Patent Office
Research and Development
Reports . . . . . No. 17

SEMI-AUTOMATIC INDEXING AND ENCODING

Prepared by

Julius Frome

Staff Member

Office of Research and Development

Patent Office

December 10, 1959

U. S. DEPARTMENT OF COMMERCE
Frederick H. Mueller
Secretary

PATENT OFFICE
Robert C. Watson
Commissioner

For sale by the U. S. Department of Commerce, Washington 25, D. C. Price 25 cents.

Acknowledgments

The author wishes to acknowledge the help of Mr. C. Dodds and Mr. J. Liptock in the writing of the RAMAC programs created by the author.

Contents

U. S. DEPARTMENT OF COMMERCE
Patent Office

RESEARCH AND DEVELOPMENT REPORT NO. 17 - SEMI-AUTOMATIC INDEXING AND ENCODING

ERRATA

Page 21 - Cancel all compounds beginning with "Acrylic Acid Esters-3", through "Xylens-2" on page 22, and substitute the following:

The folowing is a list of all solvents in 2, 334, 195:

2334195	BENZENE-2	06
2334195	BUTANOL-2	01
2334195	CARBON TETRACHLORIDE-2	00
2334195	CHLORBENZENE-2	00
2334195	CHLORINATED ALIPHATIC HYDROCARBON-2	00
2334195	DICHLORETHYLENE-2	00
2334195	ETHYLENE CHLORIDE-2	00
2334195	HALOGENATED HYDROCARBONS-2	00
2334195	METHANOL-2	06
2334195	ORGANIC SOLVENTS-2	05
2334195	ORGANIC SOLVENT-2	01
2334195	SOLVENTS-2	00
2334195	SOLVENT-2	01
2334195	TOLUENE-2	01
2334195	TRICHLORETHYLENE-2	00
2334195	XYLENE-2	00

The following is a list of all the momomers in 2, 334, 195:

2334195	ACRYLIC ACID ESTERS-3	00
2334195	ACRYLIC ACID ETHYL ESTER-3	00
2334195	BUTADIENE-3	00
2334195	ETHYLENE-3	19
2334195	ISOBUTYLENE-3	01
2334195	PROPYLENE-3	00
2334195	STYRENE-3	01
2334195	UNSATURATED SUBSTANCES-3	00
2334195	UNSATURATED COMPOUND-3	00

USCOMM--DC

SEMI-AUTOMATIC INDEXING AND ENCODING

INTRODUCTION

Most mechanized information retrieval systems are dependent upon the extraction of information from documents, processing such information in some manner and finally placing this information in a file so that the information can be retrieved by machine. Any such retrieval system is only as good as the degree of accuracy and thoroughness of the extraction of information from the document.

Analysis can be defined as the process of extracting information from a document. This information is often in the form of notions or concepts.

Indexing processes the information resulting from analysis by placing it in lists.

Coding creates and assigns specific unambiguous terms as substitutes for such notions or concepts (13). The terms may come from the document itself.

Encoding prepares a document for storage by substitution of codes or terms for the information already extracted from the document.

To date, in the development of most of the mechanized information retrieval systems, including the three mechanical retrieval systems developed in the U. S. Patent Office (1,2,3) the major portion of the effort of analyzing and encoding documents has been essentially manual. Only a small portion of this work has been mechanized.

It is desirable to attempt to shift the major burden of the analyzing and encoding from humans to machines. This paper will describe and illustrate an experimental attempt to so do. This attempt suggests a possible interim solution to these problems. The extraction, listing, indexing and encoding of the structural formulae not specifically named in the document are treated in another manner not included in this paper.

Four extremely important factors which have been considered are those of (1) personnel, (2) training, (3) product obtained and (4) cost. I have called this "a semi-automatic indexing and encoding method". In this experiment humans are used for what they are best suited, e.g. reading and recognition of concepts, and machines for what they can do best, e.g. fast, accurate, and repetitive action.

The usual two step procedure was used, consisting, first, of extracting the material from the document and, then, encoding it in a form suitable for making searches. The particular method used in the first step additionally resulted in a method for indexing. Such auxiliary indexing is of extreme importance in serving as a surrogate for the document itself for the purpose of browsing in manual retrieval systems. The first step will, accordingly, be called indexing.

An excellent start in automatic abstracting has been made by H. P. Luhn (4,5,6,7) and associates. However, these systems cannot now be used by the Patent Office. Even if the text of the documents were available in "machine-readable" form and a computer such as the IBM 704 were available, no program for performing the type of analysis we find necessary is now known. The ultimate solution awaits not only the development of machines able to read any text in any type font, but the discovery of those rules of language that would enable machines to process the text on the basis of its information content. Until such developments are forthcoming, humans will still play an important part in the task of analyzing and encoding documents. H. P. Luhn (8) and C. K. Schultz et al. (10) also suggest mechanized methods for indexing terms which they extract.

INDEXING

Indexing is the listing of words, subjects, or concepts found in, or otherwise associated with a document, which enables a person to find that document. Indexing in depth is an extremely useful tool for Patent Office use. Deep indexing by the usual manual methods is quite expensive and has required very highly trained personnel (11). In "subject" indexing particularly, extracted terms or concepts are placed under headings according to some specified scheme of classification. This requires personnel skilled both in the discipline involved and in the art of classification.

Most indexes carry no indication of the depth of the indexing, since this factor is left to the judgment of the indexer. While he is always conscious of the time required to make the index, he is, perhaps, not as conscious of all the varied requirements of the users. If a method of indexing is chosen which has definite rules, a less skilled indexer can be used. The uniform application of these rules would control the depth of indexing, and knowledge of them will indicate that depth. Our experimental method enables the user to know both the depth of indexing and the point of view from which the indexing was carried out.

Indexes in which the user can find the item desired either in the *language of the document* or under an appropriate subject heading appear desirable.

The following steps were taken in preparing the index:

1. Determination of the point of view or objective of indexing, e. g., extracting the chemical compounds from a document or chemical processes; or

in disciplines other than chemistry extracting electrical components or systems.

2. Determination of depth of indexing desired, e.g. extracting *all* chemical compounds, *all* steroid compounds, *only* phosphate compounds or *only* electrical components.

3. Determination of the rules for extracting terms or concepts, in accordance with the point of view and depth of indexing, is communicated to the indexer. The more precise and simple the rules are, the less training of personnel is needed, and the better the index.

4. Conversion of extracted terms to machinable form, e.g. converted into punched cards.

5. Elimination and counting of duplicates by machine processing. The elimination of synonyms is not done at this time.

6. Printing of lists or indexes by machine.

These first steps now result in a list of terms for each document. The machine can then arrange and print an alphabetical list of all the terms in all the documents indexed, again eliminating the duplicates, while indicating the plural documents in which each term appears.

Experimental work has been carried on in the U. S. Patent Office since April, 1959. Several specific examples of this system will now be described.

Steroid Example

All "steroid compounds," not shown by structural formulae, are extracted from each patent of the 2,100 patents in the twelve steroid subclasses (1).

The technical personnel were instructed to underline each word which recited a steroid compound as it occurred and *every time* it occurred. Thus a steroid compound occurring seven times was underlined seven times. Hence no effort or time was lost in determining whether a compound was previously underlined. Structural formulae were not underlined. The underlined patent was next given to a punch card operator who punched the underlined work in alphanumeric text. One card was punched for each steroid compound. Trailer cards were used as needed. The patent number in which the word occurred was also punched in each card. A RAMAC 305, by suitable program, accepted these cards, copied each card into its disc memory and then eliminated duplicate information. Next it punched out a new deck of cards which eliminated duplicates and printed a list of steroid compounds in alphabetical order, containing no duplicates. Each card and its listing included the patent number and the number of times the term appeared in the patent. This list constitutes an index of *all* the steroid compounds in each patent, excluding those represented by structural formulae. These could be easily included if desired. The resulting list tends to be complete and accurate because of the way it is generated. In most patents each steroid compound is usually repeated two or more times. If the indexer failed to underline one occurrence, it is probable that the second or third occurrence would be underlined. Furthermore, since the list also contains the number of times the compound occurred, this may be an indication of its importance in the document (5,6). This list can be attached to the patent to aid in manual searching of the patent. It also enables one to quickly see if a particular steroid compound is present in the patent. Specific examples of an underlined patent and such a prepared list can be seen in the appendix. (A). A flow chart of the program is also appended. (B). Another RAMAC program takes the alphabetized deck of the entire file of documents, and assigns address numbers of the RAMAC storage (9) for each different concept.

Personnel

In contrast to other methods, this method of indexing was done with personnel who have only a B. S. degree in chemistry. No other formal schooling appeared to be necessary. In fact, in two instances, personnel with only 1 year of chemistry were trained to do this work satisfactorily. However, a science degree is more desirable because it reduces the amount of training necessary. Highly trained indexers are not needed because a knowledge of steroid chemistry is not necessary. All that is required is the ability to recognize that a word describes a steroid compound.

Training Time

Relatively little time is required to train an indexer adequately to recognize steroid compounds, rarely more than a week. His training consists of supervised extraction and each of his errors are discussed in detail with him during the training period.

Cost

One of the attractive features of this system is low cost. The average trained extractor can underline thirty-two steroid patents a day. The average salary is about $2 per hour. Thus this intellectual effort costs about fifty cents per patent. Punching costs about $1.45 per patent. Machine processing on the RAMAC costs about 90 cents a document. Additional steps in processing account for an additional direct cost of $.25 a patent. Hence the average cost of the index per document is about $3.10. For this cost both the printed list or index and the punched cards, which can be used in encoding, are obtained.

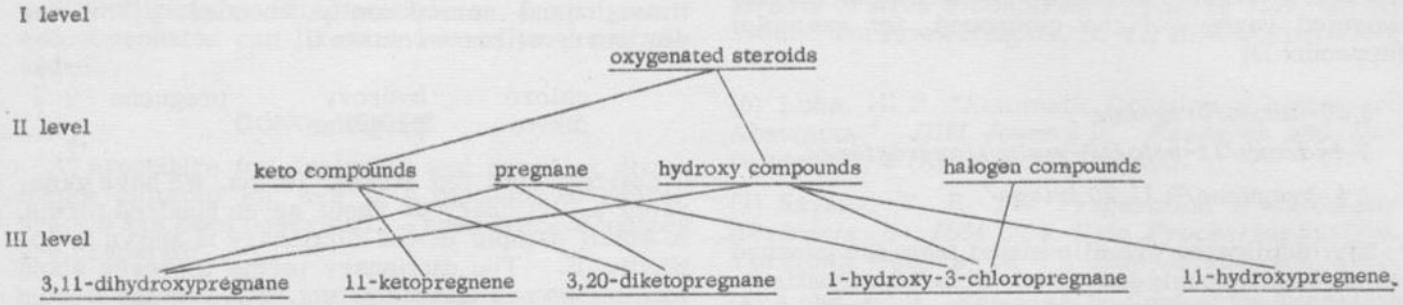
Products

As previously stated, two products result directly from this process. One of them is a printed

list of *all* the steroid compounds, excluding those disclosed only by a structural formula; the other is a set of punched cards containing the steroids with no duplicates.

Many other lists and classifications can also be printed by the machine without returning to the original document. Thus, if one desired steroid compounds which are esters, or all those which have keto groups or which have molecular weight above 500, a highly skilled chemist can take such subject headings and, consulting the addresses of the terms extracted from the documents, list all such terms which belong under each heading. The machine can then print these listings by a slight modification of the program used in (9).

Since the smallest unit of information, the compound, was extracted from the document, reference to the document for indexing, encoding, or separation of the compounds into various categories or classifications can be handled without having to refer to the original document. Thus the chart below illustrates some of the possibilities.



Level III is essentially the original index. Levels I and II can be created in a manner similar to the subject matter headings by one skilled person. It is now easy to prepare other lists such as:

keto steroids

11-ketopregnene (Pat. #)
3,20-diketopregnane (Pat. #)

hydroxy steroids

3,11-dihydroxypregnane (Pat. #)
1-hydroxy-3-chloropregnane (Pat. #)
11-hydroxypregnene (Pat. #)

halogen compounds

1-hydroxy-3-chloropregnane (Pat. #)

pregnane compound

3,11-dihydroxypregnane (Pat. #)
3,20-diketopregnane (Pat. #)
11-hydroxypregnane (Pat. #)
1-hydroxy-3-chloropregnane (Pat. #)

oxygenated compounds

3,11-dihydroxypregnane (Pat. #)
3,20-diketopregnane (Pat. #)
11-hydroxypregnene
11-hydroxy-3-chloropregnane (Pat. #)
11-ketopregnane (Pat. #)

Thus it can be seen that almost any type of index or subject heading list can be prepared without reference to the original document or without the need of having every extractor know the classification scheme.

