A Program for Rapid and Automatic Functional Group Recognition

By **Ashmeed Esack** * and **Malcolm Bersohn**, Department of Chemistry, University of Toronto, Toronto Ontario, Canada M5S 1A1

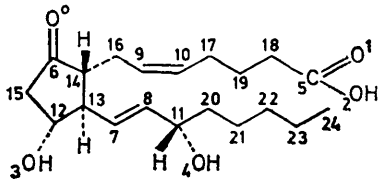
An algorithm has been devised which, when applied to a canonical connection table description of an organic molecule, produces the functional groups therein. The procedure has been programmed for all non-heterocyclic molecules containing chlorine, sulphur, phosphorus, oxygen, nitrogen, and carbon atoms.

In a computer analysis of mass spectra or in computing synthetic pathways the simulation of chemical reactions is necessary. To access the relevant and feasible reactions one must first recognize the functional groups of a molecule. This is true whether the molecule is a product from which we wish to generate the reactant giving rise to it or a reactant from which we wish to generate a product. In one approach to functional group recognition, a list of the substructures of interest is drawn up and the molecule is repeatedly queried as to the existence of members of that list.[1,2] Although screening techniques[3] are employed in narrowing the choice among particular members of the list, the large number of such queries and the high percentage of negative answers makes this a time-consuming and hence unattractive approach. The use of a canonical connection table is essential in our method and allows us to take the opposite approach wherein the asking of questions has largely been eliminated. The substructure discovery process is largely driven by the data, the canonical connection table of the molecule, and almost never by our knowledge of organic chemistry. Except in a very few cases, no routine of our program can be described as a search for any particular functional group and in this way negative answers are not possible. Corey and his co-workers[4] have developed an algorithm which involves logical manipulation of sets. Their procedure, which is knowledge-directed as well as data-driven, ideally suits their purpose but is deemed less efficient in a non-interactive synthesis design program which must make up in high speed what it lacks in human sophistication.

Our program, and this paper, is structured as follows. (1) Recognition of the 'elementary functional groups' and preparation of their machine representations, *i.e.* their machine formulae. An elementary functional group is usually the cluster of atoms immediately surrounding the nearest neighbour of a hetero-atom and some examples are listed in Table 2. (2) Discovery of 'complex functional groups,' that is, the discovery of special relationships between the elementary functional groups, such as α-, β-, *ortho*-, and *meta*-relationships. We also discover the adjacencies of elementary groups to a carbon–carbon double bond, the allylic relationship, and, to an aromatic carbocyclic ring, the benzylic relationship. Once the functional groups have been found by the method described in this paper, retrieval of reactions relevant to any functional group becomes a simple operation. The mechanism by which this is done is outlined in the final section.

Since the procedure represents an operation performed on the machine representation of the molecule, we will examine this briefly by using a member of the prostaglandin family, PGE_2, as our model.

Molecular Representation.—Numerous ways exist[5] for representing chemical structures in a binary computer. The form we have chosen is a connection table in which there is a canonical numbering of the atoms (*cf.* Table 1). A detailed description of our canonical connection table has been published.[2] (We are now omitting the columns 4 and 5 described in that paper.) Each non-hydrogen atom is described by a 64-bit row,



the rows being arranged according to the nature of the corresponding atom, in descending order of atomic number (thus chlorine before oxygen before carbon), then in descending order of degree of unsaturation of the atom (*e.g.* =O precedes –O–), then in ascending order of the number of attached hydrogen atoms, then in order of stereochemical features. Two atoms equivalent in all the above aspects are ranked according to these same features of their neighbours in the molecule. (If necessary, the examination is extended to neighbour's neighbours, *etc.*) Finally the row numbers of all adjacent atoms of each atom are listed in ascending order within each row. This procedure leads to a canonical ordering of the atoms which is similar to that obtained if one uses the Cahn–Ingold–Prelog sequence rules.[6]

All non-hydrogen atoms are described. Column 1 denotes the atomic number of the atom, column 2 its unsaturation, and column 3 shows four minus the number of attached hydrogen atoms. Columns 5—7 convey information as to whether or not the atom is a member of a three-, five-, or six-membered ring, respectively, and column 4 denotes membership in a ring

