this group (36 compounds) could be reduced by including a greater number of specified atom pairs in the simple pair bit string, or by dividing the group into two for the general pairs C–X and X–X. The other groups of identical strings for all pair types taken together contained only bits for very common pairs, accounting for the lack of differentiation of bit string representations. For example, there were 26 compounds, the bit strings of which contained only bits for simple pairs C–C and general single bond augmented pairs OC–C1 and 1C–C1, and bonded pairs C–C– and –C–C–. This finding was also borne out at the separate pair levels and for octuplets and four-atom fragments, and is in line with the general principle that structures containing rarely occurring features are easy to search for, whereas common structural features must be described in greater detail to be easily retrievable. This group analysis is similar to earlier work involving the distribution of molecular formula group sizes[5] and leads to similar conclusions.

CONCLUSIONS

Good differentiation of the majority of structural representations in the file is obtained using the set of pair screens at present available in the Sheffield substructure search system. The inclusion of other fragment types as screens promises further improvements in the direction of unique screen representations for structures, leading to improvements in systems performance.

ACKNOWLEDGMENTS

We thank Chemical Abstracts Service for provision of the data-base, and the Office for Scientific and Technical Information, London, for financial support for this work. V. A. C. acknowledges the award of an OSTI Research Studentship.

LITERATURE CITED

(1) Adamson, G. W., Cowell, J., Lynch, M. F., McLure, A. H. W., Town, W. G., and Yapp, A. M., "Strategic Considerations in the Design of a Screening System for Substructure Searches of Chemical Structure Files," *J. Chem. Doc.*, **13,** 153 (1973).

(2) Milne, M., Lefkovitz, D., Hill, H., and Powers, R., "Search of CA Registry (1.25 Million Compounds) with the TSS," *J. Chem. Doc.*, **12,** 183 (1972).

(3) Crowe, J. E., Lynch, M. F., and Town, W. G., "Analysis of Structural Characteristic of Chemical Compounds in a Large Computer-Based File. Part 1. Non-cyclic Fragments," *J. Chem. Soc. C,* 990 (1970).

(4) Adamson, G. W., Creasey, S. E., and Lynch, M. F., "Analysis of Structural Characteristics of Chemical Compounds in the Common Data Base," *J. Chem. Doc.*, **13,** 158 (1973).

(5) Bragg, J. H. R., Lynch, M. F., and Town, W. G., "The Use of Molecular Formula Distribution Statistics in the Design of Chemical Structure Registry Systems," *J. Chem. Doc.*, **10,** 125 (1970).

Semiautomatic Coding of Steroid Markush Formulas

J. FITTING, H. LEHNA, G. RIEGE, and K. SPECHT*

Research Laboratories, Schering AG, Berlin/Bergkamen, Germany

Received December 27, 1973

Manual coding of complicated Markush formulas is very time consuming. A semiautomatic method for the encoding of steroid Markush formulas is described. Manual encoding is necessary only for the basic structure and the variables. The permutation is done automatically by computer.

The Pharma-Dokumentationsring e.V. is encoding the FARMDOC and AGDOC patents of the CPI service (sections B and C) by using the Ringdoc- and Pestdoc-Codes. This work is done in cooperation among the Ring members of the corresponding working group. Years ago (in pre-CPI time), the Ring members worked out special rules for patent encoding because they were and are convinced that the expense of time and money is justified.

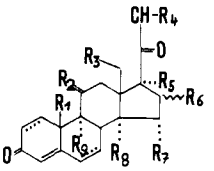
The patent coding rules of the RING allow the overcoding of chemical and biological information to a certain extent depending on RINGDOC and PESTDOC coding rules. According to these rules the encoding of Markush formulas will lead very often to a remarkably great amount of punch cards, and, therefore, it is time-consuming. As far as organic compounds other than steroids are concerned, a publication about a semiautomatic coding method has been issued from Roussel-Uclaf (H. Deforeit, A. Caric, H. Combe, S. Leveque, A. Malka, and J. Valls, "CORA—A Semiautomatic Coding System Application to the Coding of Markush Formulas," *J. Chem. Doc.*, **12,** 230 (1970)).

The method described here for semiautomatic coding of steroid Markush formulas has been developed by Schering in cooperation with the research oriented data processing section (Datenverarbeitung Forschung) and the central documentation department.