A typical list of compounds generated by this method appears in the appendix (C).

Resin Example

The methods previously described can be used additionally to extract concepts or relationships. The following abbreviated example may be applicable to the polymers for ultimate use in searching them (3,9).

Determination of the point of view and the depth of indexing resulted in the following rule:

Extract *all* compounds in a patent which are *catalysts, solvents, monomers*. In addition to the underlining, the personnel were instructed to write adjacent thereto the number (1) for catalyst, (2) for solvent and (3) for monomer. If the number of kinds of different concepts or relationships is kept small, the extractors work satisfactorily. A specific example is shown in the appendix (C). Indexes can easily be printed for all catalysts, all solvents, all monomers, or all compounds in the patent regardless of their function. See appendix C.

Phosphate Example

There are seven people now extracting information by this method from the 1,765 phosphorus ester patents in Class 260 subclass 461.

The personnel are instructed to underline all phosphorus compounds. This results in having both generic and specific compounds extracted. The resulting list is an index to all the phosphorus compounds in a patent, excluding structural formulae. See appendix A. A list can be produced containing all the phosphorus compounds in all the 1,765 patents, and under each compound listing the patents in which they occur, for example:

Trimethyl phosphate	Tributyl phosphate	Tricresyl phosphate
Pat. No.	Pat. No.	Pat. No.
Pat. No.	Pat. No.	Pat. No.
Pat. No.	Pat. No.	Pat. No.

No attempt is made to eliminate or correlate synonyms since the search system provides means to allow such redundancy. The index will be quite accurate because of the redundancy of extraction, important compounds usually appearing several times in a document.

AUTOMATIC ENCODING

From the punched cards so created the document can be automatically encoded. A detailed illustration of the coding of the steroids will now be described. Although not described in this paper, the phosphates and polymers can be semi-automatically or automatically encoded by using similar procedures and techniques.

A slight modification of the first step (extracting) requires that the compound not only be underlined but that a virgule be placed to separate the various selected parts of the compound, for example: (appendix D)

3,20-diketo/pregnane/.

3-hydroxy/11-keto/17-methyl/androstane/.

Δ^4-pregnene/3,11,20 trione/.

The duplicates are eliminated from the punched cards as previously described. Before automatically encoding, a dictionary must be generated.

Dictionary Generation

Two methods are available for generating a dictionary. In the first method one attempts to list all the possible entries. There are two serious defects to this method, (1) it is impractical and time consuming to think of all the entries, and (2) some of the entries never used must still be stored at no small cost. The second method is to use the extracts from the documents in a suitable computer program to generate the dictionary. This was experimentally and successfully tried with the steroids. Extracts from the first document were put into the machine, for example,

3,20-diketo/pregnane/.
3-hydroxy/pregnane/.
11-hydroxy/pregnene/.

From these, the program generated the following terms which were alphabetized and listed to form a dictionary:

diketo	pregnane
hydroxy	pregnene

The extracts from the next patent contained the following compounds:

3-hydroxy/pregnene/.
3,20-diketo/pregnane/.
11-hydroxy/pregnane/.
3-hydroxy/7-chloro/pregnane/.

The first three compounds were automatically processed since the dictionary already contained the terms:

diketo	pregnane
hydroxy	pregnene

and these compounds contain only dictionary terms even though different from the compounds in the first patent. However, when the fourth compound is processed, the machine recognizes that there is no *chloro* in the dictionary and it prints out on the typewriter *"Chloro—not in File."* and automatically punches out a card which contains the whole compound. At the end of the run, all the terms which the typewriter indicated "not in file" are added to the dictionary. Now the "not in file" compounds which were not processed are sent through and automatically encoded. The new dictionary will now contain

chloro	hydroxy	pregnene
diketo	pregnane	

Starting with ten steroid terms, we have generated a dictionary of about seven hundred terms. A small sample of the dictionary is shown in appendix E. The dictionary terms increase about four terms per steroid patent. The average steroid patent contains about fifty compounds. It is hoped that the number of terms added will materially decrease as the size of the dictionary increases. Trivial names such as "progesterone," if listed, are, of course, added to the dictionary. The dictionary is therefore self generating and every term therein has been used at least once. (See appendix E.)

Encoding

The punched cards generated by the dictionary are fed into the machine. For example, for

3,20-diketo/11-hydroxy/pregnene/.

the machine recognizes the first portion indicated by the / and looks up diketo in the dictionary. In the dictionary the codes for diketo are stored with the term, provisions for handling the numbers (e.g. 3, 20) having already been made. The codes are then sent to the output of the computer. The machine then recognizes and processes "11 hydroxy" and finally "pregnene." The period "." indicates the end of the compound. The codes which have been sent to output are then either printed if there is no suitable punch available or, if one is available, the codes are punched in a punched card. It does not matter how a compound is named in the document. It will always receive the same code. For example:

3,20-diketo/Δ^4-pregnene/.
Δ^4-pregnene/3,20-dione/.
progesterone/.

are all different methods of expressing the same compound, and will each receive the same codes. The machine in the first instance processes diketo and then pregnene; in the second instance it processes pregnene and dione; and, in the third it

merely puts out the code for progesterone. The dictionary can thus contain both easily recognized parts of compounds as well as their trivial names. The code for dione and diketo in the dictionary are the same. The code for progesterone, listed in the dictionary, is the same as the combined listings of (3,20-diketo) or (3,20-dione) and Δ4-pregnene. A flow chart of the program for encoding is appendixed (F).

The cards, or the listed codes are suitable for use in the steroid searching system (1).

With slightly modified procedures, the polymers and phosphates can likewise be analyzed and encoded.

CONCLUSIONS

A procedure for analyzing and encoding documents suitable for use in a mechanized search system has been described. Three types of indices are prepared:

1. Index of terms, in the language of the document, alphabetically listed for each document
2. Index of terms, in the language of the documents, of all the documents in the file, alphabetically arranged
3. Index of terms under preassigned headings

A procedure for generating a dictionary and encoding compounds has been disclosed. The methods discussed appear to offer a promising resolution to the problem of file preparation. There are several apparent advantages of this method. These are that (a) the bulk of the extraction can be performed by the use of personnel having less formal education, (b) on-the-job training can be short, (c) an accurate and reliable file is constructed, and (d) cost of file preparation is reasonable.

REFERENCES

(1) Frome, Julius, and Jacob Leibowitz. *A Punched Card System for Searching Steroid Compounds.* Patent Office Research and Development Report #7. Washington 25, D. C., Department of Commerce, 1957.

(2) Leibowitz, J., J. Frome, and D. D. Andrews, "Variable Scope Search System: VS_3." *Preprints of papers for the International Conference on Scientific Information.* Washington, D. C., National Academy of Sciences--National Research Council, 1958. Area V, pp. 291-316.

(3) Frome, Julius, Jacob Leibowitz, and Don D. Andrews. *A System of Retrieval—Compounds, Compositions, Processes, and Polymers.* Patent Office Research and Development Report #13, Washington 25, D. C., Department of Commerce, 1958.

(4) Luhn, H. P. "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," *IBM Journal of Research and Development, 4* (Oct. 1957) 309-317.

(5) *An Experiment in Auto-Abstracting—Auto Abstracts of Area 5 Conference Papers.* Nov. 16-21, 1958. Yorktown Heights, N. Y., IBM Corporation, 1958.

(6) Luhn, H. P. "Automatic Creation of Literature Abstracts," *IBM Journal of Research and Development, 2* (April 1958) 159-165.

(7) Savage, T. R. *The Preparation of Automatic Abstracts on IBM 704 Data Processing System.* Yorktown Heights, N. Y., IBM Corporation, 1958.

(8) Luhn, H. P. *Keyword-in-Context for Technical Literature* Yorktown Heights, N. Y., IBM Corporation, 1959.

(9) Leibowitz, Jacob, Julius Frome, and Don D. Andrews, *Variable Scope Patent Searching by an Inverted File Technique.* Patent Office Research and Development Report #14, Washington 25, D. C., Department of Commerce, 1958.

(10) Schultz, C. K., and J. J. O'Connor, "Designing More Efficient Indexes." Submitted for publication *UNESCO Bulletin,* Nov.-Dec., 1959.

(11) MacMillan, J. T., and Dr. I. D. Welt. "Some Personnel Problems of a Small Indexing Project," delivered at ACS 136th meeting in Atlantic City, Sept. 14, 1959. (Unpublished.)

(12) Welt, I. D. "A Combined Indexing-Abstracting System," *Preprints of Papers for International Conference on Scientific Information.* Washington, D. C., National Academy of Sciences--National Research Council, 1958, Area 2, pp. 135-145.

(13) Newman, S. M. *Linguistic Problems in Mechanization of Patent Searching.* Patent Office Research and Development Report #9, Washington 25, D. C., Department of Commerce, 1957.

(14) Luhn, H. P. *Review of Information Retrieval Methods.* Research Report #RC-59. Yorktown Heights, N. Y., IBM Corporation, Oct. 1, 1958.

Appendix A

A partial listing of all the steroid compounds in patent 2,905,678. The two digit number is one less than the number of times the compound occurred in the patent.

2905678	6 ACETATE 4 PREGNENE	00
2905678	5A 6B DIHYDROXY 3 KETO PREGNANE	00
2905678	5A 6B DIHYDROXY 3 KETO PREGNANES	00
2905678	5A 6B DIHYDROXY PREGNANE	00
2905678	5A 6B DIHYDROXY PREGNANES	01
2905678	5A 6B DIHYDROXY 3 11 20 TRIKETO PREGNANE	00
2905678	5A 6B DIHYDROXY 3 11 20 TRIKETO PREGNANE 6 ACETATE	00
2905678	5A 6B DIHYDROXY 3 11 20 TRIKETO PREGNANE 6 FORMATE	00
2905678	5A HYDROXY 6 ACETOXY PREGNANE	00
2905678	5A HYDROXY 6 ACETOXY PREGNANES	00
2905678	5A HYDROXY 6 FORMOXY PREGNANE	00
2905678	5A HYDROXY 6B ACETOXY PREGNANE	00
2905678	5A HYDROXY 6B FORMOXY PREGNANE	00
2905678	5A HYDROXY 6B FORMOXY PREGNANES	01
2905678	5A 6B 17A TETRAHYDROXY 3 11 20 TRIKETO PREGNANE 6 FORMATE 21 ACETATE	00
2905678	5A 6B 17A 21 TETRAHYDROXY 3 11 20 TRIKETO PREGNANE 21 ACETATE	02
2905678	5A 6B 17A 21 TETRAHYDROXY 3 11 20 TRIKETOPREGNANE 21 ACETATE	01
2905678	5A 6B 17A TRIHYDROXY 3 11 20 TRIKETO PREGNANE	00
2905678	5A 6B 17A TRIHYDROXY 3 11 20 TRIKETO PREGNANE 6 FORMATE	03
2905678	5A 6B 21 TRIHYDROXY 3 11 20 TRIKETO PREGNANE 6 FORMATE 21 ACETATE	03
2905678	5A 6B 21 TRIHYDROXY 3 11 20 TRIKETO PREGNANE 6 FORMATE 21 BENZOATE	02
2905678	5A 6B 21 TRIHYDROXY 3 11 20 TRIKETO PREGNANE 6 FORMATE 21 BUTYRATE	00
2905678	5A 6B 21 TRIHYDROXY 3 11 20 TRIKETO PREGNANE 6 FORMATE 21 CAPROATE	00
2905678	5A 6B 21 TRIHYDROXY 3 11 20 TRIKETO PREGNANE 6 FORMATE 21 HEMISUCCINAT	00
2905678	$E	00
2905678	6B 21 DIHYDROXY 3 11 20 TRIKETO 4 PREGNENE 6 21 DIACETATE	01
2905678	6B 17A DIHYDROXY 3 11 20 TRIKETO 4 PREGNENE 6 FORMATE	05
2905678	6B 21 DIHYDROXY 3 11 20 TRIKETO 4 PREGNENE 6 FORMATE 21 ACETATE	09
2905678	6B 21 DIHYDROXY 3 11 20 TRIKETO 4 PREGNENE 6 FORMATE 21 BENZOATE	03
2905678	6B 21 DIHYDROXY 3 11 20 TRIKETO 4 PREGNENE 6 FORMATE 21 BUTYRATE	01
2905678	6B 21 DIHYDROXY 3 11 20 TRIKETO 4 PREGNENE 6 FORMATE 21 CAPROATE	00
2905678	6B 21 DIHYDROXY 3 11 20 TRIKETO 4 PREGNENE 6 FORMATE 21 ESTERS	00
2905678	6B 21 DIHYDROXY 3 11 20 TRIKETO 4 PREGNENE 6 FORMATE 21 HEMISUCCINATE	01