[1] J. Figueras, *J. Chem. Documentation*, 1972, **12**, 237.
[2] M. Bersohn and A. Esack, *Chemica Scripta*, 1974, in the press.
[3] G. W. Adamson, J. Cowell, M. F. Lynch, A. H. W. Mclure, W. G. Town, and A. M. Yapp, *J. Chem. Documentation*, 1973, **13**, 153.
[4] E. J. Corey, W. T. Wipke, R. D. Cramer, and W. J. Howe, *J. Amer. Chem. Soc.*, 1972, **94**, 431.
[5] M. F. Lynch, J. M. Harrison, W. G. Town, and J. E. Ash, 'Computer Handling of Chemical Structure Information,' McDonald, London, 1972.
[6] R. S. Cahn, *J. Chem. Educ.*, 1964, **41**, 116.

other than three-, five-, or six-membered. Stereochemical features are described in column 8, and columns 9 and 10 state whether or not the atom is positively or negatively charged, respectively. *R*- and *S*-Chirality are indicated in columns 11 and 12, respectively; columns 13—16 list the row numbers of the adjacent atoms in ascending order. The number 255 is our notation for a blank.

TABLE 1

Canonical connection table for the prostaglandin PGE_2

Row no.	Column no. 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	8	2	4	0	0	0	0	0	0	0	0	0	6	6	255	255
1	8	2	4	0	0	0	0	0	0	0	0	0	5	5	255	255
2	8	0	3	0	0	0	0	0	0	0	0	0	5	255	255	255
3	8	0	3	0	0	0	0	8	0	0	0	0	12	255	255	255
4	8	0	3	0	0	0	0	0	0	0	0	0	11	255	255	255
5	6	2	4	0	0	0	0	0	0	0	0	0	1	1	2	18
6	6	2	4	0	0	1	0	0	0	0	0	0	0	0	14	15
7	6	2	3	0	0	0	0	8	0	0	0	0	8	8	13	255
8	6	2	3	0	0	0	0	0	0	0	0	0	7	7	11	255
9	6	2	3	0	0	0	0	0	0	0	0	0	10	10	16	255
10	6	2	3	0	0	0	0	6	0	0	0	0	9	9	17	255
11	6	0	3	0	0	0	0	12	0	0	0	1	4	8	20	255
12	6	0	3	0	0	1	0	12	0	0	1	0	3	13	15	255
13	6	0	3	0	0	1	0	12	0	0	1	0	7	12	14	255
14	6	0	3	0	0	1	0	0	0	0	1	0	6	13	16	255
15	6	0	2	0	0	1	0	8	0	0	0	0	6	12	255	255
16	6	0	2	0	0	0	0	5	0	0	0	0	9	14	255	255
17	6	0	2	0	0	0	0	0	0	0	0	0	10	19	255	255
18	6	0	2	0	0	0	0	0	0	0	0	0	5	19	255	255
19	6	0	2	0	0	0	0	0	0	0	0	0	17	18	255	255
20	6	0	2	0	0	0	0	0	0	0	0	0	11	21	255	255
21	6	0	2	0	0	0	0	0	0	0	0	0	20	22	255	255
22	6	0	2	0	0	0	0	0	0	0	0	0	21	23	255	255
23	6	0	2	0	0	0	0	0	0	0	0	0	22	24	255	255
24	6	0	1	0	0	0	0	0	0	0	0	0	23	255	255	255

The Concept of a Central Atom.—Table 2 lists some examples of elementary functional groups recognized by our program. In our scheme each group is thought to possess one central atom (see Table 2, where the asterisk denotes the central atom of the group). Our concept of a central atom is in some ways related to the concept of a 'focus' employed in the DARC system.[7] By definition, central atom *i* is central to some elementary functional group in the sense that all the atoms that comprise that group have their row numbers listed in columns 13—16 of the *i*th row of the connection table. For example, in PGE_2 carbon 5 is the central atom of the carboxy-group and oxygens 1 and 2 have their row numbers listed in the 5th row of the connection table. However we have deliberately created exceptions to this definition in the case of the nitro-, nitroso-, hydroxylamine, and oxime groups. According to the definition of a central atom, the nitrogen atom would appear to be the central atom in each case. The reason for choosing the carbon atom instead will become clear shortly.

