Patent Office
Research and Development
Reports . . . . . No. 7

A PUNCHED CARD SYSTEM FOR SEARCHING STEROID COMPOUNDS

Prepared by

Julius Frome
Jacob Leibowitz

Patent Research Specialists
Staff Members

Office of Research and Development

U. S. Patent Office

July 8, 1957

Robert C. Watson
Commissioner of Patents

Sinclair Weeks
Secretary of Commerce

A PUNCHED CARD SYSTEM FOR SEARCHING STEROID COMPOUNDS

The success attending the mechanization of 370 steroid patents in a single subclass of the Patent Office classification, encouraged us to extend the search system to include all steroid compounds contained within the 12 subclasses which pertain to steroid patents.

These totaled about 2,350 patents, but of this number over 900 were duplicate patents appearing as cross-references in these 12 subclasses. As a result it was necessary to code only 1,420 patents.

The expansion of the project necessitated the establishment of additional descriptors and the modification of some of the former descriptors.

After careful study, 24 descriptors were selected which were to be associated with particular positions of substitution on the steroid nucleus and 8 or 10 additional descriptors were selected which were general in nature and not specifically identified with a particular position in the nucleus. The choice of descriptors was based on past experience as to the type of search questions most likely to occur in the steroid art. The list is, of course, modifiable in accordance with the knowledge gathered through experience in the actual operation of the system.

Composite Formulas

Patents frequently contain as a part of their disclosures a type of formula representation known as the "Markush" formula. The Markush formula represents a generic portrayal of a class of compounds. However, in contrast to what may be called a natural genus in compounds such as "amines," "phenols," etc., the Markush genus is an artificial, or synthetic genus. It is usually expressed by a structural formula representation wherein a portion of the molecule is represented as a constant while other portions are variables.

Fig. 1 illustrates a Markush formula. Some of the portions of the generic formula thus represented are fixed and unvarying while others like the members of each set of embodiments for each of R_1, R_2 and R_3 are varying. Only one member of a set of embodiments is simultaneously present in the molecule with only one member of each of the other sets.

It is easily seen that given only a few such variables, the number of specific compounds within the scope of such a formula may be fantastically high. However, each embodiment of the genus, whether actual or theoretical, is regarded as a valid disclosure of the concept of the compound and is of potential pertinence in the examination of patent applicatcns. It is therefore desirable that each and all of these many

possible embodiments be retrievable in any search on the basis of these embodiments. In addition to the Markush formulation the patent document may disclose generic formulas of the "true" genus type and specific compounds as illustrative of the various classes.

To cope with this situation, the coding system adopted treated each of the substituent groups at each position as if it were present at the same time as other substituents in the same group although this leads very often to a chemically impossible formulation.

By this method, the Markush formula is treated as a simple formula of constant configurations with each of the variable groups being simultaneously present, thereby converting, in effect, the variables into constants. In addition, where a series of related compounds are disclosed, a formula of Markush type was synthesized preparatory to encoding by combining the variables with the chemical configuration common to these compounds into a composite formula.

This method not only made the task of encoding the multitudinous numbers of combinations within the scope of the patents comparatively easy but it also permitted the retrieval by machine of each of the combinations within the scope of the disclosure. It will be obvious that the formula as coded is ordinarily an impossible one, from theoretical chemical considerations. Since, however, a search request will not usually be made for theoretically impossible compounds, this does not lead to any great difficulties. It will also be noted that in view of the composite formula principle, any retrieval of disclosures in response to a search request may include some non pertinent disclosures. However, all references of pertinence, if any, are, according to the principles of the system, concomitantly retrievable.

A further principle employed in the system is the principle of multiple designation which means that any compound or class of compounds is described by as many descriptors, both general and specific as are applicable.

The system is limited to the steroid art and the questions to be asked of the file are confined to said art. Therefore, every compound coded and to be searched must contain at least the steroid nucleus. This permits the use of a fixed numbering system, the fundamental building block being the steroid nucleus with its fixed positions for nuclear substitution.

Descriptors

The descriptors used are shown in Figs. 2a and 2b which illustrate the work sheet used for coding each patent document. Fig. 2a shows a list of terms or symbols such as: " = "(double bond), "OH", "NH_2", etc., of variable scope as to genericity. In the same row with each of these terms is a list of numbers from 1 to 22 which refer to the positions of substitution on the steroid nucleus according to a

selected numbering system. The sheet also indicated the code for each term and its position of substitution to be translated into the corresponding punches on the punched card, see Fig. 4.