2905678	6B 21 DIHYDROXY 3 11 20 TRIKETO 4 PREGNENE 6 FORMATE 21 &HENYLACETATE	00
2905678	6B 32	00
2905678	6B 21 DIHYDROXY 3 11 20 TRIKETO 4 PREGNENE 6 FORMATE 21 ISOVALERATE	01
2905678	6B 21 DIHYDROXY 3 11 20 TRIKETO 4 PREGNENE 6 FORMATE 21 PHENYLACETATE	00
2905678	6B 21 DIHYDROXY 3 11 20 TRIKETO 4 PREGNENE 6 FORMATE 21 PROPIONATE	01
2905678	6B 21 DIHYDROXY 3 11 20 TRIKETO 4 PREGNENE 6 FORMATE 21 T BUTYLACETATE	01
2905678	6B 21 DIHYDROXY 3 11 20 TRIKETO 4 PREGNENE 6 FORMATE 21 TRICARBALLYLAT	00
2905678	♦E	00
2905678	6B 21 DIHYDROXY 3 11 20 TRIKETO 4 PREGNENE 6 FORMATE 21 VALERATE	01
2905678	6B 21 DIHYDROXY 3 11 20 TRIKETO 4 PREGNENE 21 A	00
2905678	6B 21 DIHYDROXY 3 11 20 TRIKETO 4 PREGNENE 21 ACETATE	06
2905678	6B 21 DIHYDROXY 3 11 20 TRIKETO 4 PREGNENE 21 BENZOATE	04
2905678	6B 21 DIHYDROXY 3 11 20 TRIKETO 4 PREGNENE 21 T BUTYLACETATE	00
2905678	6B 21 DIHYDROXY 3 11 20 TRIKETO 4 PREGNENE 21 T BUTYL ACETATE	01
2905678	6B 21 DIHYDROXY 3 11 20 TRIKETO 4 PREGNENE 21 BUTYRATE	02
2905678	6B 21 DIHYDROXY 3 11 20 TRIKETO 4 PREGNENE 21 HEMISUCCINATE	02
2905678	6B 21 DIHYDROXY 3 11 20 TRIKETO 4 PREGNENN 21 PHENYLACETATE	00

The following patent and list is an example of the underlining and listing in a phosphate patent.

United States Patent Office

2,870,190

Patented Jan. 20, 1959

2,870,190

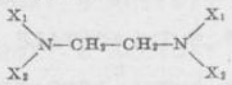
PHOSPHONATE COMPOUNDS

Bill E. Burgert and Henry Tolkmith, Midland, Mich., assignors to The Dow Chemical Company, Midland, Mich., a corporation of Delaware

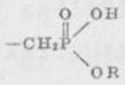
No Drawing. Application July 3, 1957
Serial No. 669,718

5 Claims. (Cl. 260—461)

The present invention relates to phosphonate compounds having the formula

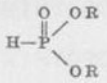


In this and succeeding formulas, X_1 represents X_2 or hydrogen and X_2 represents a radical having the formula



wherein R is a lower alkyl radical. The expression "lower alkyl" is employed in the present specification and claims to refer to the alkyl radicals containing from 1 to 4 carbon atoms, inclusive. These new compounds are very viscous liquids which are readily soluble in water and somewhat soluble in many organic solvents. They have been found to be useful as parasiticides and are adapted to be employed as active toxic constituents of dust and liquid compositions for the control of many insect and fungal pests such as *Alternaria solani, Fusarium solani, Pythium spp.* and *Rhizoctonia solani.*

The new compounds may be prepared by causing a reaction between ethylenediamine, formaldehyde and a dialkyl phosphite of the formula

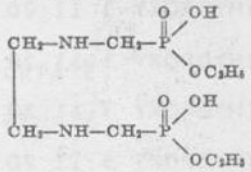


Where it is desired to introduce two $—CH_2PO(OH)(OR)$ groups into the molecule, one molecular proportion of ethylenediamine is employed with two molecular proportions of formaldehyde and two molecular proportions of dialkylphosphite. Where it is desired to introduce four $—CH_2PO(OH)(OR)$ groups into the molecule, one molecular proportion of ethylenediamine is employed with four molecular proportions of each of the formaldehyde and dialkyl phosphite reactants. Although the details of the reaction mechanism are not completely understood, the reaction takes place smoothly at the temperature range of from 0° to 50° C. with the formation of the desired product and alkanol of reaction. The reaction is somewhat exothermic and the temperature may be controlled by regulating the rate of contacting the reactants and by external cooling. Upon completion of the reaction, the reaction mixture is fractionally distilled under reduced pressure to separate low boiling constituents and obtain the desired product as a liquid residue.

In carrying out the reaction, a mixture of ethylenediamine and dialkyl phosphite is contacted portionwise with an aqueous solution of formaldehyde. This operation is carried out with stirring and at a temperature of from 0° to 50° C. Upon completion of the reaction, the reaction mixture is fractionally distilled under reduced pressure at gradually increasing temperatures up to a temperature of 150° C. to separate low boiling constituents and obtain the desired product as a liquid residue. Since the desired products are somewhat unstable at temperatures in excess of 150° C., exposure to such elevated temperatures for any appreciable period should be avoided.

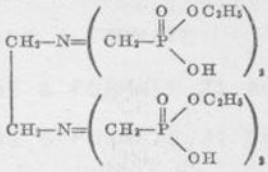
The following examples illustrate the invention but are not to be construed as limiting:

Example 1.—Ethylenediamine N,N'-bis(O-ethyl methanephosphonate)



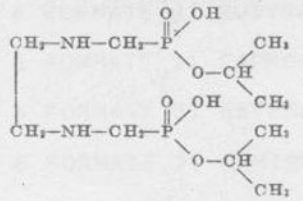
A 37 percent aqueous solution of formaldehyde containing 16.5 grams (0.55 mole) of formaldehyde was added portionwise with stirring to a mixture of 15 grams (0.25 mole) of ethylenediamine and 75.9 grams (0.55 mole) of O,O-diethylphosphite. The addition was carried out in 15 minutes and at a temperature of from 25° to 35° C. Following the addition, stirring was continued for 40 minutes, and the reaction mixture thereafter diluted with 70 grams of methylcyclohexane. The resulting mixture was fractionally distilled under reduced pressure at temperatures gradually increasing up to 150° C. to remove low boiling constituents and obtain an ethylenediamine N,N'-bis(O-ethylmethanephosphonate) product as a dark brown, very viscous, liquid residue having a density of 1.4 at 25° C. This product was found to contain 9.27 percent nitrogen, 20.50 percent phosphorus and 7.09 percent hydrogen compared to the calculated values 9.21, 20.40 and 7.29 percent, respectively.

Example 2.—Ethylenediamine N,N,N',N'-tetra (O-ethyl methanephosphonate)



A 37 percent aqueous solution of formaldehyde containing 24 grams (0.80 mole) of formaldehyde was added portionwise with stirring to a mixture of 9 grams (0.15 mole) of ethylenediamine and 109.1 grams (0.80 mole) of O,O-diethyl phosphite. The addition was carried out in 15 minutes and at a temperature of from 20° to 30° C. Following the addition, stirring was continued for 90 minutes to complete the reaction. The reaction mixture was then processed as described in Example 1 to obtain an ethylenediamine N,N,N',N'-tetra(O-ethyl methanephosphonate) product as a light tan, viscous, liquid residue having a density of 1.2 at 25° C. This product contained 5.49 percent nitrogen and 20.86 percent phosphorus compared to the theoretical values of 5.11 percent and 22.16 percent, respectively.

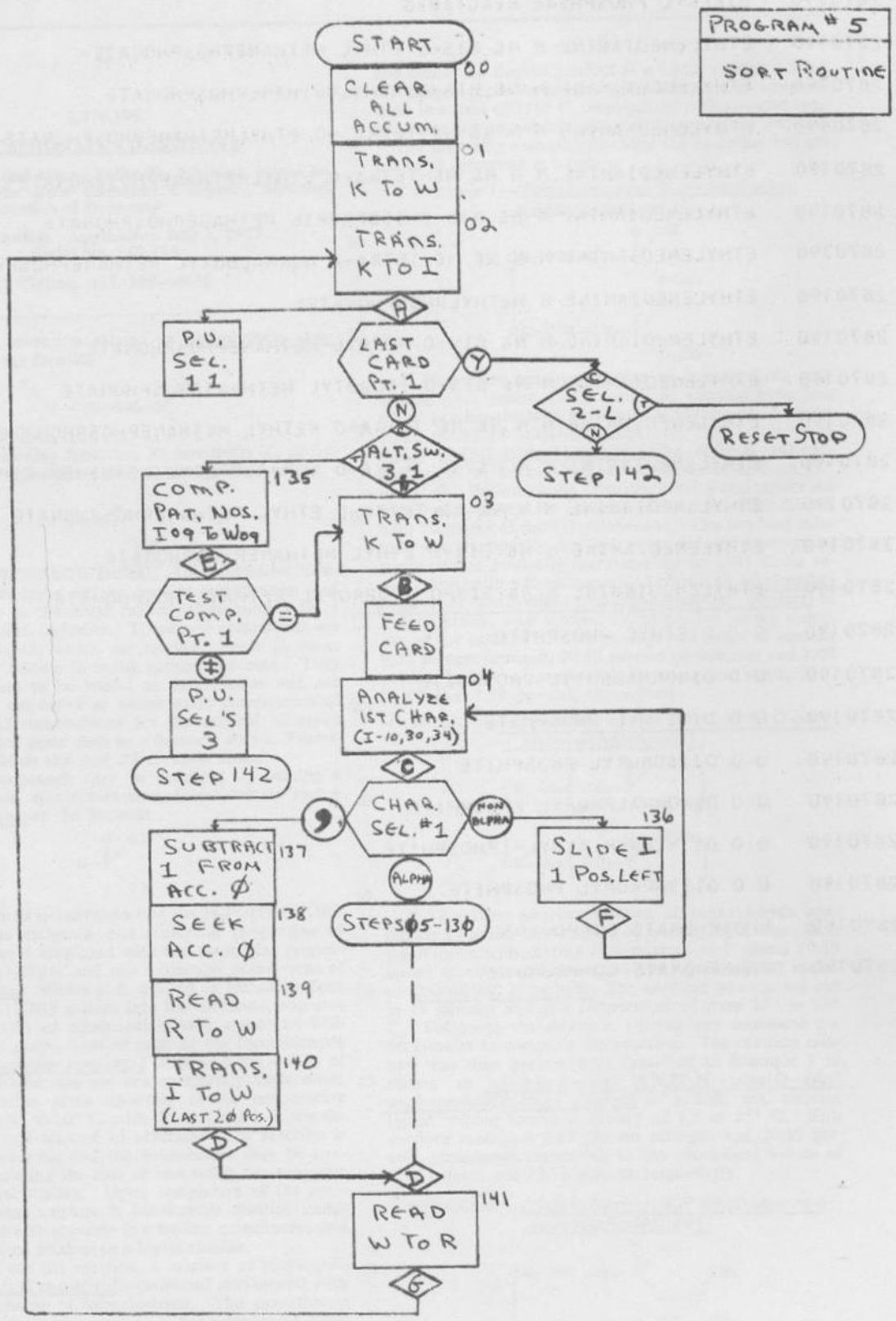
Example 3.—Ethylenediamine N,N'-bis(O-isopropyl methanephosphonate)