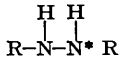
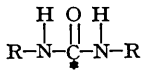
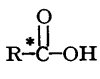
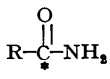
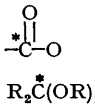
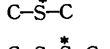
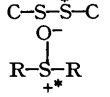
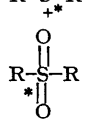
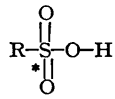
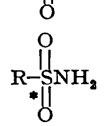
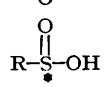
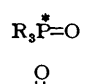
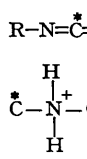
The concept of a central atom, which can be sulphur, phosphorus, oxygen, nitrogen, or carbon, has proved very useful for three reasons. (1) Detection of a central atom implies discovery of an elementary functional group. (2) Detection of central atoms is fast. Sulphur, phosphorus, oxygen, and nitrogen central atoms rank high in the connection table so that in a top-down scan of the table they are encountered early. A carbon central atom is likewise readily perceived since it appears in the connection table as a neighbour of the heteroatom to which it is bonded. That is, the row number of the carbon central atom is listed in column 13 of the heteroatom's row and so is retrieved on examination of this row. In PGE_2 the central atoms are 5, 6, 11, and 12. Examination of Table 1 shows that central atom 6 appears in column 13 of row number 1 of the table. Similarly, central atom 11 is retrieved on examination of row number 4. The rules by which sulphur, phosphorus, oxygen, and nitrogen central atoms are detected are specific and simple and are detailed in a later section. (3) α-, β-, *ortho*-, *meta*-, allylic, and benzylic relationships between elementary functional groups are deduced from the relative locations of these central atoms.

In the present version of this program sulphur, phosphorus, oxygen, and nitrogen central atoms serve only in the discovery and formulation of the elementary functional groups to which they belong and the relationships between these and other groups in the molecule are not sought. Thus, if a molecule possessed the peroxide group, one of the peroxide oxygen atoms (it is immaterial which oxygen is chosen) would normally be selected as its central atom and employed in the preparation of the peroxide group machine formula, that is, in reporting the discovery of the peroxide group in the molecule. However the oxygen central atom is not employed in the discovery of special relationships, such as α or β, between the peroxide and other groups which may be present in the molecule. Since the number of such special relationships involving elementary functional groups with heteroatom central atoms is small, this weakness does not appear to be severe at present. The relationships that the nitro-,

[7] J. E. Dubois, *J. Chem. Documentation*, 1973, **13**, 8.

TABLE 2

Examples of elementary functional groups recognized by the program

Name	Elementary group	Name	Elementary group
Aldehyde		Hydrazine	
Ketone		Amine	
Alcohol		Imine	
Phenol		Urea	
Acid		Amide	
Ester		Nitro	
Lactone		Nitroso	
Ketal		Nitrile	
Acetal		Thiol	
Hemiacetal		Sulphide	
Ether		Disulphide	
Epoxide		Sulphoxide	
Peroxide		Sulphone	
Allene		Sulphonic acid	
Keten		Sulphonamide	
Chloride		Sulphinic acid	
Acid chloride		Phosphine	
Carbodi-imide		Phosphine oxide	
Isocyanate		Phosphinic acid	
Amine oxide		C–C double bond	
Oxime		C–C triple bond	
Hydroxylamine		Anhydride	
Azo		Imide	

* Denotes the central atom of the group.

nitroso-, hydroxylamine, and oxime groups might share with other groups present in the molecule are desired. Thus in each case, the first carbon atom bonded to the heteroatom central atom (nitrogen in this case) is labelled the central atom and employed in the discovery of α-, β-, *ortho-*, *meta-*, allylic, and benzylic relationships.

Table 2 also lists the elementary functional groups for which central atoms are not defined. These are carbon–carbon double and triple bonds and the anhydride and imide substructures. For these anomalous cases, specific queries concerning their existence must be made. This represents a departure from the normal procedure of not asking questions and so only in these few cases can negative answers result. Anomalous substructures are handled by partitioning the program into canonically ordered chloride, sulphur, phosphorus, oxygen, nitrogen, and carbon modules with specific and detailed tests for their presence instituted in the appropriate atom module. More than one group may share the same central atom, as in the case of a *gem*-dichloride, and such situations are dealt with in the normal context of our procedure.

Having detected a central atom, the program reports the elementary functional group to which it belongs. This is done by preparing the elementary group machine formula. The detection of a central atom and the formulation of the elementary group proceed simultaneously.