Fig. 2b the second part of the work sheet show descriptors which are not correlated with any particular position of substitution on the steroid nucleus. Each term has a particular bit position on the punched card. Thus O-acyl is in column 60 as indicated at the heading and in row 0 as indicated to the left of the term. A carboxylic O-acyl is in the same column, row 1, etc. The work sheets of Figs. 2a and 2b represent the encoded data for a composition formula Fig. 3, the encircled portions represent the codes. Thus, a halogen is shown in positions 2, 4 and 21. By the principle of multiple coding in Fig. 3 under column 66, the term "halogen" is encircled. The terms "bromine" and "chlorine" are also encircled. This will permit the searching and finding of a steroid containing a halogen generically or a bromine or chlorine specifically regardless of position of substitution.

On the other hand, a question may be asked wherein it is required that a halogen be on the 2 position of the steroid nucleus and said question will retrieve the disclosure illustrated in Fig. 3.

This method of coding permits economy of space and time. The numerous details of each patent were recorded on one punched card per patent.

The Arrangement of the Punched Card (Fig. 4)

The punched card is roughly divided into two sections. Columns 1 to 48 are reserved for recording the descriptors for substituents associated with the steroid nucleus at a designated location thereof. The second portion, columns 60 to 69 is reserved for general descriptors not identifiable with any particular position on the nucleus.

The first field is further divided into 24 fields of 2 columns each. These 24 fields correspond with the 24 substituents listed in the left hand column of Fig. 2a. Each of the 2 column fields has 24 locations for punches, 12 in each column. These are used to designate the 22 positions of attachment of the substituent to the steroid nucleus. Punches 1 to 9 of the first column of each pair of columns correspond to positions 1 to 9 of the nucleus and punches 0 to 9 of the second column of each pair designate the positions from 10 to 19. Position 20 is indicated by the 11 punch in the second column, 21 by the 12 punch in the second column, with the 22 position going back to the 0 punch in the first column, (the appearance of position 22 out of sequence was due to a belated entry of this term).

The general descriptors in columns 60 to 69 are shown in Fig. 2b and a punch in any column and row in this field indicates the presence of the portion of the compound meeting the qualifications set out by the descriptors.

Searching

By requiring that the search select those cards having at least a prescribed pattern of holes, it is possible to process the punched cards on a conventional sorting machine. If more than one descriptor is to be searched at a single pass of the cards, it is necessary to use a multiple column sorting machine.

Operation of the System

To date 64 new patent applications have been mechanically searched and thereafter fully acted upon by the Patent Office. Since these applications have never been searched manually, the effectiveness of the mechanical search must be judged by the past experience with actions made on specifically different steroid applications. A comparison indicates that there is a considerably greater number of direct anticipations retrieved by machine selection than is ordinarily found by manual searching. At least 10 of the 64 applications were felt to be fully anticipated whereas manual searching only rarely uncovered a full anticipation.

The examiner performing the mechanized search had about 6 months' experience in steroid art and $3\frac{1}{2}$ years' experience in the Patent Office. This examiner ordinarily would be expected to examine only 3 or 4 new applications each 40 hour week, whereas with the machine search system, the examiner was able to produce at the rate of 25 applications a week. It is important to note that each search was conducted through the entire steroid art whereas the average manual search is made through only a portion of the art. Comparison on an equal number of patents basis would therefore further multiply the productivity factor stated above.

In addition to speed of searching, there are other features of significance in the machine method which point toward an improvement in quality of the search. Firstly, each search is more complete and comprehensive in its scope than is possible as a practical matter of a manual search method. Secondly, prior to encoding, each patent is carefully and thoroughly analyzed in all its details. The results of this exhaustive analysis is then recorded on the punched cards. Each machine search thereafter, while performed very rapidly, involves in effect the same careful scrutiny of each patent as was performed ab initio.

Coding

The coding of the steroid patents is now completed so that it can be used for actual searches. The patents were analyzed by a group of patent examiners who carefully read the patents and marked the work sheets. Approximately 100 man-days of effort were expended by the Patent Office in performing this analysis and coding. As a further check of the accuracy of the file, the whole group of patents is now being

re-analyzed to insure that there have been a minimum of errors. All work has been done on the patented art insofar as steroid compound disclosures are concerned with no non-patent literature included.

Although it is realized that the coding of approximately 2,000 patents may be insufficient on which to base conclusions, it is felt that some tentative conclusions may be made.

