 12/06 12/07 13/12 13/11 16/07 19/00

Figure 2. Semiautomatic coding of patents

PATENT'S EXAMPLES FORMULAS

The formulas of the patent's examples can be coded with the same method. According to the examples to be coded, two possibilities can occur:

Fragmentation coding similar to that of Markush formulas, the structures being linked by *and* or by *or*.

Individual coding of the structures, the structure cards being linked only by *or*. In this case, punch position 35/04 is added to prevent part of the program from acting.

Description of the Program. The program CORA, summarized in the diagram (Figure 3) is split into 3 phases.

Phase 1. The input stage where descriptions are given of the structures (fragment cards and condition cards) and of the biological properties, together with the thesaurus terms.

All the information is put on magnetic disks, and two files plus one table are created:

The fragment file with it's boolean logic which enables us by linking them to one another to build up the various possible combinations corresponding to the patent's equation.

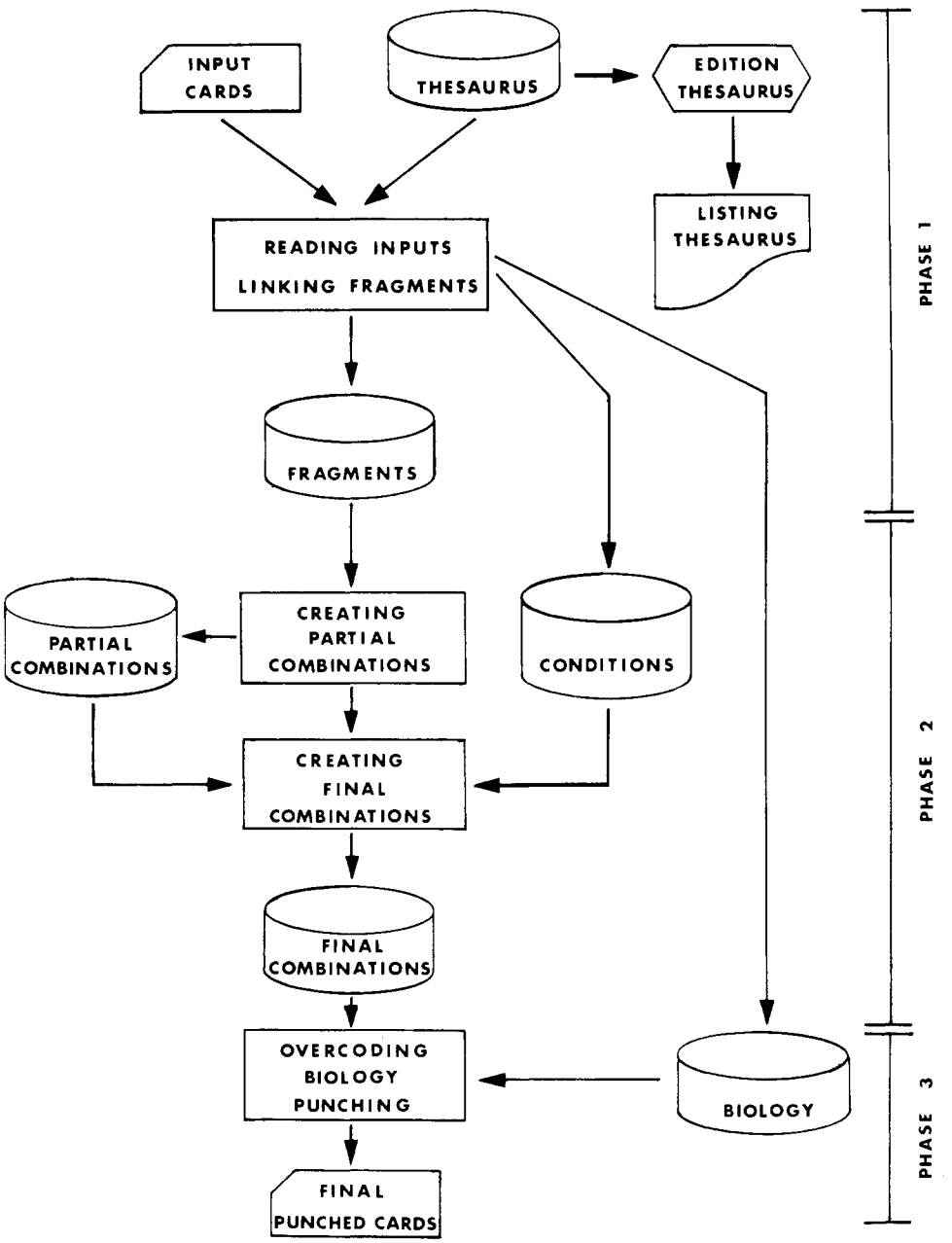
The conditions table which will be needed in phase 2.

The biological file.

The thesaurus may be updated at this stage.

Phase 2. Its purpose is to find out the various possible combinations between fragments and is made of two stages:

Progressive determination and recording of partial combinations.