Description of Atom and Group Machine Formulas.—The machine formula of an elementary functional group is the sum of the atom machine formulae of the heteroatom neighbours of the central atom plus the atom machine formula of the central atom itself. Atom machine formulae are represented by the illustrated 16-bit row. Initially all bits are set to zero. Two bits are allocated to each column except for column **5** which has **4** bits. The presence of a single sulphur, nitrogen, phosphorus, oxygen, or chlorine atom is indicated by placing a non-zero number in columns **1**—**4** or **6**, respectively. The value of that number is determined solely by the unsaturation of the particular heteroatom. The atom machine formula of a heteroatom is then made complete by placing in column **7** a number equal to the number of hydrogen atoms directly attached to the heteroatom. Since only 2 bits are allocated to columns **1**—**4** and **6**, no heteroatom is allowed an unsaturation greater than three. Thus in our connection table description of a molecule containing the sulphonic acid group, all the heteroatoms of this group would be described as being saturated. This in no way hinders our ability to recognize and simulate reactions of this group. Furthermore, since we are constrained to 2 bits in each of columns **1**—**4** and **6**, the maximum number of any one kind of heteroatom that can be represented in a particular column is dependent on the unsaturation of that heteroatom. Thus the number **3** in column **4** may represent **3**

Bit pattern	SS	NN	PP	OO	CCCC	XX	HH
Column number	1	2	3	4	5	6	7

saturated oxygen atoms, as in the sulphonic acid group, or 1 unsaturated and 1 saturated oxygen atom, as in the ester group. Overlap of the bits between columns is possible and allowed but only in those cases where no ambiguities result. For example, consider the machine formula of the carbodi-imide group, the last entry in Table 3. This machine formula may also be interpreted as representing a group composed of a sulphur and a carbon atom, for example the thiol or the sulphide group. However, note that in these groups the sulphur atom is the central atom. Since the machine formula of an elementary group is the sum of the atom machine formulae of the heteroatom neighbours of the central atom plus the atom machine formula of the central atom itself, then these groups would have machine formulae different from that of the carbodi-imide group and no ambiguities would be created.

The preparation of the carbon atom machine formula consists of placing in column 5 the number 1 if the carbon atom is aliphatic and of any unsaturation, or the number 6 if the carbon atom is aromatic. We do not include the number of attached hydrogen atoms in the carbon atom machine formula.

TABLE 3

Examples of atom and group machine formulae

Structure	SS	NN	PP	OO	CCCC	XX	HH
Cl	00	00	00	00	0000	01	00
OH	00	00	00	01	0000	00	01
C≡N	00	11	00	00	0001	00	00
COCl	00	00	00	10	0001	01	00
–N=C=N–	01	00	00	00	0001	00	00

TCAFORM is an intermediate binary table consisting of 16-bit blocks accessed by the row number of the central atom of an elementary functional group. Initially all 16 bits are set to zero. The ith block in TCAFORM contains the group machine formula of an elementary functional group with central atom i. Indexing the table in this manner allows for very rapid formulation of both elementary and complex functional groups.

Modular Operations.—In a single top-down scan of the ordered connection table all the central atoms are discovered with simultaneous formulation of the elementary functional groups to which they belong. The atomic number (value in column 1) of the atom being investigated determines the particular atom module to be executed. Execution continues within this module until an atom of a different element is encountered, at which time control passes to the appropriate atom module. In this way the nitrogen module is executed only if N atoms are present in the molecule. In order to be systematic the atom modules are canonically ordered, but we do not proceed through them sequentially. Let us examine the operations implemented within each atom module.

(1) *The chlorine module.* Central atom i is the entry in column 13 of the current chlorine atom's row. Prepare and add the chlorine atom machine formula to the contents of the ith block in the table TCAFORM.

(2) *The sulphur module.* Central atom i is the current sulphur atom only if its row number is numerically smaller than the entry in column 13 of the current row. Otherwise central atom i is the entry in column 13. In this way we avoid repetitions as in disulphides and structures in which sulphur is directly bonded to phosphorus. Prepare and add the sulphur atom machine formula to the contents of the ith block in TCAFORM.

(3) *The phosphorus module.* The operations performed here are analogous to those in the sulphur module.

(4) *The oxygen module.* Central atom i is the entry in column 13 of the current oxygen atom's row, provided that the current oxygen is not itself a central atom. Prepare and add the oxygen atom machine formula to the contents of the ith block in TCAFORM. Perform specific tests for the presence of the anhydride substructure.

