

MECHANIZED SEARCHING OF PHOSPHORUS COMPOUNDS

II. A PUNCHED CARD SYSTEM FOR PHOSPHORUS COMPOUNDS (Card Automatic Mechanization of Phosphorus—CAMP)*

By J. FROME

U. S. Department of Commerce, Patent Office, Washington, D. C.

INTRODUCTION

In the field of information retrieval, several important factors may be considered, such as the cost of the machine, the availability of machines and the field of interest of the user.

One or all of these factors may be an important consideration to any corporation, large or small, contemplating the installation of machines for information retrieval purposes. The cost factor of the machine in the majority of cases is the biggest hurdle facing users today. Therefore it is desirable to be able to mechanically search and retrieve documents for as low a cost as possible. One of the machines that may be able to render an effective mechanized search is the common punch card machine (sorter).

In these days of specialization, user interest may be confined to specific fields of organic chemistry, for which this system was primarily created, *i.e.*, organic phosphorus compounds.

The objectives of this system are much the same as the ones described in RAMP (Paper I), with this exception, to create and develop a chemical compound mechanized search system that can be effectively and expeditiously used for certain chemical subject matter using a low-cost punch card machine. The machine used is the Census Multicolumn sorter.

SUBJECT MATTER

The field of chemistry selected for "CAMP" includes all the patents in class 260, subclass 461, including official and unofficial cross references as classified in the U. S. Patent Office.

System.—A system for retrieval of organic compounds should have at least two features: (a) the division of the compounds into building blocks (*i.e.*, NH_2 , COOH , SO_3H); (b) a method of showing relation between these. The building blocks used in this system are those ordinarily recognized by chemists and for this particular system are listed later, the list being called the fragment dictionary.

Relationships are extremely important and are shown in this system by the use of a matrix, *i.e.*

| | O ₁ | O ₂ | O ₃ | S ₁ |
|-------|----------------|----------------|----------------|----------------|
| alkyl | X | | | |
| aryl | | | | |
| | | | | |

Thus if an alkyl group is attached to a single oxygen group (O₁) this relationship is indicated in the matrix by a cross in the intersecting box. This box is given a column and row number. If an aryl group is connected to a sulfur atom (S₁) the box corresponding to the intersection of the horizontal row representing aryl and the vertical row representing S₁ is then marked, *i.e.*

| | O ₁ | O ₂ | O ₃ | S ₁ |
|-------|----------------|----------------|----------------|----------------|
| alkyl | | | | |
| aryl | | | | X |

This box is also given a column and row number. By means of this matrix we are able to show the connection between the organic phosphorus nucleus and the fragments which are either directly or indirectly attached to it. By a further extension of this device we are able to show further relationship as described later.

DEFINITION OF TERMS

1. A fragment is a chemical element or a collection of chemical elements treated as a unit for chemical or information retrieval purposes. It is a component part of a structural formula, and a series of designated fragments will constitute a structural formula.

2. A node is a collection of at least two fragments. For the purpose of this system, there will be three nodes, first, second and terminal.

GENERAL CODING PRINCIPLES

"CAMP" relies on a matrix to show relationships; in addition it describes fragments contained in the organic phosphorus compounds disclosed in the documents. Two coding sheets were used in most cases for every document analyzed, the first for specific formulas disclosed, the second coding sheet for "Markush" formulations only.

It is important to note that all specific formulas disclosed in each document are composited and coded on one coding sheet, and only one punch card is used to describe all the specific formulas disclosed. The same is true for all Markush formulas.

The "CAMP" coding sheet at present is divided into five sections. The first three sections include the I, II and terminal nodes, the fourth

*Presented before Division of Chemical Literature, ACS National Meeting, New York, September 13, 1960.

section is the fragment dictionary and the fifth section shows both the specific and generic combinations and permutations of the organic phosphorus nucleus which at present encompass the organic phosphates and thiophosphates.