2870190	DIALKYL PHOSPHITE	01
2870190	DIALKYLPHOSPHITE	00
2870190	DIALKYL PHOSPHITE REACTANTS	00
2870190	ETHYLENEDIAMINE N N& BIS-O ETHYL METHANEPHOSPHONATE-	00
2870190	ETHYLENEDIAMINE N N& BIS-O ETHYLMETHANEPHOSPHONATE	00
2870190	ETHYLENEDIAMINE N N N& N& TETRA -O ETHYLMETHANEPHOSPHONATE	00
2870190	ETHYLENEDIAMINE N N N& N& TETRA-O ETHYLMETHANEPHOSPHONATE- PRODUCT	00
2870190	ETHYLENEDIAMINE N N& BIS-O ISOPROPYL METHANEPHOSPHONATE	00
2870190	ETHYLENEDIAMINE N N N& N& TETRA-O NORMALBUTYL METHANEPHOSPHONATE	02
2870190	ETHYLENEDIAMINE N METHYLPHOSPHONATES	00
2870190	ETHYLENEDIAMINE N N& BIS-O METHYL METHANEPHOSPHONATE	00
2870190	ETHYLENEDIAMINE N N& BIS-O ISOBUTYL METHANEPHOSPHONATE	00
2870190	ETHYLENEDIAMINE N N N& N& TETRA-O METHYL METHANEPHOSPHONATE-	00
2870190	ETHYLENEDIAMINE N N N& N& TETRA-O NORMALPROPYL METHANEPHOSPHONATE	00
2870190	ETHYLENEDIAMINE N N N& N& TETRA-O ETHYL METHANEPHOSPHONATE	01
2870190	ETHYLENEDIAMINE N N& BIS-O ETHYL METHANEPHOSPHONATE	00
2870190	ETHYLENEDIAMINE N N& BIS-O ISOPROPYL METHANEPHOSPHONATE	00
2870190	O O DIETHYL PHOSPHITE	00
2870190	O O DINORMALBUTYL PHOSPHITE	00
2870190	O O DIMETHYL PHOSPHITE	00
2870190	O O DIISOBUTYL PHOSPHITE	00
2870190	O O DINORMALPROPYL PHOSPHITE	00
2870190	O O DI -LOWER ALKYL- PHOSPHITE	00
2870190	O O DIISOPROPYL PHOSPHITE	00
2870190	PHOSPHONATE COMPOUNDS	00
2870190	PHOSPHONATE COMPOUND	00

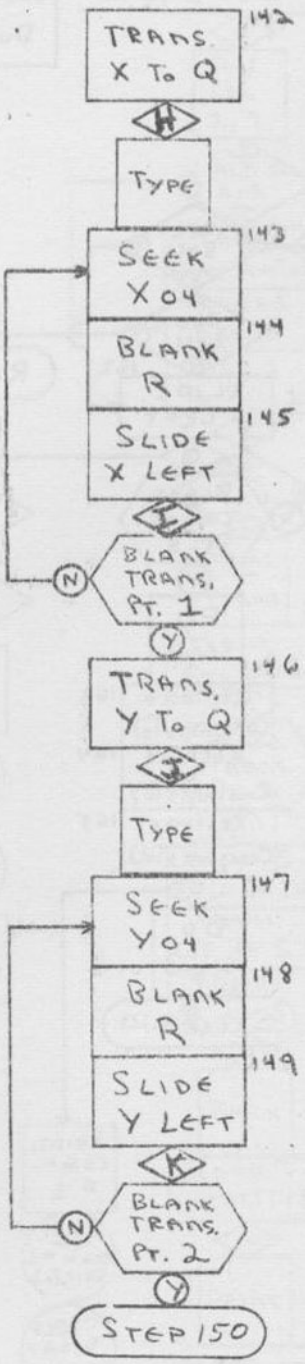
Appendix B

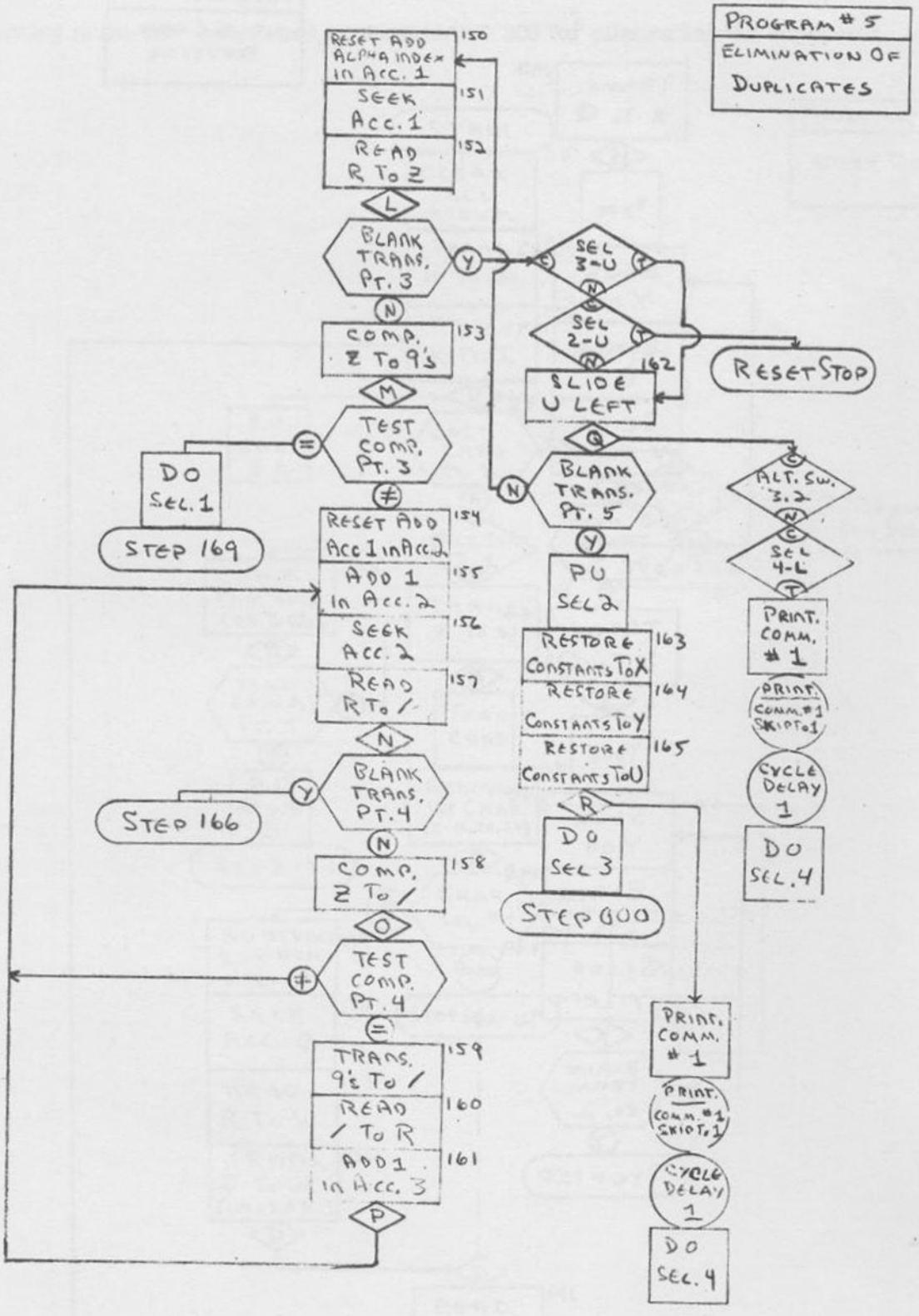
The following is the flow chart used on the RAMAC 305 for eliminating the duplicates.