(5) *The nitrogen module.* Prepare the nitrogen atom machine formula. Central atom i is the entry in column 13 provided that the current nitrogen is not itself a central atom. The latter occurs in cases where the nitrogen is bonded to one or more oxygen atoms as in the case of the nitro-group. In such cases, central atom i is the immediate neighbouring carbon of the nitrogen. In addition, the nitrogen atom machine formula has added to it the formulae of the oxygen atoms to which it is bonded. If the current nitrogen atom has a positive charge, we arbitrarily add the number 4 to column 5 of its atom machine formula. This is done to reduce the number of isomeric formulae produced. Finally, we add the nitrogen atom machine formula to the contents of the ith block in TCAFORM and perform specific tests for the presence of the imide substructure.

(6) *The carbon module.* In order to facilitate the search for C–C double and triple bonds, the carbon atom module is segmented into unsaturation sub-modules. The unsaturation of the carbon atom currently being analysed determines the particular sub-modules to be executed so that we do not proceed sequentially through the sub-modules. In our notation, carbon atoms with unsaturation of 0, 1, 2 or 3 correspond to a saturated or aromatic or doubly bonded or triply bonded atom, respectively. A carbon atom with unsaturation of 4 describes an atom with two geminal double bonds, for example, the central carbon atom in an allene. The following are the operations performed within each sub-module.

(a) Carbon atoms with unsaturation of 4. Central atom i is the current carbon atom. Prepare and add the carbon atom machine formula to the contents of the ith block in TCAFORM.

(b) Carbon atoms with unsaturation of 3. Update a table of all carbon–carbon triple bonds. If the current carbon atom i is a central atom, prepare and add the carbon atom machine formula to the contents of the ith block in TCAFORM.

(c) Carbon atoms with unsaturation of 2. Update

a table of carbon–carbon double bonds and carbon–carbon double bonds which are members of six-membered rings. If the current carbon atom *i* is a central atom, prepare and add the carbon atom machine formula to the contents of the *i*th block in TCAFORM.

(d) Carbon atoms with unsaturation of 1. Make an addition to a table of all aromatic carbon atoms. If the current carbon atom *i* is a central atom, prepare and add the carbon atom machine formula to the contents of the *i*th block in TCAFORM.

(e) Carbon atoms with unsaturation of 0. If the current saturated carbon atom *i* is a central atom, prepare and add the carbon atom machine formula to the contents of the *i*th block in TCAFORM.

We exit from the atom modules when the last atom has been analysed. When execution of the atom modules has terminated, all the elementary functional groups have been detected and formulated. At this point, for the case of PGE_2, the group machine formulas of the carboxy-, ketone, and both alcohol functions have been prepared.

This method of formulation produces some ambiguities. Some pairs of functions such as ketone and aldehyde, or ketonic and aldehydic acetals, have isomeric machine formulae. By comparing the machine formulae produced in the atom modules with those known to produce ambiguities and subsequent testing of the central atom in question the program differentiates among the structures and readjusts the machine formulae accordingly. In this way the machine formulae of the elementary groups are all unique, fixed, and predictable. This in turn means that the machine formulae of substructures obtained by combinations of these elementary groups will also be fixed and predictable, although not necessarily unique. This predictability allows for discovery of special groups. For example, the substructures CH–C=O, a hydrogen α to a carbonyl group, and $CH–CO_2H$, a hydrogen α to a carboxy-group, can be deduced by comparing machine formulae produced above with those known for the carbonyl and carboxy-functions. When a match is made, subsequent examination for hydrogen on the carbon α to the central atom would reveal the existence or non-existence of the desired substructure.