With regard to the overcoding rules, a computer program has been developed for the semiautomatic coding. The "Basic-Structure" and the variable structures are coded once only. The permutation will be done by the program. By indicating additional conditions, some of the resulting cards can be changed following permutation.

Sometimes, the encoding of steroid Markush formulas requires 100, 1000, or more punch cards. This is due to the fact that Markush formulas sometimes cover 100,000 or even 1 million mathematically possible combinations, and that—for retrieval purposes with a minimum of false

* Author to whom correspondence should be addressed.



R_1 = H, CH_3

R_2 = Oxo, ß-Hydroxy, ß-Halogen (if R_9 is chlorine, R_2 always is a halogen)

R_3 = H, lower alkyl

R_4 = H, NH_2

R_5 = H, OH

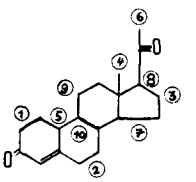
R_6 = H, ∝ CH_3, ß CH_3

R_7 = H or combined with R_8 14, 15 Epoxy

R_8 = H, OH or combined with R_7 14, 15 Epoxy

R_9 = H, Cl

Figure 1.



• : VARIATION OF THE BASIC STEROID

NO ADDITIONAL PUNCH POSITIONS

① Vb1 - •
Vb2 Δ 1

② Vc1 - •
Vc2 Δ 6

③ Vd1 - •
Vd2 - 16∝CH_3
Vd3 16ß CH_3

④ Ve1 - •
Ve2 18 Alkyl

⑤ Vg1 NOR, 10ß H
Vg2 - •

⑥ Vg1 - •
Vg2 21 NH_2

⑦ Vh1 - •
Vh2 14∝OH
Vh3 14∝, 15∝ epoxy

⑧ Vi1 17∝ H
Vi2 17∝ OH

⑨ Vj1 11ß OH; 11 Keto ••
Vj2 11ß Halogen ••

⑩ Vk1 - •
Vk2 9∝ Cl

Figure 2.

$$K \wedge (Vb1 \vee Vb2) \wedge (Vc1 \vee Vc2) \wedge (Vd1 \vee Vd2 \vee Vd3) \wedge (Ve1 \vee Ve2) \wedge (Vf1 \vee Vf2) \wedge (Vg1 \vee Vg2) \wedge (Vh1 \vee Vh2) \wedge (Vi1 \vee Vi2) \wedge (Vj1 \vee Vj2) \wedge (Vk1 \vee Vk2)$$

Figure 3.

drops—the possibilities of overcoding are restricted. Figure 1 shows an example of such a Markush formula. The encoding of this formula will require 1728 punch cards as based on the restrictions of the overcoding rules. The encoding of such "big patents" naturally is very time-consuming.

The use of the program can diminish the necessary time remarkably. Figure 2 shows the various files which have to be coded only. The basic structure is marked with a "K," the variables with a "V." The variables marked with a star are assigned to the BASIC structure; they bear only the special punch positions for the Steroid Code (ϕ1/& Λ ϕ1/-) and an internal file number. These cards are necessary because not only those structures are to be produced which differ from the basic structure K. The various possible combinations are produced by the permutation program. The corresponding logical links of all data sets are shown in Figure 3.

Biological information which is common to all possible compounds are punched on the K-cards. If biological activities are linked to special structures these activities are punched in the respective V-cards. If biological activities are only valid in combinations with special structures, then this information can be punched following permutation. This procedure is done automatically too by help of special condition cards called "Qualifiers." The Qualifiers are necessary, *e.g.*, if permuted combinations need further punches as "POLY-punches," etc.

Following combination and consideration of "Qualifier conditions," the cards are punched automatically.

This semiautomatic coding method should be used only if the coder has an advantage using it. This fact will always occur if one basic structure has to be combined with a lot of variables. In the future, the permutation formerly executed manually can be done by computer. The program is now running under test conditions. So far we have coded 16 steroid patents, automatically generating about 2000 punch cards. The CPU time needed for all these patents was about 3 min. In our opinion, much more money and time are necessary for manual keypunching than for the computerized execution. In addition, this system makes it very easy to check the coding of Markush formulas; in the past, it was nearly impossible, as far as the coding of the so-called "big patents" is concerned.