1. Coders should have knowledge and experience in Patent Office principles and practice.

2. Motivation is of extreme importance.

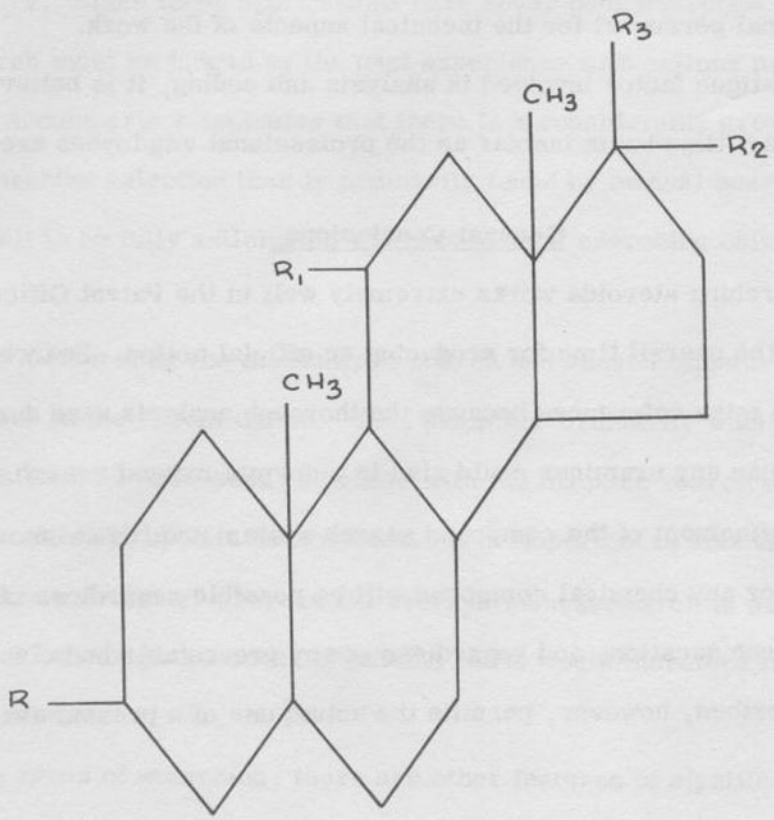
3. It is desirable to relegate as much as possible of the routine work to clerical personnel, using the skill of the professional personnel for the technical aspects of the work.

4. Because of the fatigue factor involved in analysis and coding, it is believed the task could be more effectively done on a part time basis insofar as the professional employees are concerned.

General Conclusions

The system for searching steroids works extremely well in the Patent Office giving a factor of at least 5 to 1 improvement in the overall time for producing an official action. Searches are more thorough and there is less tendency to miss references because the thorough analysis used during the coding operation is much more detailed than any examiner could give in a normal manual search operation. Work is continuing toward greater refinement of the compound search system and toward more universality. By this universality, a search for any chemical compound will be possible regardless of the type of art involved or the type of the search question, and regardless of any pre-established classification scheme. The steroid system just described, however, permits the actual use of a present available tool for a most pressing current problem.

FIG. 1



wherein R is α-hydroxy or β-hydroxy; R_1 is hydrogen, α-hydroxy or an acyloxy ester thereof such as, for example, formyloxy, acetoxy, benzoyloxy, propionyloxy, butyryloxy, valeryloxy, hexanoyloxy, phenylacetoxy, octanoyloxy, or the like, especially lower alkanoyloxy, or β-hydroxy, or ketonix oxygen; R_2 is hydrogen or hydroxy; and R_3 is acetyl, acyloxyacetyl , e.g., acetoxyacetyl, propionoxyacetyl, butyryloxyacetyl, octanoyloxyacetyl, benzoyloxyacetyl, or the like, especially acyloxyacetyl wherein the acyloxy group is a lower-alkanoyloxy group, or haloacetyl, e.g., bromoacetyl, chloroacetyl, or the like.