Let us see why the application of the rules for central atom discovery to the anhydride and imide substructures leads to erroneous results. With regard to the component atoms of the anhydride substructure, the carbonyl oxygen atoms are the highest ranking atoms so that the oxygen module is the first to be executed. According to the instructions within this module, both carbonyl carbon atoms, say carbons *i* and *j*, would be selected as central atoms and the carbonyl oxygen atom machine formula is added to the contents of the *i*th and the *j*th block in TCAFORM, respectively. Furthermore, if carbon *i* has a higher priority than carbon *j*, then the row number of carbon *i* would appear in column 13 of the saturated oxygen atom's row. When we next visit the saturated oxygen atom's row, carbon *i* is again chosen as a central atom and the saturated oxygen atom machine formula is added to the contents of the *i*th block in TCAFORM. Thus, after formulation of the central atoms in the carbon module, the program would report the presence of two elementary functional groups, an ester whose central atom is carbon *i* and a ketone whose central atom is carbon *j*. In addition, when the procedure for complex functional group discovery is later invoked, the presence of an α-keto-ester group would be reported. In the case of the imide substructure the program would report the discovery of the amide, ketone, and α-keto-amide groups. These results are clearly misleading so that specific tests are carried out to detect the presence of the anhydride and imide substructures. It is interesting to see why ureas are correctly perceived. Within the oxygen module, the carbonyl carbon is selected as a central atom. But since this carbon has a higher priority than all other carbon atoms in the substructure, its row number appears in column 13 of each of the nitrogen atoms' rows. Thus when the nitrogen module is subsequently executed, this same carbonyl carbon is selected as the central atom and the formulation of the ureas proceeds in the correct manner.

Data Organization.—FXNLIST is the list of all functional groups, both elementary and complex, present in the molecule and comprises the output of the program. FXNLIST is divided into 20-character blocks, each being sequentially populated with each new functional group the program discovers. Each block contains both the machine formula of the group and its component atoms in the molecule. Schematically we may represent each block as:

Character	0	1	2	3	X1	5	6	7	8
X2	10	11	12	13	X3	15	16	17	18	19

Initially each character contains a blank. Characters 0—1 are initially left blank while characters 2—3 contain the machine formula of the functional group. Characters 4—19 contain the row numbers of the atoms that comprise that group. In the case of elementary functional groups, character 4, named X1, has the group's central atom while characters 5—8 list the adjacent atoms of the central atom in the same canonical order in which they appear in the connection table. With this canonical ordering we can predict with certainty which character refers to which atom of the group. Let us take the carboxy-group of PGE_2 as an example and refer to Table 4 which lists the contents of each FXNLIST block. There X1 contains 5, the row number of the carbon central atom of the carboxy-group, character 5 has the row number of the unsaturated oxygen 1, the double bond to this oxygen results in a repetition in character 6, character 7 has the hydroxy-oxygen atom 2, and character 8 the carbon atom 18 α to the carboxy-group. Characters 5—8 thus correspond to columns 13—16 of row number 5 of the connection table. Knowledge of the kind of functional group, its location, and also its component

atoms is thus obtained. The last is vital since our program subsequently accesses reactions which directly affect these same atoms. Characters **9**—**19** are used in representing special and complex functional groups and their use will become clearer in the next section.

Discovery of the Complex Functional Groups.—Complex functional groups are the result of combinations of two elementary groups. For example, an α-hydroxy-ketone is a complex functional group composed of two elementary groups, hydroxy and ketone, sharing an α-relationship. The discovery of complex functional groups is necessary as they often undergo reactions different from those of the elementary groups of which they are composed. α- and *ortho*-Relationships among elementary groups are deduced as follows. Elementary functional group 1 with central atom i and elementary group 2 with central atom j possess an α-relationship if atom j is one of the adjacent atoms of i. Thus if atom i is a central atom we proceed to the ith row of the connection table and examine the entries in columns **13**—**16**. If none of the adjacent atoms of i is itself a central atom, then group 1 does not bear an α- or *ortho*-relationship to any other group. If, however, some adjacent atom j is a central atom, then α- or *ortho*-relationships exist. Group 1 is involved in more than one α- or *ortho*-relationship if more than one of its adjacent atoms are themselves central atoms. The machine formula of the complex group so discovered is obtained by adding the machine formulae of the elementary groups contained in the ith and jth blocks in TCAFORM and is subsequently stored in the first available block in FXNLIST. If the machine formula of elementary group 1 is numerically smaller than that of group 2, then X1 has i followed by the row numbers of its adjacent atoms while X2 has j followed by the row numbers of its adjacent atoms. Characters **14**—**19** are left blank. For α-hydroxy-ketone, X1 has the hydroxy while X2 has the carbonyl central atom, respectively. Again no uncertainty is left as to which entry in FXNLIST refers to which atom in the structure. In PGE_2 no α- or *ortho*-relationships were discovered. Elementary groups 1 and 2 possess a β- or *meta*-relationship if there exists an atom k such that central atoms i and j are both among the adjacent atoms of k. That is, we inspect columns **13**—**16** of the kth row of the connection table. If two of its adjacent atoms, say i and j, are themselves central atoms, then a β- or *meta*-relationship exists between groups 1 and 2 with atom k as the intermediate atom. Clearly more than one β-relationship can share the same intermediate atom. The FXNLIST block is constructed in a manner similar to that for the α-relationships except that instead of a blank, X3 has k followed by the row numbers of the adjacent atoms of k in characters **15**—**18**. Since the machine formulae of α- and β-hydroxy-ketones are the same, this provides us with an easy means of distinguishing between the two.