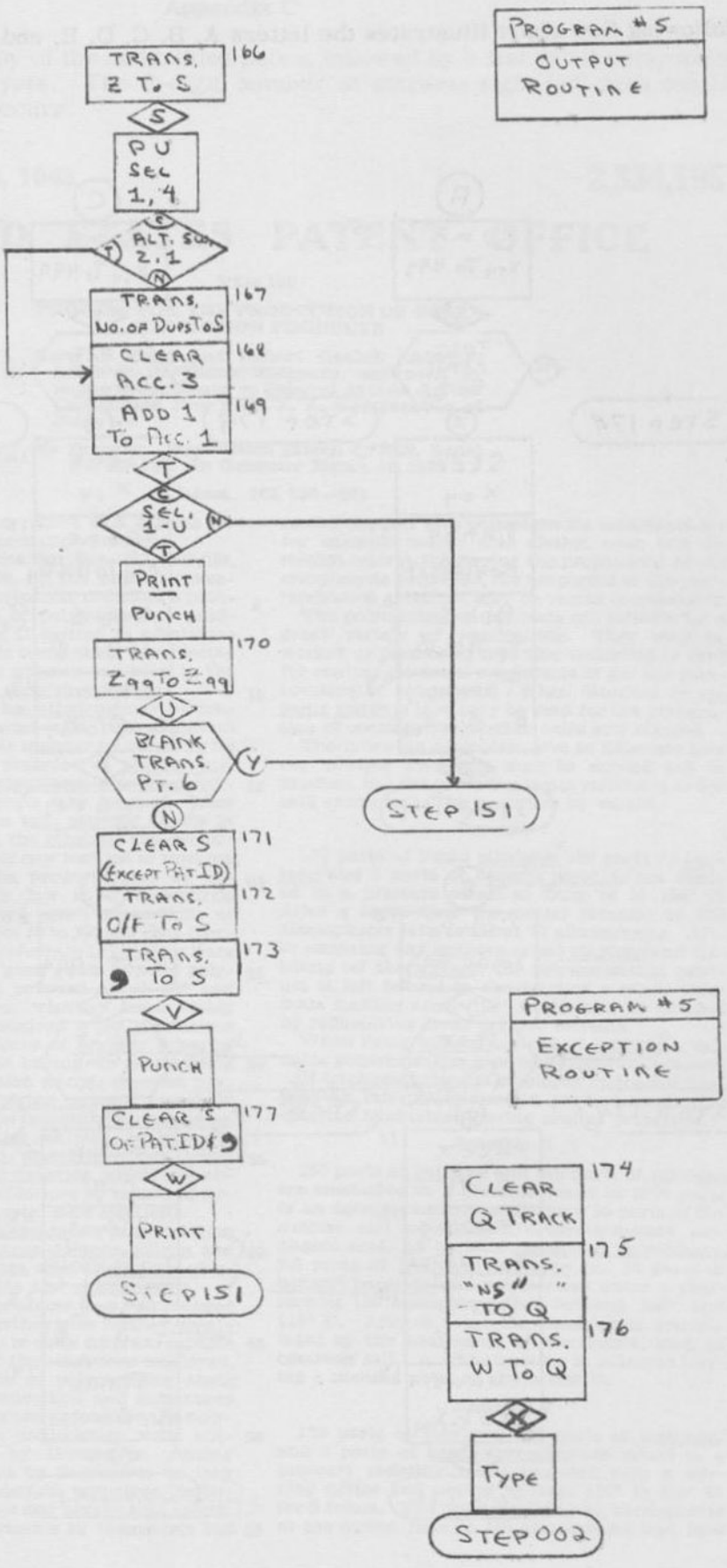


PROGRAM #5

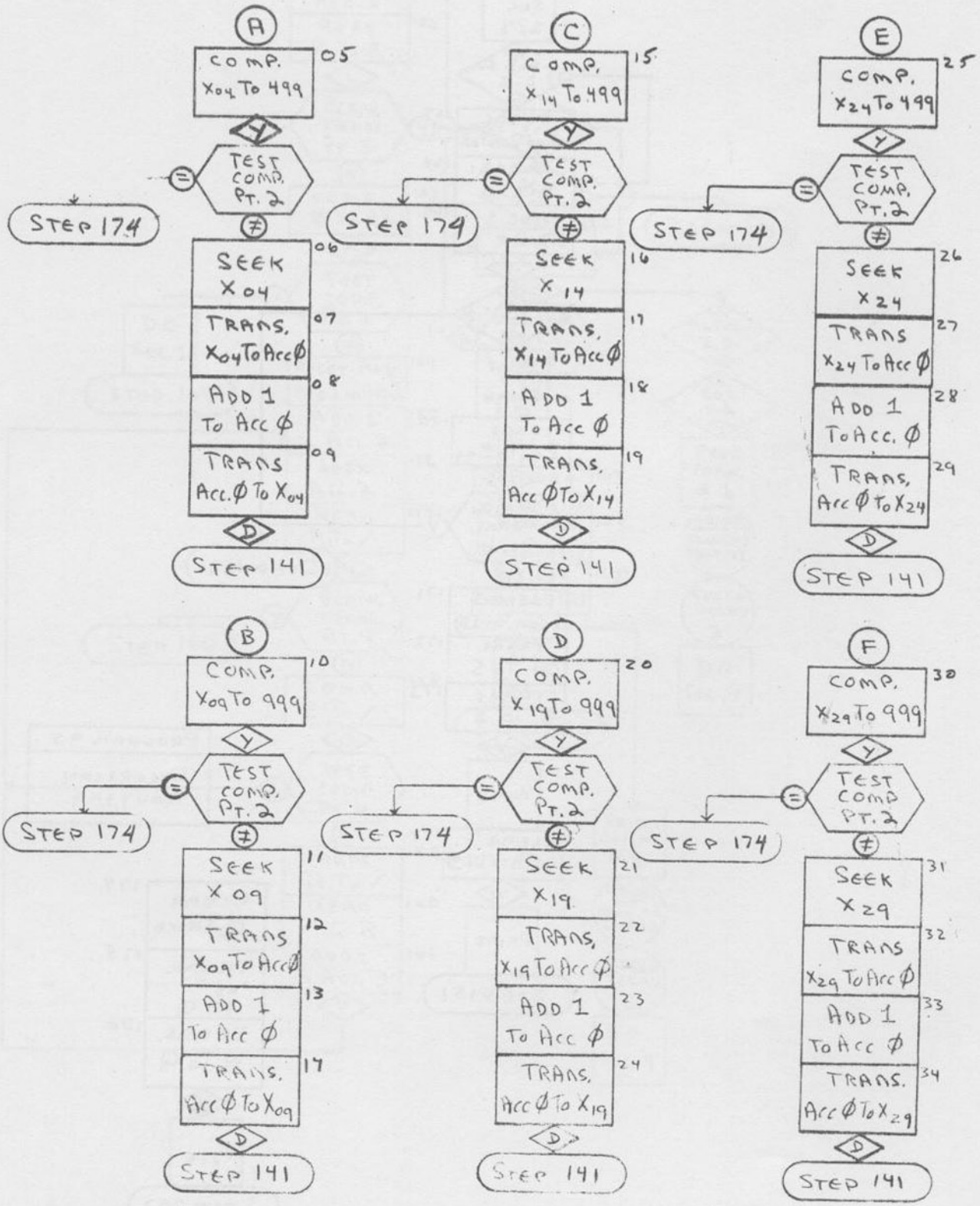
LAST CARD ROUTINE




PROGRAM # 5
ELIMINATION OF DUPLICATES
RESET ADD ALPHA INDEX IN ACC. 1
150
SEEK ACC. 1
151
READ R TO Z
152
L
BLANK TRANS. PT. 3
Y
N
COMP. Z TO 9's
153
M
TEST COMP. PT. 3
=
DO SEL. 1
STEP 169
≠
RESET ADD ACC 1 IN ACC 2
154
ADD 1 IN ACC. 2
155
SEEK ACC. 2
156
READ R TO /
157
N
BLANK TRANS. PT. 4
Y
STEP 166
N
COMP. Z TO /
158
O
TEST COMP. PT. 4
≠
=
TRANS. 9's TO /
159
READ / TO R
160
ADD 1 IN ACC. 3
161
P
SEL 3-U
SEL 2-U
RESET STOP
SLIDE U LEFT
162
Q
BLANK TRANS. PT. 5
N
Y
PU SEL 2
RESTORE CONSTANTS TO X
163
RESTORE CONSTANTS TO Y
164
RESTORE CONSTANTS TO U
165
R
DO SEL 3
STEP 000
ALT. SW. 3.2
SEL 4-L
PRINT. COMM. # 1
PRINT. COMM. # 1 SKIP T. 1
CYCLE DELAY 1
DO SEL. 4
PRINT. COMM. # 1
PRINT. COMM. # 1 SKIP T. 1
CYCLE DELAY 1
DO SEL. 4


PROGRAM #5
OUTPUT ROUTINE
TRANS. Z TO S
166
S
PU SEL 1, 4
ALT. SUB. 2.1
T
N
TRANS. NO. OF DUPS TO S
167
CLEAR ACC. 3
168
ADD 1 TO ACC. 1
169
T
SEL. 1-U
N
T
PRINT PUNCH
TRANS. Z99 TO Z99
170
U
BLANK TRANS. PT. 6
Y
N
STEP 151
CLEAR S (EXCEPT PAT ID)
171
TRANS. C/F TO S
172
TRANS. , TO S
173
V
PUNCH
CLEAR S OF PAT ID & ,
177
W
PRINT
STEP 151
PROGRAM #5
EXCEPTION ROUTINE
CLEAR Q TRACK
174
TRANS. "NS" " TO Q
175
TRANS. W TO Q
176
X
TYPE
STEP 002

The following flow chart illustrates the letters A, B, C, D, E, and F as examples.



Appendix C

This appendix contains a copy of the underlined patent followed by a list of all compounds which are solvents, monomers, and catalysts. The 2-digit number at extreme right indicates one less than the number of times the compound occurs.

Patented Nov. 16, 1943 2,334,195

UNITED STATES PATENT OFFICE

2,334,195

PROCESS FOR THE PRODUCTION OF POLYMERIZATION PRODUCTS

Heinrich Hopff and Siebert Goebel, Ludwigshafen-on-the-Rhine, Germany, assignors, by mesne assignments, to General Aniline & Film Corporation, New York, N. Y., a corporation of Delaware

No Drawing. Application March 6, 1940, Serial No. 322,547. In Germany March 14, 1939

5 Claims. (Cl. 260—80)

The present invention relates to a process for the production of polymerization products.

The copending application Ser. No. 309,238, filed December 14, 1939, by the present inventors and Curt W. Rautentrauch, describes a process for the production of polymerization products in which ethylene, if desired in admixture with other polymerizable compounds, is subjected to polymerization in an aqueous emulsion in the presence of substances supplying oxygen.

We have now found that ethylene may be converted into solid or semi-solid polymerization products in a very simple manner by effecting its polymerization in the presence of organic solvents and oxygen or substances supplying oxygen. The organic solvents may be used either alone or in combination with aqueous liquids in which the solvents and the ethylene are emulsified or dissolved. By our new method of working the solid polymerization products may be obtained under relatively low superatmospheric pressures. Comparatively small proportions of organic solvents, say from 10 to 100 per cent (percentage by weight with reference to ethylene) are sufficient for obtaining good yields of solid polymerization products at between about 100 and about 300 atmospheres, whereas considerably higher pressures are required when conducting the process in the absence of organic solvents.

Among the substances having an accelerating effect in the polymerization we may mention oxygen or substances supplying oxygen, especially peroxidic substances, for example persulphates, percarbonates, perborates as well as benzoylperoxide, peracetic acid, diacetylperoxide, toluic acid peroxide and oleic peroxide which are obtainable from the acid chlorides by means of hydrogen peroxide and caustic soda solution.

As suitable organic solvents we may mention by way of example benzene, toluene, xylene and halogenated hydrocarbons, such as ethylene chloride, carbon tetrachloride and chlorbenzene.

Interpolymerization products may be produced by treating ethylene together with further unsaturated substances which in their turn are capable of polymerization under the conditions employed, i. e. substances capable of polymerizing alone under the conditions concerned and substances which polymerize only when exposed to the polymerizing conditions in conjunction with substances polymerizable by themselves. Among substances polymerizable by themselves we may mention by way of example propylene, isobutylene, butadiene, styrene and acrylic acid esters; compounds not polymerizable by themselves but in conjunction with polymerizable substances are for example maleic acid diethyl ester and dimethyl ester. By varying the proportions of the components employed, the properties of the polymerization products may be varied considerably.

The polymerization products are suitable for a great variety of applications. They may be worked as plastics by injection-moulding or used for coating electrical conductors or for the manufacture of condensers. When dissolved in organic solvents they may be used for the preparation of coatings resistant to acids and alkalies.

The following examples serve to illustrate how our present invention may be carried out in practice, but the invention is not restricted to the said examples. The parts are by weight.

Example 1

100 parts of liquid ethylene, 100 parts of benzene and 2 parts of benzoyl peroxide are heated in a pressure vessel at from 80 to 100° C. After a short time the initial pressure of 100 atmospheres falls to about 70 atmospheres. After removing any unpolymerized ethylene and distilling off the solvent, the polymerization product is left behind in the form of a white waxy mass melting above 100° which may be purified by redissolving from organic solvents.

When using toluene instead of benzene a valuable polymerization product is likewise obtained.

If trichlorethylene is employed instead of benzene an interpolymerization product containing chlorine is obtained having similar properties.

Example 2

250 parts of benzene and 500 parts of ethylene are emulsified in a pressure vessel in 1000 parts of an aqueous solution containing 20 parts of the sodium salt of alpha-hydroxyoctodecane sulphonic acid, 1.5 parts of potassium persulphate, 7.5 parts of hydrogen peroxide and 10 parts of benzoyl peroxide and polymerized under a pressure of 150 atmospheres at between 100° and 110° C. After 2 hours the emulsion is precipitated by the addition of an electrolyte, such as common salt. A white powder is obtained having a melting point of above 100° C.

Example 3

100 parts of ethylene, 200 parts of methanol and 2 parts of benzoylperoxide are mixed in a pressure resisting vessel provided with a stirring device and heated to from 110° to 120° C. for 2 hours. The pressure, being 80 atmospheres at the outset, rises to 160 atmospheres and falls

2334195	ACRYLIC ACID ESTERS-3	00
2334195	ACRYLIC ACID ETHYL ESTER-3	00
2334195	BENZOYL PEROXIDE-1	05
2334195	BENZENE-2	06
2334195	BUTADIENE-3	00
2334195	BENZOYLPEROXIDE-1	00
2334195	BUTANOL-2	01
2334195	CARBON TETRACHLORIDE-2	00
2334195	CHLORBENZENE-2	00
2334195	CHLORINATED ALIPHATIC HYDROCARBON-2	00
2334195	DIACETYLPEROXIDE-1	00
2334195	DICHLORETHYLENE-2	00
2334195	ETHYLENE-3	19
2334195	ETHYLENE CHLORIDE-2	00
2334195	HALOGENATED HYDROCARBONS-2	00
2334195	HYDROGEN PEROXIDE-1	01
2334195	ISOBUTYLENE-3	01
2334195	METHANOL-2	06
2334195	ORGANIC SOLVENTS-2	05
2334195	OXYGEN-1	03
2334195	OLEIC PEROXIDE-1	01
2334195	ORGANIC SOLVENT-2	01
2334195	PEROXIDIC SUBSTANCES-1	00
2334195	PERSULPHATES-1	00
2334195	PERCARBONATES-1	00
2334195	PERBORATES-1	00
2334195	PERACETIC ACID-1	00
2334195	PROPYLENE-3	00
2334195	POTASSIUM PERSULPHATE-1	01
2334195	PEROXIDE-1	00
2334195	SUBSTANCES SUPPLYING OXYGEN-1	04
2334195	SOLVENTS-2	00
2334195	STYRENE-3	01
2334195	SOLVENT-2	01
2334195	TOLUIC ACID PEROXIDE-1	00
2334195	TOLUENE-2	01

2334195	TRICHLORETHYLENE-2	00
2334195	UNSATURATED SUBSTANCES-3	00
2334195	UNSATURATED COMPOUND-3	00
2334195	XYLENE-2	00

The following is a list of all of the catalysts in 2,334,195. The 2-digit number plus one at the extreme right is the number of times the compound occurs in the patent.

2334195	BENZOYL PEROXIDE-1	05
2334195	BENZOYLPEROXIDE-1	00
2334195	DIACETYLPEROXIDE-1	00
2334195	HYDROGEN PEROXIDE-1	01
2334195	OXYGEN-1	03
2334195	OLEIC PEROXIDE-1	01
2334195	PEROXIDIC SUBSTANCES-1	00
2334195	PERSULPHATES-1	00
2334195	PERCARBONATES-1	00
2334195	PERBORATES-1	00
2334195	PERACETIC ACID-1	00
2334195	POTASSIUM PERSULPHATE-1	01
2334195	PEROXIDE-1	00
2334195	SUBSTANCES SUPPLYING OXYGEN-1	04
2334195	TOLUIC ACID PEROXIDE-1	00

The folowing is a list of all solvents in 2,334,195:

2334195	BENZENE-2	06
2334195	BUTANOL-2	01
2334195	CARBON TETRACHLORIDE-2	00
2334195	CHLORBENZENE-2	00
2334195	CHLORINATED ALIPHATIC HYDROCARBON-2	00
2334195	DICHLORETHYLENE-2	00
2334195	ETHYLENE CHLORIDE-2	00
2334195	HALOGENATED HYDROCARBONS-2	00
2334195	METHANOL-2	06
2334195	ORGANIC SOLVENTS-2	05
2334195	ORGANIC SOLVENT-2	01
2334195	SOLVENTS-2	00

2334195	SOLVENT-2	01
2334195	TOLUENE-2	01
2334195	TRICHLORETHYLENE-2	00
2334195	XYLENE-2	00

The following is a list of all the momomers in 2,334,195:

2334195	ACRYLIC ACID ESTERS-3	00
2334195	ACRYLIC ACID ETHYL ESTER-3	00
2334195	BUTADIENE-3	00
2334195	ETHYLENE-3	19
2334195	ISOBUTYLENE-3	01
2334195	PROPYLENE-3	00
2334195	STYRENE-3	01
2334195	UNSATURATED SUBSTANCES-3	00
2334195	UNSATURATED COMPOUND-3	00

Appendix D

A copy of a steroid patent underlined for automatic encoding and a list of the compounds extracted therefrom. Two digit number is one less than the number of times the compound occurs in the patent.