Finishing the encoding for the final combinations taking into account the conditions recorded in phase 1.

Phase 3. At the beginning of that part, a sorting is made of key-punched cards to help comparison of combinations and overcoding. Then to each final chemical coding the biological coding is added, and the final cards are punched.

Checking. The CORA system includes:

A program to check the codings and find out the main errors, which are clearly specified on a listing produced by the computer.

A program to compare codings of the same patent made in different ways, either purely manually or by using the semiautomatic approach even by splitting up the Markush formula differently.

Efficiency. Some indications on the performance of the program can be given:

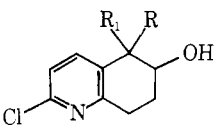
For "medium sized" patents with several hundreds of combinations of fragments (100 to 10,000), the CPU time (IBM 370) needed is about 1 to 4 seconds.

For "large" patents with 10,000 to 100,000 combinations, the CPU time (IBM 370) needed ranges between 1 and 5 minutes.

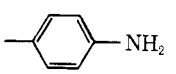
The average time needed for the manual encoding work preceding computer processing is 30 minutes per patent (instead of 1 hour if the patent was coded in an entirely manual way).

Using the Ring codes and overcoding rules gives an average number of six punched cards per patent. The yearly increase of the file for Farmdoc and Agdoc is consequently about 60,000 cards.

Example. In the following very simple example of the input manual coding part of the semiautomatic coding system,

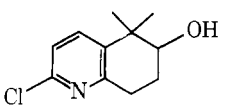


R and R1 can be H— or



the structure is split in 3 parts:

I



II

III R

R1

A. *Structure Cards.*

Fragment ① 1st Coding Sheet:
The skeleton shown in part I.
The position of the substituents on the alicycle is not given at this stage.

Fragment ② 2nd Coding Sheet: R = H

Fragment ③ 3rd Coding Sheet: R = —C_6H_4—NH_2
The coding is put in temporarily, the coding of this fragment being the same as R1 fragment 5.

Fragment ④ 4th Coding Sheet: R1 = H

Fragment ⑤ 5th Coding Sheet: R1 = —C_6H_4—NH_2
Coding by using the temporary memory. In addition, the position of the substituents on the alicycle when R (and, or) R1 = aryl, is indicated.

B. *Condition Cards.*

1st Coding Sheet: It gives the position of substituents on the alicycle when R and R1 = H. The condition is: "If there are fragment ② and fragment ④, add punch 09/03."

2nd Coding Sheet: It gives the geminal substitution and the quaternary carbon in α-position to OH when R and R1 = phenyl. The condition is: "If there are fragment ③ and fragment ⑤, add punches 09/00 and 13/11."

CONCLUSIONS

For patents with simple Markush formulas and even more so in the rather frequent cases where the Markush formulas correspond to hundreds or thousands of possible combinations, the CORA system enables the considerable power and speed of the computer to be used and results in an easy, speedy, accurate, and inexpensive encoding of patents, which otherwise are very difficult to encode and lead too easily to errors and omissions.

Though we have been, up to now, involved mostly in the Ring fragmentation codes, the system can be applied easily to other fragmentation codes.

The purpose of this paper is to describe a computer program useful to encode patents—that is to say, deals only with the input stage. We have left out describing and discussing the use and retrieval efficiency of patents files encoded according to the Ring fragmentation codes for which various computer programs, using either sequential or inverted files, are being used. This retrieval problem is a separate and important development and will be covered in a future publication. Up to now it has been discussed only at Derwent Publications' Meetings.[4]

ACKNOWLEDGMENT

The authors gratefully acknowledge the help of C. Rosenberg and E. Bonnel de Mezieres and the Staff of the Central Documentation Department.

LITERATURE CITED

(1) Nübling, W., and Steidle, W., "The Dokumentationsring der Chemisch-pharmazeutischen Industrie; Aims and Methods," *Angew. Chem. Internat. Edit.*, **9**, 596 (1970).

(2) Pharma-Dokumentationsring e.v. and Derwent Publications Ltd., Ringdoc Instruction Bull., No. 5, 3rd ed., 1971.

(3) The Pharma-Dokumentationsring's member firms are: Farbenfabriken Bayer AG, Leverkusen (Germany); Boehringer Mannheim GmbH, Mannheim (Germany); Ciba-Geigy AG, Basel (Switzerland); Chemie Grünenthal GmbH, Stolberg (Germany); Knoll AG, Ludwigshafen (Germany); E. Merck, Darmstadt (Germany); Metabio, Boulogne sur Seine (France); N.V. Philips-Duphar, Weesp (Netherlands); Rhone-Poulenc, Vitry sur Seine (France); Roussel-Uclaf, Romainville (France); Schering AG, Berlin (Germany); K.Thomae GmbH, Biberach (Germany); Troponwerke Dinklage & Co, Köln-Mülheim (Germany).

(4) Bork, "Retrieval of information from C.P.I., Sections B and C (Farmdoc and Agdoc)," Report of the Derwent 1971 Autumn Meeting, p. 37, 1971.