Nodal Designation and Description.—The first node (example 2) has ten generic chemical

| PAT. NO. | | | | | | | | | |
|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| 1st NODE | 12 | 11 | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| | O ₁ | O ₂ | O ₃ | S ₁ | S ₂ | S ₃ | X ₁ | X ₂ | X ₃ |
| Alkyl 1 | | | | | | | | | |
| Sub. Alkyl 2 | | | | | | | | | |
| Alkenyl 3 | | | | | | | | | |
| Alkynyl | | | | | | | | | |
| Sub. Alkenyl 4 | | | | | | | | | |
| Sub. Alkynyl | | | | | | | | | |
| Unsub. Aryl 5 | | | | | | | | | |
| Sub. Aryl 6 | | | | | | | | | |
| H or Metal 7 | | | | | | | | | |
| Misc. 8 | | | | | | | | | |

terms and a miscellaneous in the "Y" or horizontal axis as shown by columns 1-8, and twelve descriptive characters in the "X" or vertical axis as shown in rows 12-6.

The descriptive characters in rows 12-6 refer directly to the generic chemical terms in columns 1-8. In rows 12, 11 and 0, the O₁, O₂, O₃ represent the oxygen atom or atoms directly attached to the phosphorus atom either once, twice or three times. In rows 1, 2 and 3, S₁, S₂, and S₃ represent the sulfur atom or atoms directly attached to the phosphorus atom either once, twice or three times. In rows 4, 5 and 6, X₁, X₂ and X₃ are generic characters and represent either an oxygen or sulfur atom or atoms directly attached to the phosphorus atom either once, twice or three times.

The rows 12-6 in the "X" or vertical axis in the first node are designed to show the nodal relationship with the generic chemical terms in columns 1-8.

In row 7 "SY" represents symmetrical and in row 8 "USY" represents unsymmetrical. These descriptive characters are used only in the first node and are designed to further distinguish each of the three first nodal connections.

Row 9 is used to define further the generic chemical terms in columns 1-8.

The 2nd node has 25 specific and generic chemical fragments and terms in the "Y" or horizontal axis (columns 9-28). Column 29 covers every other possible chemical term or fragment not included in columns 9-28. There are in the second node 12 descriptive characters in the "X" or vertical axis. Each chemical term in the "Y" axis is further defined by row 9 of the "X" axis.

The rows 12-6 in the II node are designed to show the nodal relationship with the generic chemical terms in columns 9-28 and are the same as the first node. In row 7, the abbreviation "CH" represents the attachment of the fragment being coded to a chain and in row 8 "R" represents the attachment of the fragment being coded to a ring.



| 2nd NODE | O ₁ | O ₂ | O ₃ | S ₁ | S ₂ | S ₃ | X ₁ | X ₂ | X ₃ | CH | Ring |
|------------------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----|--------------|
| SH; SR 9 | | | | | | | | | | | R |
| SO; SO ₂ 10 | | | | | | | | | | | SO |
| NR 11 | | | | | | | | | | | Het. |
| SO ₂ NR | | | | | | | | | | | |
| OH-(Acyl-O-) 12 | | | | | | | | | | | Acyl |
| Alkenyl 13 | | | | | | | | | | | Alk- inyl |
| Alkynyl | | | | | | | | | | | Het. |
| R) Amine 14 | | | | | | | | | | | |
| N) | | | | | | | | | | | |
| R) Salt | | | | | | | | | | | |
| NO ₂ 15 | | | | | | | | | | | |
| Alkyl 16 | | | | | | | | | | | High |
| Aryl 17 | | | | | | | | | | | Poly |
| Alkene 18 | | | | | | | | | | | High |
| Arylene 19 | | | | | | | | | | | Poly |
| Halogen 20 | | | | | | | | | | | IF |
| -S- 21 | | | | | | | | | | | |
| -O- 22 | | | | | | | | | | | |
| COOR; H 23 | | | | | | | | | | | Ester |
| R 24 | | | | | | | | | | | Hetero |
| CON R | | | | | | | | | | | |
| CN 25 | | | | | | | | | | | |
| OR 26 | | | | | | | | | | | R |
| R R | | | | | | | | | | | Hetero |
| N R | | | | | | | | | | | |
| R R | | | | | | | | | | | |
| = O; = S 28 | | | | | | | | | | | = S |
| Misc. 29 | | | | | | | | | | | Het. A |