United States Patent Office

2,871,246

Patented Jan. 27, 1959

2,871,246

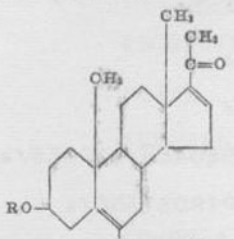
PROCESS FOR 3-HYDROXY-6-ALKYL-5,16-PREGNADIEN-20-ONES AND ESTERS THEREOF

Bjarte Löken, Roosevelt, Puerto Rico, assignor to G. D. Searle & Co., Chicago, Ill., a corporation of Delaware

No Drawing. Application May 28, 1958
Serial No. 738,305

5 Claims. (Cl. 260—397.4)

The instant invention relates to the preparation of 3β-hydroxy-6-alkyl-5,16-pregnadien-20-ones and their esters of the general structural formula



In the foregoing structural formula the lower alkyl radical in the 6-position can be a straight-chain or branched propyl, butyl, amyl or hexyl group, but it is preferably a methyl or ethyl radical. The radical R can represent hydrogen or an acyl radical of a hydrocarbon carboxylic acid, preferably of less than 7 carbon atoms. Examples of such acyl groups are formyl, acetyl, propionyl, butyryl, valeryl and hexanoyl.

The foregoing compounds are progestationally active. They also serve as starting materials for the preparation of 6-alkylprogesterones and 6-alkyl-17α- hydroxyprogesterones and their 17-esters which are potent progestational agents.

The procedures in the literature for conversion of diosgenin or cholesterol to pregnadiene derivatives all suffer from serious disadvantages in specific steps of the reaction sequence. Actually only the first conventional step of isomerizing a 6-alkyldiosgenin to its corresponding pseudoderivative, 3β,26-diacetoxy-6-alkyl-5,20(22)-furostadiene can be conducted by a procedure analogous to that generally employed for diosgenin namely reaction with acetic anhydride for 6 hours at 190–195° C. The subsequent steps of the sequence had to be altered materially and in some steps, as in cleavage of the δ-acetoxy-γ-methylpentanoyl ester, entirely different procedures had to be devised.

Thus, the second step, namely oxidation of the furostadiene to the 16-(δ-acetoxy-γ-methyl)pentanoate by chromium trioxide in an acetic acid medium is advantageously effected in the presence of a chloralkane solvent amounting to 15–60% by volume of the oxidation medium. The chlorinated solvents contemplated are the chlorinated hydrocarbons of one to three carbon atoms including, for example, ethylene chloride, chloroform, carbon tetrachloride, ethylene trichloride and propylene chloride. Effecting the oxidation in this manner facilitates recovery of the steroid from the reaction mixture. Thus the chromium salts and acetic acid are readily removed in aqueous phase while the 3-acyloxy-16-(δ-acetoxy-γ-methylpentanoyloxy)-5-pregnen-20-one / stays in the chlorinated hydrocarbon organic phase; this ester

can then be recovered by evaporating the chlorinated hydrocarbon.

The subsequent cleavage is effected by a novel solvolysis with at least three equivalents of alkali in aqueous acetone. A rapid reaction occurs with very strong alkali, e. g. caustic soda and potash. More time is needed with weaker bases such as the alkali metal carbonates. The acetone-water ratio can vary from about 1:3 to 3:1. The alkali, e. g. sodium hydroxide, can constitute from 1–20% by weight of the alkaline medium. The time period for the treatment is relatively short, ranging from about 5–60 minutes. Treatment temperatures are in the range of 20° C. to about 65° C. Temperature is relatively critical because two different results are attainable. In the lower range of about 20–33° C., cleavage of the substituted pentanoate results in a high yield of 3β-acyloxy-6-alkyl-5,16-pregnadien-20-one. On the other hand, in the range of about 45–65° C. there also occurs saponification of the 3-acylate to yield 3β-hydroxy-6-alkyl-5,16-pregnadien-20-one.

It is within the contemplation of the instant invention to carry out the same alkaline acetone-water treatment further at higher temperature to hydrolyze 3β-acyloxy-6-alkyl-5,16-pregnadien-20-one to 3-hydroxy-6-alkyl-5,16-pregnadien-20-one.

These compounds are valuable as key intermediates in the overall conversion of diosgenin into medically active 6-alkyl steroids. The 3β-hydroxy-6-alkyl-5,16-pregnadien-20-ones have certain advantages over their 3-esters as intermediates. Epoxidation to 3β-hydroxy-6-alkyl-16α,17α-epoxy-5-pregnen-20-ones can be effected by reaction with alkaline hydrogen peroxide in methanol solution to give a better yield. Also the 3-hydroxy compounds can be converted by the Oppenauer oxidation to the 6-alkyl-5,16-pregnadien-3,20-diones then by Raney nickel reduction to progestationally active 6-alkylprogesterones or epoxidation as above outlined to the 16α,17α-epoxides, followed by conversion to the 17α-hydroxy compounds according to conventional practices. The resulting esters of 17-hydroxyprogesterone are highly potent progestational agents. It has also been found that the 3-hydroxy-6-alkyl-5,16-pregnadien-20-ones are progestational agents.

The following examples describe the invention in further detail but are not to be construed as limiting in spirit or in scope. Quantities are given in parts by weight.

Example 1

To an ice-cooled solution of 100 parts of diosgenin in 1340 parts of dichloromethane there is added a solution of 160 parts of acetone, 270 parts of dichloromethane, 2 parts of sodium acetate and 75 parts of peroxyacetic acid. After completion of the addition, the mixture is left at room temperature for 5 hours and then washed with 1000 parts of a 5% ferrous sulfate solution which serves to reduce the excess peroxyacetic acid. The organic layer is separated and carefully washed with water. The washings are reextracted with dichloromethane and the combined dichloromethane extracts are taken to dryness on a steam bath. Then 900 parts of heptane are added and a small head fraction is distilled off in order to remove residual dichloromethane. The mixture is cooled to 30° C. under agitation and the crystalline material is collected on a filter. This material consists of a mixture of the α- and β-epoxides; the optical rotation α_D of −120° in chloroform solution indicates predominance of the α-epoxide. The mother liquor is evaporated to a small volume and a second crop is harvested with a rotation α_D of −107°, which is also a mixture of the two epimeric epoxides, but in this crop the β-epoxide predominates. The first crop is dissolved