TABLE 4

Contents of FXNLIST for the PGE_2 molecule

		FXNLIST blocks															
Structure	Machine formula	X1	5	6	7	8	X2	10	11	12	13	X3	15	16	17	18	19
H H | | –C=C–	22	7	8	8	13		8	7	7	11							
H H | | –C=C–	22	9	10	10	16		10	9	9	17							
C–OH	111	11	4	8	20												
C–OH	111	12	3	13	15												
H H | | –C=C–C–OH	133	11	4	8	20		8	7	7	11		7	8	8	13		
C=O	210	6	0	0	14	15											
CH–C=O	221	14	6	13	16		6	0	0	14	15						
CH–C=O	221	15	6	12			6	0	0	14	15						
O || –C–OH	311	5	1	1	2	18											
OH | –C–C–C=O	321	12	3	13	15		6	0	0	14	15	15	6	12			
O || CH–C–OH	322	18	5	19			5	1	1	2	18						

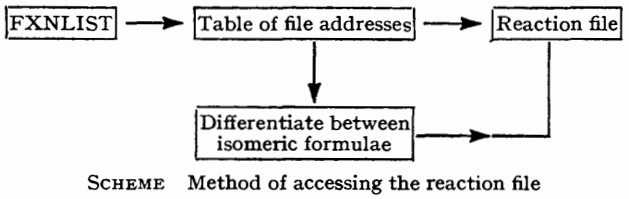
In its next step our program deals with carbon–carbon double bonds and assigns machine formulae according to the total number of hydrogen atoms on both carbons and to their disposition among them. Carbon–carbon double bonds in six-membered rings are given a special machine formula and characters **4**—**9** of the FXNLIST block carry the members of that ring in a fixed order. If atoms i and j are the terminal atoms of a carbon–carbon double bond and there exists a central atom k external to the double bond, which is one of the adjacent atoms of atom i (or j), then an allylic relationship exists between the double bond and the functional group with central atom k. The formula of the allylic structure is the sum of the formulae of the carbon–carbon double bond and the external functional

group. In the FXNLIST block, X1 has k, X2 has the carbon of the double bond directly bonded to k, and X3 the remaining double-bond carbon.

In its next step, our program formulates carbon–carbon triple bonds if any are present. Starting from the aromatic carbon of highest rank, benzylic relationships are deduced. If atom i is aromatic and devoid of hydrogen atoms and atom j is a non-aromatic central atom, then a benzylic relationship exists between the aromatic ring containing i and the external functional group with central atom j provided that j is one of the adjacent atoms of i. Aromatic carbons with hydrogens are considered here as functional groups and formulae are subsequently prepared. The program has now concluded the discovery and formulation of all the elementary and complex functional groups present in the molecule.

Since all α-, β-, *ortho-*, *meta-*, allylic, and benzylic type combinations are indiscriminately produced, a check must be made as to the validity, that is the synthetic value, of these substructures. To this end a bit table is employed. The ith bit is set to 1 if the group with machine formula i is valid. Thus each entry in FXNLIST is checked against the bit table and deletions are made if necessary. The entries in FXNLIST are then sorted out with respect to the numerical values of the machine formulae. In this way like functional groups scattered about the molecule are clustered together in FXNLIST. This is necessary since reactions which produce such groups may affect all occurrences of the same. The routine has now completed its task. Table 4 lists the output of the program for PGE_2. The machine formulae are in hexadecimal notation.

Our own interests lie in the generation of synthetic pathways by computer. At each synthetic step it is necessary to recognize the functional groups and perform the reactions which transform or give rise to these groups. The large and ever increasing number of reactions in which any one group can participate presents us with a major problem of maintaining and accessing a massive and growing reaction file. The overall method we employ in indexing such reactions in the reaction file is shown in the Scheme. Part of the machine formula of the functional group under consideration is used to calculate a displacement in a table of reaction file addresses. These machine addresses point to the actual position in the reaction



SCHEME Method of accessing the reaction file

file where the first reaction that transforms or gives rise to the functional group in question is described. We then proceed to this address and perform the reaction.