The terminal node (example 4) is almost the same as the II node, the only difference being that there are fewer generic and specific terms and fragments included within it.

































| Last NODE | O ₁ | O ₂ | O ₃ | S ₁ | S ₂ | S ₃ | X ₁ | X ₂ | X ₃ | CH | Ring |
|------------------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----|--------------|
| -S-H; -S-R 30 | | | | | | | | | | | R |
| SO; SO ₂ 31 | | | | | | | | | | | SO |
| R 32 | | | | | | | | | | | Het. |
| SO ₂ N R | | | | | | | | | | | |
| Alkenyl 33 | | | | | | | | | | | Alk- inyl |
| Alkynyl | | | | | | | | | | | O-Acyl |
| OH; O-Acyl 34 | | | | | | | | | | | = S |
| = O; = S 35 | | | | | | | | | | | Het. |
| R) Amine 36 | | | | | | | | | | | |
| N) | | | | | | | | | | | |
| R) Salts | | | | | | | | | | | |
| NO ₂ 37 | | | | | | | | | | | |
| Alkyl 38 | | | | | | | | | | | Higher |
| Aryl 39 | | | | | | | | | | | Poly |
| Halogen 40 | | | | | | | | | | | If |
| COOR; H 41 | | | | | | | | | | | Ester |
| R 42 | | | | | | | | | | | Het. |
| CON R | | | | | | | | | | | |
| CN 43 | | | | | | | | | | | |
| R R | | | | | | | | | | | Het. |
| N R | | | | | | | | | | | |
| R R | | | | | | | | | | | |
| OR 45 | | | | | | | | | | | Aryl |
| Misc. 46 | | | | | | | | | | | Hetero |

The fourth section of the coding sheet is the fragment dictionary which is used to define further the fragments in the II and terminal nodes. In the situation where the structural formula being coded has more than three nodes, the fragments in the formula not contained in either the II or terminal nodes will be checked in the fragment dictionary also.

Fragment Dictionary

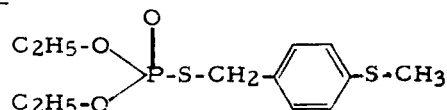
| | 50 | 51 | 52 |
|------------|---------------------------|--|-------|
| Alkyl | 12 -S- | 2 Amine | 11 |
| Low Alkyl | 11 -S | 3 NH ₂ | 0 |
| High Alkyl | 0 SO ₂ N< | 4 -N<H | 1 |
| Alkylene | 1 -SO- | 5 -N< | 2 |
| 1 | 2 -SO ₂ - | 6 >N< | 3 |
| 2 | 3 SO ₃ ; H; Me | 7 Het. +  | 4 |
| 3+ | 4 SH; SMe | 8 Het.  | 5 |
| Alkenyl | 5 -S-S- | 9 NO ₂ | 6 |
| C=C | 6 SO ₄ | 52/12 =N (imine) | 7 |
| Alkynyl | 7 | 52 Het.-A | 53 |
| C≡C | 8 -O- | 8 N-Hetero | 2 |
| Aryl | 9 =O | 9 S-Hetero | 3 |
| Benzene | 51/12 OH | 53/12 O-Hetero | 4 |
| Napthalene | 11 OME | 11 N-O-Hetero | 5 |
| Cycloalkyl | 0 | 0 N-S-Hetero | 6 |
| Cyclohexyl | 1 | 1 S-O-Hetero | 7 |
| | 54 | 55 | |
| Halogen | 0 N-Hetero | 12 O-Hetero | 8 |
| Cl-1 | 1 Pyridine | 11 Furan | 9 |
| Cl-2 | 2 Piperidine | 0 Mis-O-HET. | 56/62 |
| Cl-3 | 3 Pyrrolidine | 1 | |
| Cl-4+ | 4 Morpholine | 2 Hetero-Sat-N-6M | 11 |
| F | 5 Thiazole | 3 Misc.Het. | 0 |
| Br | 6 Misc.-N HET. | 4 Quinoline | 1 |