2871246 6 ALKYL/PROGESTERONES/. 01

2871246 6 ALKYL/17A/17HYDROXY/PROGESTERONES/. 00

2871246 6 ALKYL/DIOSGENIN/. 00

2871246 3 ACYLOXY/16-D ACETOXY G METHYLPENTANOYLOXY-/5 PREGNEN/20ONE/. 01

2871246 6 ALKYL/5 16PREGNADIEN/3 20DIONES/. 00

2871246 6 A/6 METHYL/16A/17A/1617EPOXY/4 PREGNENE/3 20DIONE/. 00

2871246 6 A/6 METHYL/17A/17HYDROXY/PROGESTERONE/. 00

2871246 3 ACYLATES/3 B/3 HYDROXY/6 METHYL/5 16PREGNADIEN/20ONES/. 00

2871246 3 ACYLOXY/26ACETOXY/6 METHYL/5 20-22-FUROSTADIENE/. 00

2871246 3 ACETOXY/6 METHYL/16-D ACETOXY G METHYLPENTANOYLOXY-/5 PREGNEN/20ONES 00

2871246 ./. 00

2871246 3 B/3 HYDROXY/6 ALKYL/5 16PREGNADIEN/20ONES/. 01

2871246 3 B/3 26DIACETOXY/6 ALKYL/5 20-22-FUROSTADIENE/. 00

2871246 3 B/3 ACYLOXY/6 ALKYL/5 16PREGNADIEN/20ONE/. 01

2871246 3 B/3 HYDROXY/6 ALKYL/5 16PREGNADIEN/20ONE/. 00

2871246 3 B/3 HYDROXY/6 ALKYL/16A/17A/1617EPOXY/5 PREGNEN/20ONES/. 00

2871246 3 B/3 HYDROXY/5 A/6 A/5 6 EPOXY/22A/22ALLOSPIROSTANE/. 00

2871246 6 B/6 METHYL/22A/22ALLOSPIROSTANE/3 B/5 A/3 5 DIOL/. 02

2871246 3 B/3 ACETOXY/6 B/6 METHYL/22A/22ALLOSPIROSTAN/5 A/5 OL/. 03

2871246 6 B/6 METHYL/22A/22SPIROSTANE/3 B/5 A/3 5 DIOL/. 01

2871246 3 B/3 ACETOXY/16B/16-D ACETOXY G METHYLPENTANOYLOXY-/6 METHYL/5 PREGNE 01

2871246 ,N/20ONE/. 01

2871246 3 B/3 HYDROXY/6 METHYL/5 16PREGNADIEN/20ONE/. 05

2871246 3 B/3 ACETOXY/5 A/6 A/5 6 EPOXY/22A/22SPRIOSTANE/. 00

2871246 6 B/6 ETHYL/22A/22ALLOSPIROSTANE/3 B/5 A/3 5 DIOL/. 00

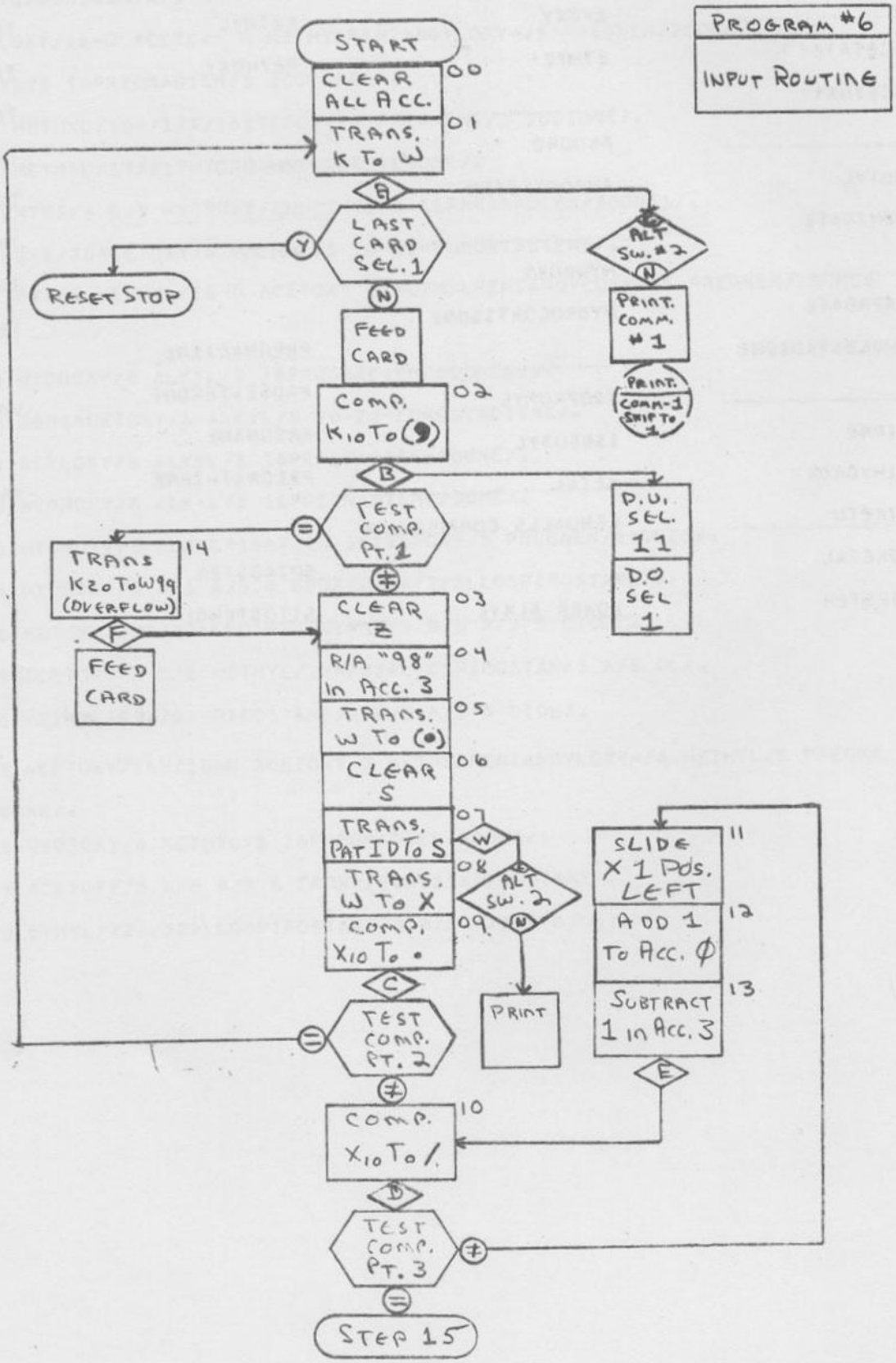
Appendix E

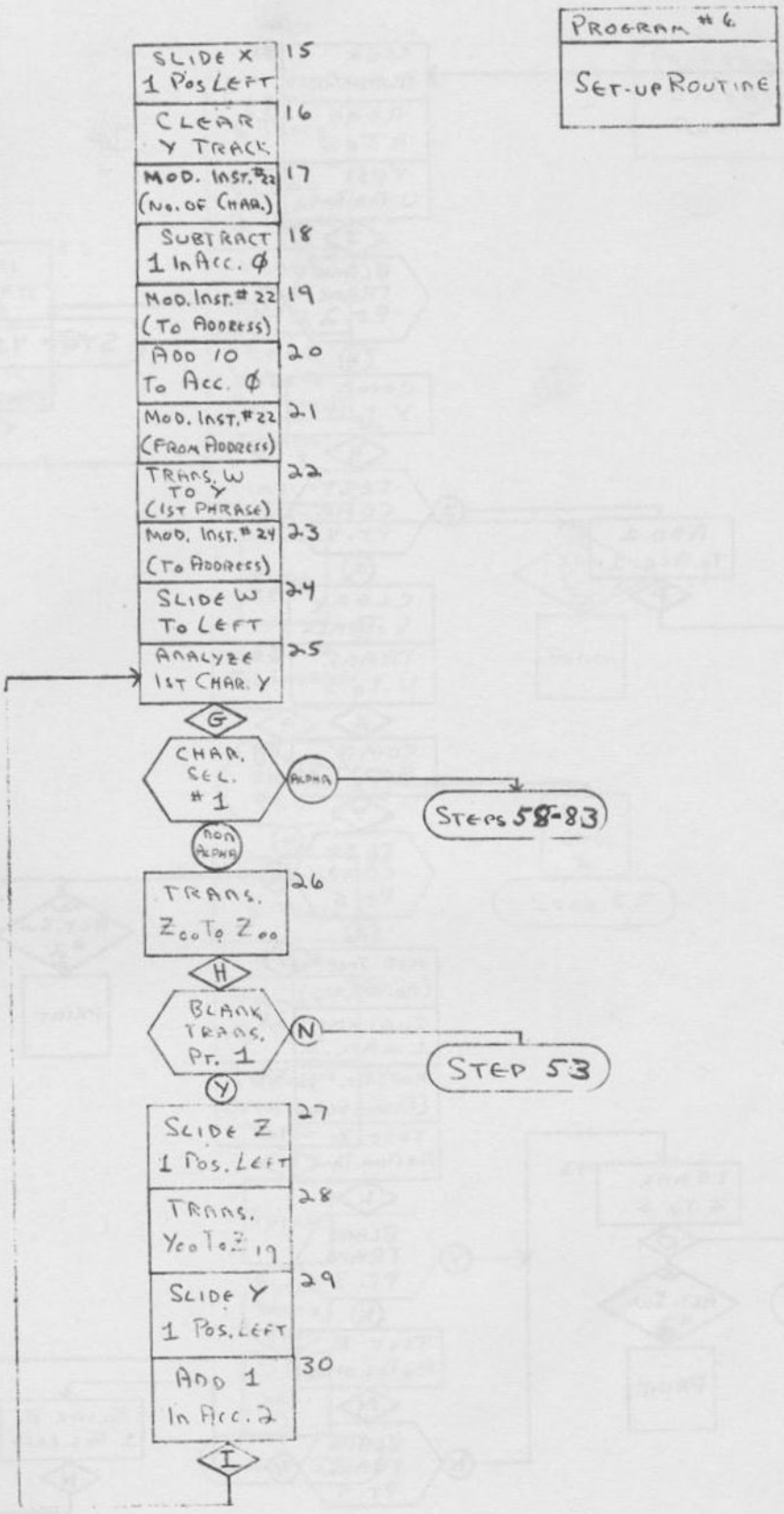
A partial list of the steroid dictionary generated from the document extracts.

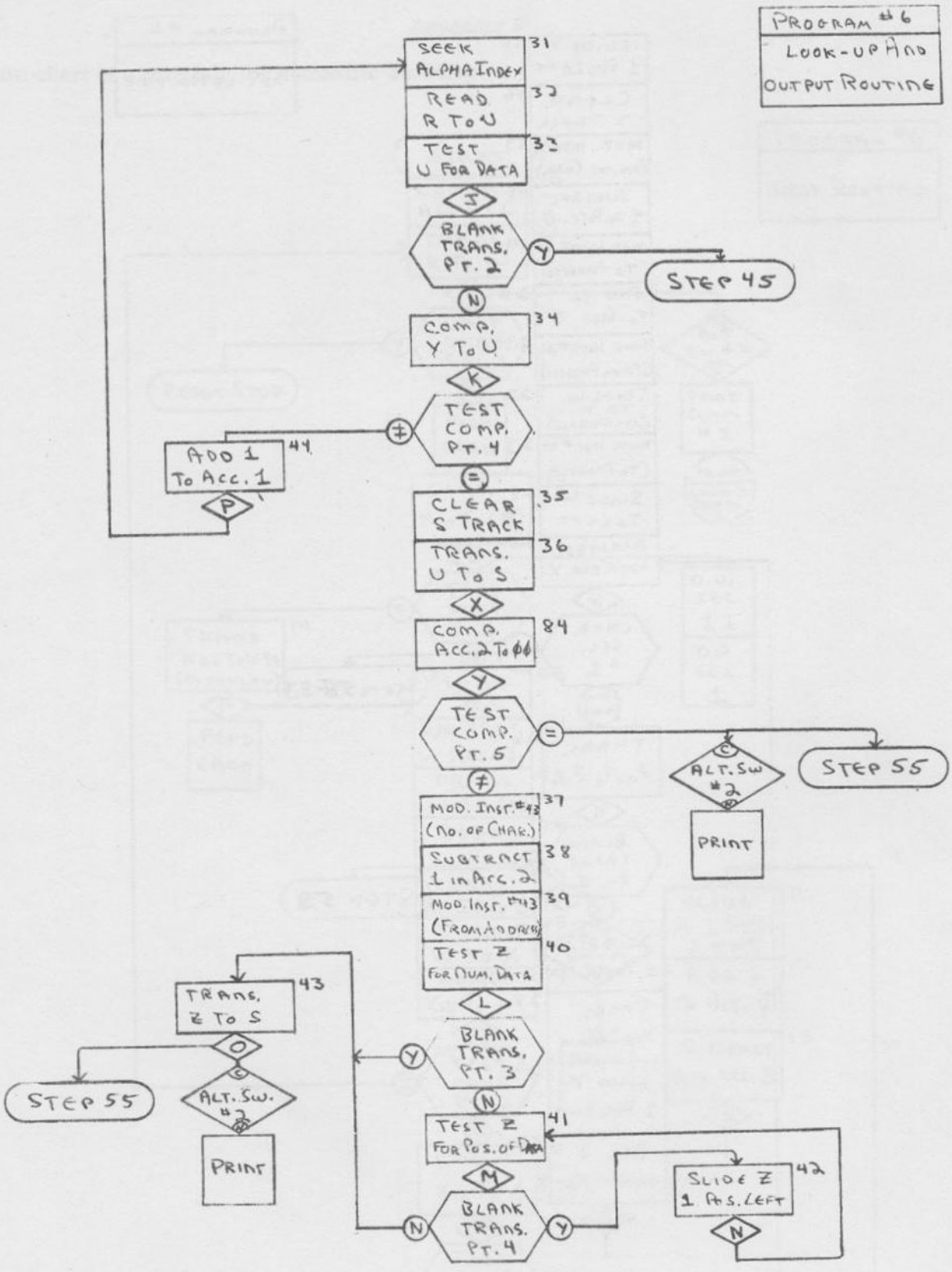
A	EPOXY	METHYL	TRIONE
ACETATE	ETHYL	METHOXY	TRIHYDROXY
ACETOXY			TESTOSTERONE
	FLUORO	ONE	TRIOL
BUTYL	FUROSTADIEN	OL	TRIKETO
BENZOATE		OXIDO	TRIFLUOROACETYL
	HYDROXY		TETRAHYDROXY
CAPROATE	HYDROCORTISONE		TETROL
CHOLESTADIENE		PREGNADIENE	
	ISOPROPYL	PROGESTERONE	VALERYL
DIONE	ISOBUTYL	PREGNANE	
DIHYDROXY	KETAL	PREGNATRIENE	YAMOGENIN
DIKETO	KENDALLS COMPOUND A		
DIKETAL		SPIROSTEN	
DIESTER	LOWER ALKYL	SITOSTEROL	