The table of reaction file addresses is structured as follows. It consists of 256 blocks, each 64 characters long. Each block is in turn divided into 8 segments, each 8 characters long. Schematically we may represent each block as:

Character	0	1	2	3	4	5	6	7
	0	1	2	3	4	5	6	7
	0	1	2	3	4	5	6	7
	.							
	.							
	.							

Within each segment, characters 0—3 have the same function as characters 0—3 in each FXNLIST block, that is, characters 0—1 are left blank while characters 2—3 contain the group machine formula. Furthermore, within the same block, the content of character 2 is the same for each segment. That is, we record within a given block all those functional groups, if any, which have machine formulae equivalent with respect to the content of character 2. A maximum of 8 such possibilities is currently allowed; hence there are 8 segments per block. The content of character 2 can range from 0 to decimal 255, hence the need for 256 blocks. The table of reaction file addresses is thus sparsely populated but the efficiency that results from its use more than compensates for the storage wasted. Normally characters 4—7 of each segment contain the machine address of the actual position in the reaction file where the first reaction of the particular group is described.

The structuring of the table of reaction file addresses in this manner allows for quick movement from the FXNLIST block to the reaction file. For example, suppose we were interested in simulating the reactions of a given group described in a FXNLIST block. Furthermore, suppose that the contents of characters 2—3 of the FXNLIST block were decimal i and j, respectively. The following procedure is used to gain access to the reactions of this group. Multiplying i by 64 produces $64i$, which is the displacement to the ith block in the table of reaction file addresses. Each segment in the ith block contains all those groups which have machine formulae such that the contents of character 2 is equal to i. We then examine each segment for the one in which the content of character 3 is equal to j. When a match is made, characters 4—7 give us the address of the desired reaction and we proceed there.

Since occasionally more than one complex functional group has the same machine formula, it is necessary to differentiate among those groups where ambiguities exist. Since the machine formulae of the elementary functional groups are known and fixed, those complex groups which have isomeric machine formulae are thus predictable and the Scheme indicates how use is made of this fact. In the table of reaction file addresses, those file addresses which normally would be accessed by isomeric machine formulae are replaced by an address to another table where specific tests are done to differentiate among the conflicting complex groups. We

distinguish addresses which point to the reaction file from those which point elsewhere by placing a flag in characters 0—1 of the appropriate segment. The absence of a flag indicates that the address specified is a reaction file address.

Finally we comment on the structure of the reaction file itself. The file is a compact and unblocked listing of the reactions. The reactions are operations to be performed on the connection table representation of the molecule and are listed in a sequential manner. For any given group, the reactions which transform or give rise to it may be described at randomly distributed points in the file. As part of the description of each reaction, we include a displacement parameter which is the displacement to the next reaction of this group in the file. If the displacement is zero then the reaction under consideration is either the last or the only reaction known to the program. In this way a new reaction is simply added to the end of the file, its displacement parameter is set equal to zero, and we readjust the displacement parameter of the parent reaction of this group, if any. The displacements are calculated and put in place by the machine.

In summary then, we go first from the machine formula in the FXNLIST block to the appropriate block in the table of reaction file addresses. We next search among the segments of this block until a match is made between the machine formula described in the segments and that in the FXNLIST block. When a match is made, the flag field is checked. If the flag is absent, we have the address of the first reaction of this group in the file and proceed there and perform the reaction. If the displacement parameter of the reaction is zero, then no more reactions of this group are known to the program. If the displacement parameter is non-zero we add it to the reaction file address and proceed to the next reaction. If when we test the flag field, the flag is present, we then proceed to an area where the conflicting structures are resolved. When this has been done we have the address of the first reaction of the group in the reaction file and proceed as described above.

The entire program was written in the IBM 370 assembly language and executed on an IBM 370/165 machine. The execution time for a single store instruction on this computer is on the average 0·32 μs. Run times of 0·59 ms for PGE_2 and 1·45 ms for the complex polyfunctional molecule tetracycline indicate the efficiency of the procedure. Our times include the time required to sort FXNLIST, the table of all the elementary and complex functional groups discovered in the molecule.

This work was supported by the National Research Council of Canada, from whom one of us (A. E.) acknowledges the receipt of a scholarship.

[4/908 *Received*, 7*th May*, 1974]