The fifth section of the coding sheet represents the organic phosphorus nucleus. This section is the starting point in the coding of every structural formula. All of the essential combinations and permutations of the organic phosphorus nucleus both generic and specific are included here. It is well to note that these combinations and permutations only include the organic phosphates and thiophosphates.

| | | 47 | | | 47 |
|---|------------------------|----|---|--|-------|
|  | | |  | | |
|  | | 12 |  | | 6 |
|  | } 1-S NO Triaryl | |  | | |
|  | | |  | | 7 |
|  | | |  | | |
|  | | 0 |  | | 8 |
|  | | |  | | |
|  | | 1 |  | | 9 |
|  | | |  | | |
|  | | 2 |  | | 48/12 |
|  | | |  | | |
|  | | 3 |  | | 11 |
|  | | |  | | |
|  | | 4 |  | | 0 |
|  | | |  | | |
|  | | 5 |  | | 1 |

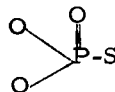
All the nodes, the fragment dictionary and the organic phosphorus nucleus will be described further by a specific example.

Example:



The encoding of structural formulas in the "CAMP" system begins by separating the organic phosphorus nucleus from the rest of the compound. The organic phosphorus nucleus is defined as the phosphorus atom and all the elements directly connected to it.

For coding purposes every structural formula will be completely coded both specifically and generically in the organic phosphorus nucleus and also in the first, second and terminal nodes as well as the fragment dictionary. In the example noted it would be done in this manner



Column 47 Row 5

The analyst would code column 47, Row 5 on the coding sheet to show the specific organic phosphorus nucleus, and also check Col. 47 Row 12, Col. 47 Row 11, Col. 47 Row 8, and Col. 48 Row 12. The reason the generic phosphorus nucleus is coded in addition to coding the specific configuration is to enable the searcher to ask a more generic question. So in this example the phosphorus atom has directly connected to it a [=O], [O], [O], and an [S].

To designate the first nodal relationship, the fragment or fragments which are directly attached to either the oxygen or sulfur atoms are by the combination of the two (oxygen or sulfur and the fragment) the first node. In the example cited, then, the first node would be coded on the coding sheet in this manner.

O₂-(C₂H₅) or [O₂-(Alkyl)] Col. 1 Row 11

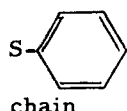
SY Col. 1 Row 7

S1-CH₂ or [S1-Alkyl] Col. 2 Row 1

USY Col. 1 Row 8

The reason O₂-(C₂H₅) or O₂-Alkyl was coded (Col. 1 Row 11) is that both oxygen atoms are directly connected to the ethyl or alkyl fragment, SY (symmetrical). Col. 1 Row 7 was coded because the ethyl fragments attached to the oxygen atoms are identical. Col. 2 Row 1 is self explanatory. USY (unsymmetrical) (Col. 2 Row 8) was coded because the sulfur and the alkyl fragment are not identical with the other first node connections. The first node also will be coded generically for reasons stated above, so Col. 1 Row 5 is coded and also Col. 2 Row 4.