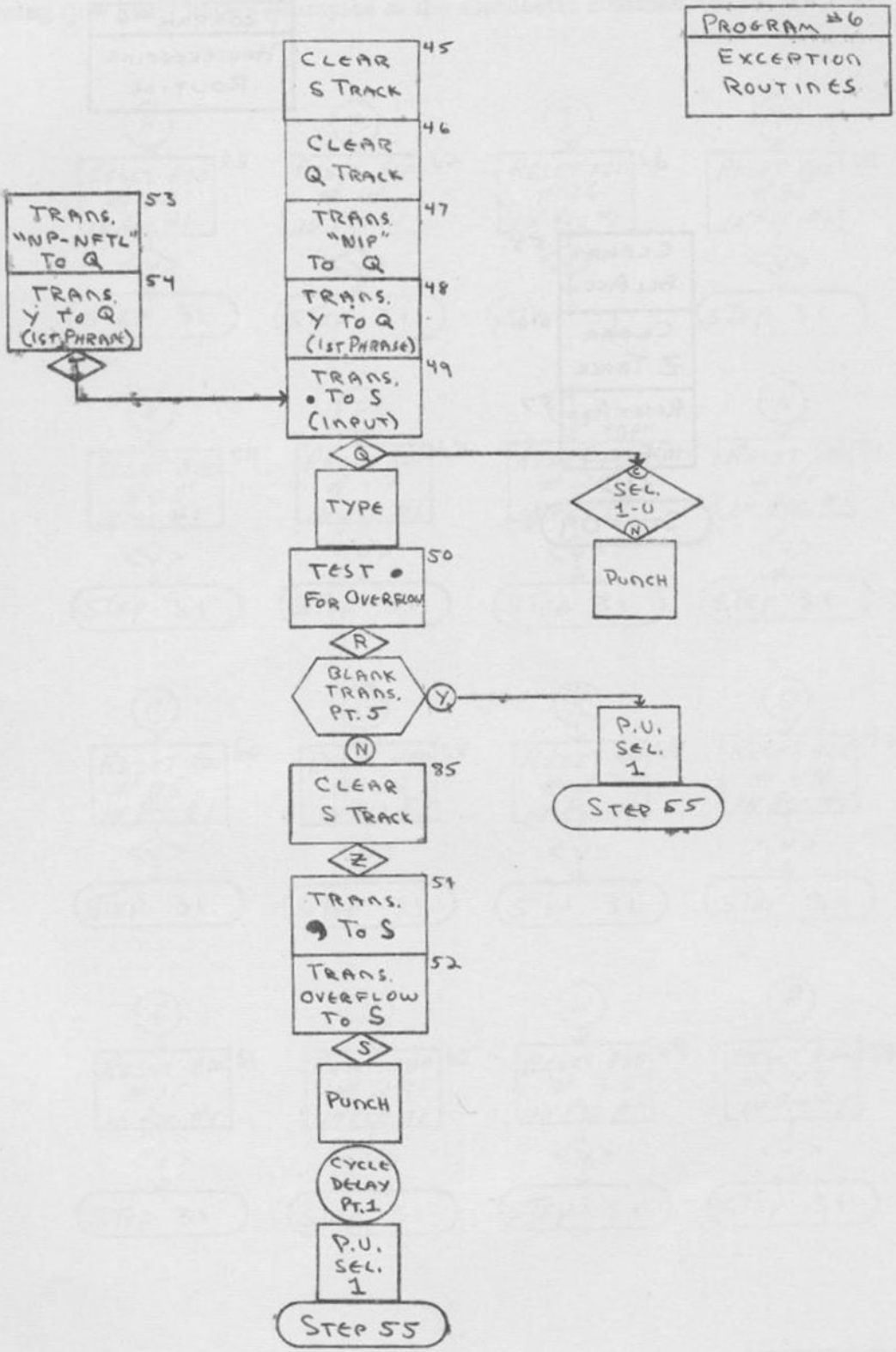
Appendix F

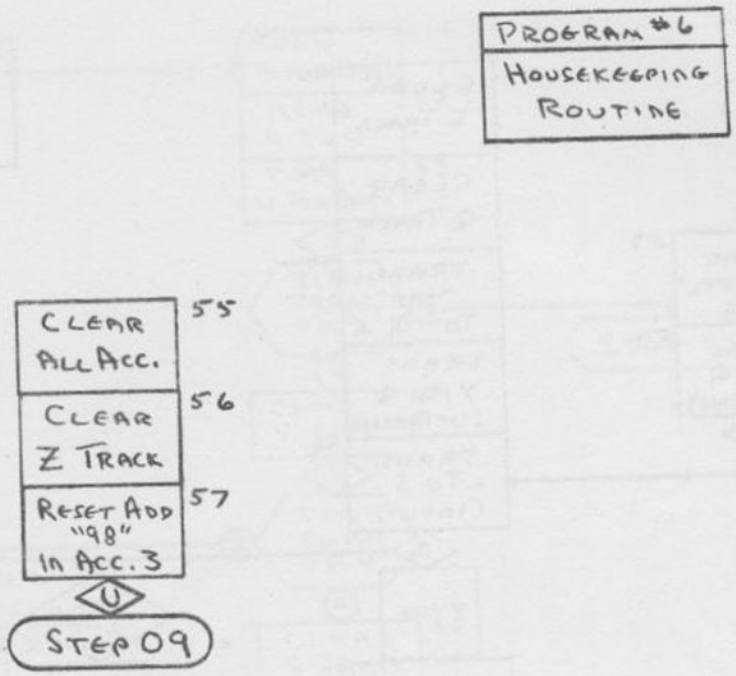
A flow chart of a program for automatic encoding.



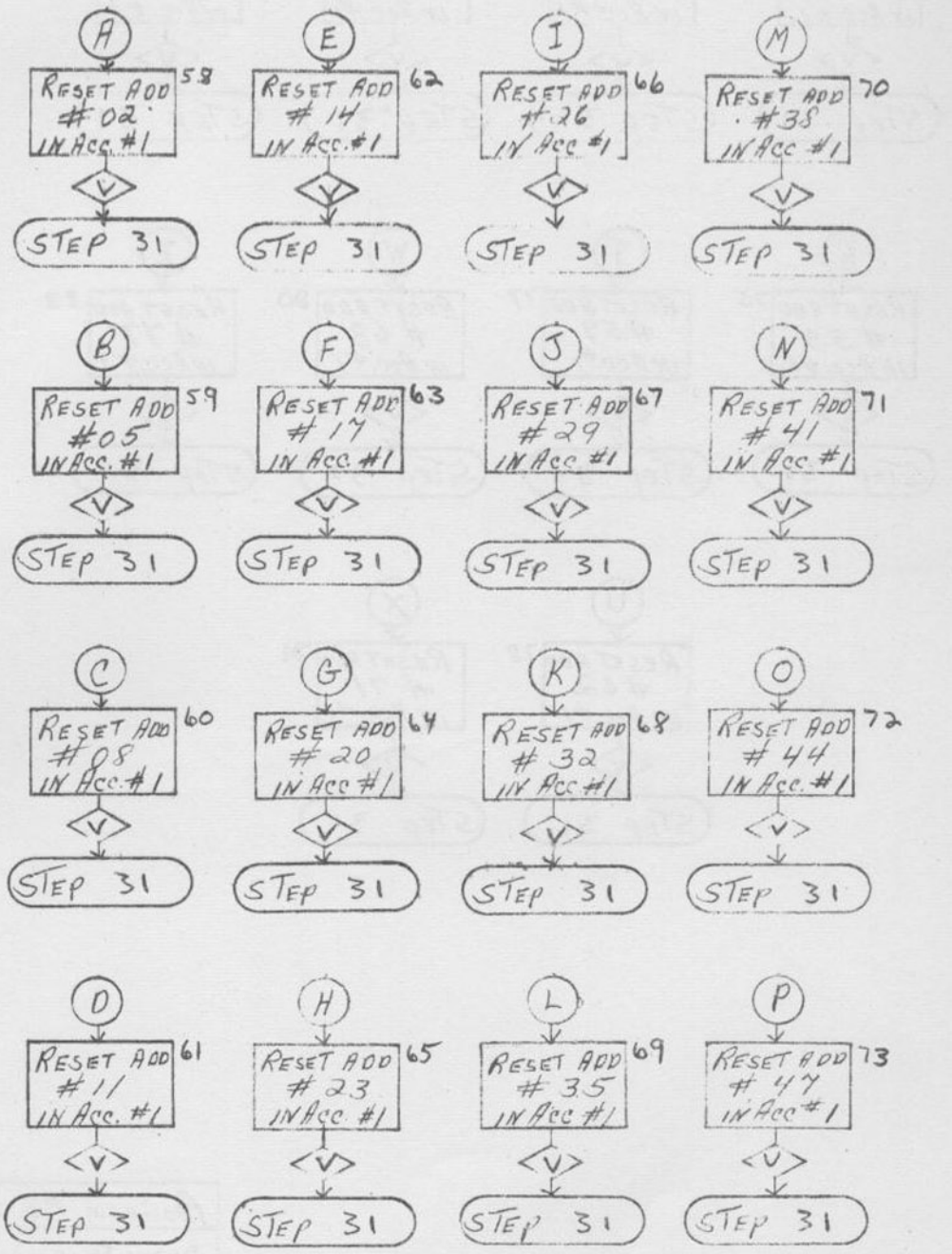

PROGRAM #6
SET-UP ROUTINE
SLIDE X 1 POS LEFT
15
CLEAR Y TRACK
16
MOD. INST. #22 (No. OF CHAR.)
17
SUBTRACT 1 IN ACC. Ø
18
MOD. INST. # 22 (TO ADDRESS)
19
ADD 10 TO ACC. Ø
20
MOD. INST. #22 (FROM ADDRESS)
21
TRANS. W TO Y (1ST PHRASE)
22
MOD. INST. # 24 (TO ADDRESS)
23
SLIDE W TO LEFT
24
ANALYZE 1ST CHAR. Y
25
G
CHAR. SEL. # 1
ALPHA
STEPS 58-83
NON ALPHA
TRANS. Z00 TO Z00
26
H
BLANK TRANS. PT. 1
N
STEP 53
Y
SLIDE Z 1 POS. LEFT
27
TRANS. Y00 TO Z19
28
SLIDE Y 1 POS. LEFT
29
ADD 1 IN ACC. 2
30
I


PROGRAM #6
LOOK-UP AND OUTPUT ROUTINE
SEEK ALPHA INDEX
31
READ R TO U
32
TEST U FOR DATA
33
J
BLANK TRANS. PT. 2
Y
STEP 45
N
COMP. Y TO U
34
K
TEST COMP. PT. 4
≠
ADD 1 TO ACC. 1
44
P
=
CLEAR S TRACK
35
TRANS. U TO S
36
X
COMP. ACC. 2 TO ØØ
84
Y
TEST COMP. PT. 5
=
C
ALT. SW #2
PRINT
STEP 55
≠
MOD. INST. #43 (NO. OF CHAR.)
37
SUBTRACT 1 IN ACC. 2
38
MOD. INST. #43 (FROM ADDRESS)
39
TEST Z FOR NUM. DATA
40
L
BLANK TRANS. PT. 3
Y
N
TEST Z FOR POS. OF DATA
41
M
BLANK TRANS. PT. 4
N
Y
SLIDE Z 1 POS. LEFT
42
N
TRANS. Z TO S
43
O
STEP 55
C
ALT. SW. #2
PRINT

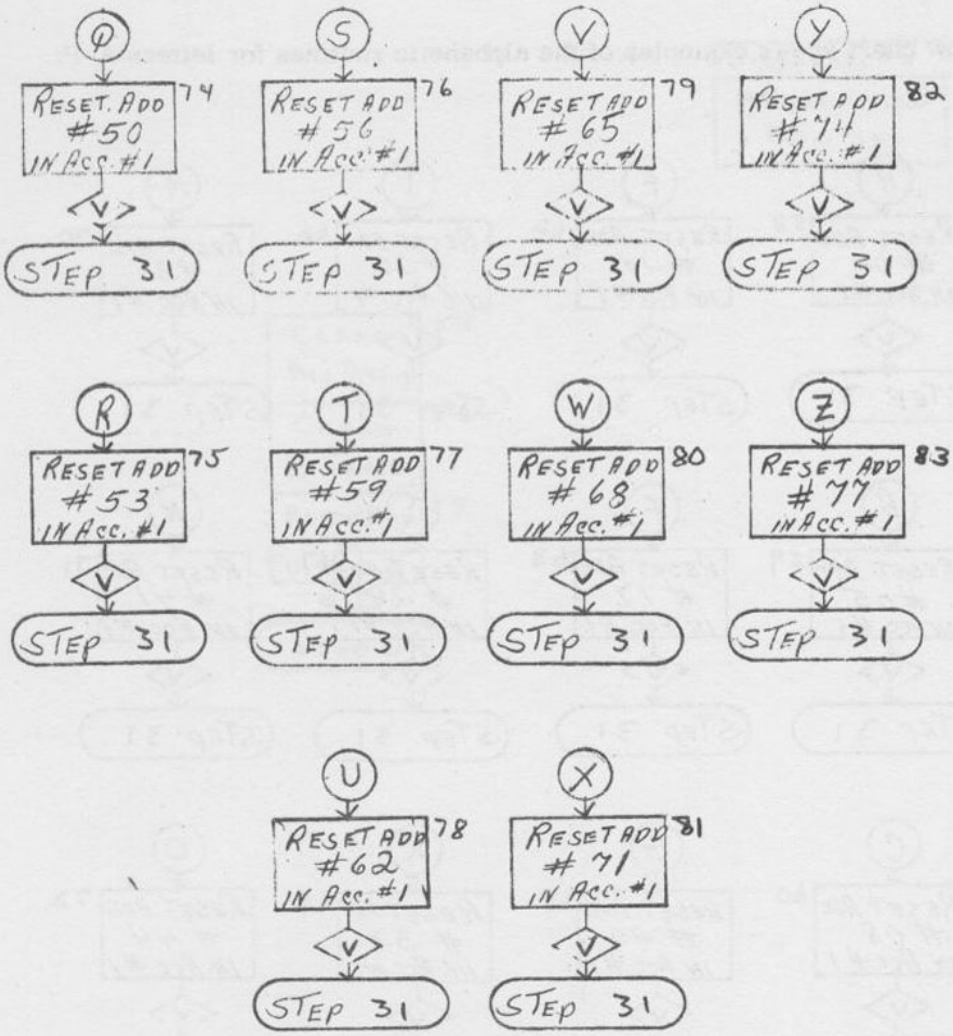

PROGRAM #6
EXCEPTION ROUTINES
CLEAR S TRACK
45
CLEAR Q TRACK
46
TRANS. "NP-NFTL" TO Q
53
TRANS. "NIP" TO Q
47
TRANS. Y TO Q (1ST PHRASE)
54
TRANS. Y TO Q (1ST PHRASE)
48
TRANS. • TO S (INPUT)
49
Q
TYPE
TEST • FOR OVERFLOW
50
C
SEL. 1-U
N
PUNCH
R
BLANK TRANS. PT. 5
Y
P.U. SEL. 1
STEP 55
N
CLEAR S TRACK
85
Z
TRANS. , TO S
51
TRANS. OVERFLOW TO S
52
S
PUNCH
CYCLE DELAY PT. 1
P.U. SEL. 1
STEP 55


PROGRAM #6
HOUSEKEEPING ROUTINE
CLEAR ALL ACC.
55
CLEAR Z TRACK
56
RESET ADD "98" IN ACC. 3
57
U
STEP 09

The following flow chart shows examples of the alphabetic routines for letters A-P:



PROGRAM #6
ALPHA ROUTINES



PROGRAM #6

ALPHA ROUTINES (CON'T)