FIG. 2A

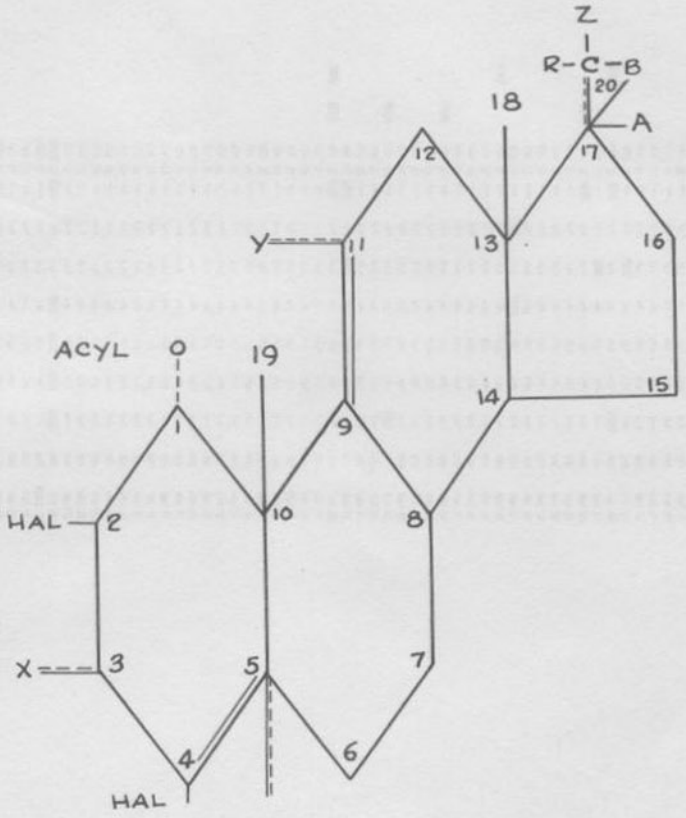
	0																					11	12
=	22	1	2	3	(4)	5	6	7	8	(9)	10	11	12	13	14	15	16	(17)	18	19	20	21	
α or allo	22	(1)	2	(3)	4	(5)	6	7	8	9	10	(11)	12	13	14	15	16	(17)	18	19	20	21	
-C≡C-	22	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
CH_3	22	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
CN	22	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
COOH or COOR	22	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	(20)	21	
-C- sub	22	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
-H	22	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
NH_2 or N<	22	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
O H	22	1	2	(3)	4	5	6	7	8	9	10	(11)	12	13	14	15	16	(17)	18	19	20	(21)	
=O	22	1	2	(3)	4	5	6	7	8	9	10	(11)	12	13	14	15	16	17	18	19	(20)	21	
-Se-	22	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
-S-R	22	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
Hal	22	1	(2)	3	(4)	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	(21)	
Hydrocarbon	22	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
Ketal	22	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	(20)	21	
Ketone reagent	22	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
Epoxy	22	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	(17)	18	19	(20)	21	
-O hydrocarbon	22	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
-O acyl	22	(1)	2	(3)	4	5	6	7	8	9	10	(11)	12	13	14	15	16	17	18	19	(20)	(21)	
-O-hetero	22	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
-N-hetero	22	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
S-hetero	22	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
Miscellaneous	22	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	

FIG. 2B

60		61		62		63	
(0)	O-Acyl	0	O-Hetero	0	N-Hetero	0	S-Hetero
(1)	Carboxylic	1	Morpholine	1	Morpholine	1	Thiophene
2	Poly	2	Furan	2	Piperidine	2	Thiazole
3	Unsat	3	Lactone	3	Pyridine	3	
(4)	Aromatic	4	Spirostane	4	Pyrimidine	4	
(5)	Aliphatic	5	Sub in O spiro ring	5	Pyrrole	5	
(6)	Substi	6	Psuedosapo.	6	Thiazole	6	
(7)	St. chain	7		7		7	
8	Cyclo-alkyl	8		8		8	
9	Branched	9	Misc.	9	Misc.	9	Misc.
11	Hetero-cyclic	11		11		11	
12	Inorganic except hal						

64		65		66		67	
(0)	Bile Cpds.	(0)	Sterols	(0)	Hal.	0	
(1)	Acids	1	Ergosterol	1	Fl	(1)	Androstane
(2)	Cholanic	(2)	Cholesterol	(2)	Br	2	Addition compound
3	Norcholanic	3	Vitamin D_3	3	I	3	Maleic adducts
4	Bisnor-cholanic	4		(4)	Cl	(4)	Pregnane
					Double Bonds		
5		5		5	5 (6)	(5)	21 Unsub-stituted
6		6		6	5 (10)	6	21 diazo
7		7		7	8 (9)	7	
8		8		8	8 (14)	8	
9	Misc.	9	Misc.	9	1 (2)	9	Misc.
11		11		11	1 (10)	11	
12		12		12		12	

FIG. 3



The dotted line denotes the alpha or allo configuration.

R is keto, cyclic ketal, acyloxy, hydrogen, or together with the carbon forms a carboxy group.

B is a double bond, oxygen or zero.

Z is hydrogen, methyl, hydroxymethyl, acyloxymethyl, halomethyl

A is hydrogen or α-hydroxy.

X is α-hydroxy, β-hydroxy, keto or acyloxy.

Y is α-hydroxy, β-hydroxy, keto or acyloxy

Hal is halogen

Acylo is acyloxy

FIG. 4