The II node in the example is the combination of the sulfur or oxygen atom or atoms and the fragment twice removed from it, which in this example would be S-C₆H₄-only, since there is only one fragment connected to the oxygen atoms. The second node would then be coded specifically as



or [S1-Sub. Aryl] Col. 17 Row 1

Col. 17 Row 7

S- was coded specifically in Col. 17 Row 1 because the sulfur atom was connected to the phenyl fragment. CH (chain Col. 17 Row 7) was coded since the phenyl is attached to a chain. As stated above in the II and terminal nodes, the fragment will also be coded in the fragment dictionary. So Col. 50 Row 9 and Col. 51 Row 12 also will be coded.

The second node must be coded generically so Col. 17 Rows 4 and 7 must be coded also.

The terminal nodal relationship is a combination of the sulfur or oxygen atom or atoms with the fragment furthest removed in the structural formula. The terminal node is the final fragment, but in order to be terminal it must be further removed than the second node. In this case it would be S1-[S-CH₃] or S1-[SR].

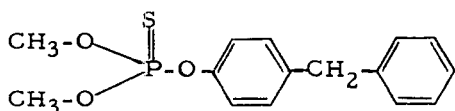
S1-[S-CH₃] or S1-[SR] Col. 30 Row 1
R Col. 30 Row 8
and 9

The reason we have combined two fragments here is that in Col. 30 the S-R represents sulfur which has directly attached to it a hydrocarbon radical, whether it is a hydrocarbon ring or chain. Col. 50 Row 11 and Col. 51 Row 2 also will be coded in the fragment dictionary. Col. 30 Row 4, 8 and 9 are the generic codes for the terminal node.

ASKING THE QUESTION

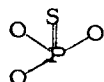
The procedure for asking the search question is this: The examiner, after analyzing the applicant's claims, determines what the essence of invention is and by using the matrix system of nodal relationships previously described, asks the question by checking the appropriate blocks on the coding sheet, depending on whether a specific or generic answer is desired. The board is then wired by the machine operator and the file of cards is fed into the Multicolumn sorter and the cards representing the documents are retrieved. A specific illustration of an actual search would be:

Example: It is desired to find the compound



so it will be coded specifically in this way:

Organic Phosphorus Nucleus



Col. 47 Row 2

I Node

O₂-(Alkyl) Col. 1 Row 11

SY Col. 1 Row 7

O₁-Sub. Aryl Col. 6 Row 12

USY Col. 6 Row 8

II Node

O1-Alkylene Col. 18 Row 12

Ring Col. 18 Row 8

Terminal Node

O₁-Aryl Col. 39 Row 12

Chain Col. 39 Row 7

It is well to note that when we say the compound is being asked specifically we mean the organic phosphorus nucleus and those elements directly attached to the phosphorus atom are asked specifically but all other fragments in the formula are usually generic to those elements except in a few instances insofar as the nodal relationship is concerned.

To ask the above question generically Col. 47 Row 12 would be coded in the organic phosphorus nucleus section. In all the nodes the same column would always be checked but instead of checking the rows O₁ or O₂ or S₁ one would check X₁ and X₂.

At present this system is only being used experimentally in the U. S. Patent Office. The following are actual statistics for approximately 150 searches which have been made up to this time.

PHOSPHORUS MACHINE SEARCHES (CAMP)

| | |
|------------------------------------|-----------|
| 1. Applications Searched | 52 |
| 2. Total Searches Made | 104 |
| 3. Searches per Application | 2 |
| 4. Time to Prepare Search Question | 3 minutes |
| 5. Machine Time per Search | 2 minutes |
| 6. Time to Wire Machine | 3 minutes |
| 7. Patents Retrieved per Search | 21 |

It is well to note that "CAMP" already has been revised in order to take care of all organic phosphorus compounds other than phosphates and thiophosphates in class 260 subclass 461. The coding manual for the revised "CAMP" sheet will be available at the U. S. Patent Office in the near future.

Conclusion.—It is felt that the matrix system using nodal relationships on a low cost punch card machine may be for certain chemical subject matter one of the answers to low cost mechanized searching.