et at the Graph Theory Conference in Kalamazov. Mich 2-1972. To appear in the proceedings.

SOME RECENT RESULTS IN CHEMICAL ENUMERATION.

Ronald C. Read

University of Waterloo

Abstract

A method is described for enumerating all acyclic chemical compounds and radicals consisting of carbon and hydrogen alone, classified according to the number of carbon atoms, and the numbers of double and triple bonds. This work extends enumerations previously carried out by Cayley, Polya, and others. Some numerical results are given.

Section 1. Preliminaries.

Before outlining the background and scope of the problem that we shall consider, it will be as well to review briefly the basic For our purposes we can regard a facts about chemical compounds. molecule of a chemical compound as an assemblage of atoms in which some atoms are linked to others by 'valency bonds'. These bonds may be single, double or triple (other kinds will not concern us). A structural formula is a method of representing the way in which the atoms in a molecule are linked together. In a structural formula

each atom is represented by a symbol, usually the initial letter or two of its name, e.g. C for carbon, H for hydrogen. A single bond is represented by a line drawn between the symbols for the two atoms that it links, a double bond by a double line, and a triple bond by a three-fold line. Examples of structural formulae are given in figure 1. Note that a structural formula makes no attempt to indicate how the atoms are situated in space relative to each other; it simply indicates which atoms are linked to which, and by what sort of valency bond. Thus a structural formula, from our point of view, is essentially a multigraph in which the nodes are of several different kinds. In this paper we shall be concerned only with hydrogen.

The valency of an atom in a molecule is the number of bonds by which it is linked to other atoms, double and triple bonds counting as 2 and 3 respectively. Carbon atoms have valency 4, and hydrogen atoms have valency 1.

An acyclic compound is one for which the structural formula, regarded as a graph, has no cycles, i.e. is a tree. The term "cycle" is here used in its usual graph-theoretical sense, on the understanding that double and triple bonds are to be regarded as single edges of the graph. Thus, although for many purposes a double edge in a multigraph is regarded as forming a cycle of length 2, we shall not so regard a double bond. The compounds (a) and (b) in figure 1 are therefore acyclic, while (c) is a cyclic compound.

By a radical we shall mean an incomplete molecule -incomplete in the sense that there is exactly one valency bond

(single, double or triple) one end of which does not have an atom. Thus a radical is not a molecule, but something from which molecules can be constructed by placing an atom, or another radical, at the end of this <u>free</u> bond. If the free bond is single, double or triple, then the radical is said to be monovalent, divalent or trivalent respectively. Figure 2(a) shows the isopropyl radical, having a free single bond. If we put an iodine atom at the end of the free bond, we get the isopropyl iodide molecule of figure 1(a). If, instead, we put another isopropyl radical there we shall get the hydrocarbon of figure 2(b).

It is easily verified that if an acyclic compound consists entirely of carbon and hydrogen, and has no double or triple bonds, then its general formula will be C_nH_{2n+2} . That is, if the number of carbon atoms is n, then the number of hydrogen atoms is 2n+2. Such compounds used to be called paraffins, but are now usually called "alkanes", and the corresponding radicals are called "alkyl radicals". Thus figure 2(a) shows an alkyl radical; figure 2(b) shows an alkane.

The above discussion of nomenclature and notation is greatly oversimplified, and would not satisfy a professional chemist.

However, it will be sufficient for our needs.

A problem that has provided a great stimulus to the development of many tools and results in Combinatorial Analysis is that of enumerating all chemical compounds with formulae of a given kind. Problems of this type were studied by Cayley [2, 3] who succeeded in enumerating the alkyl radicals and the alkanes.

Cayley's results were later extended by other investigators, notably Blair and Henze [4, 5, 6], and Perry [12]. Enumerative results for other types of chemical compounds (mostly acyclic) were obtained, but the methods whereby this was done remained similar to those used by Cayley. A new tool for working with problems of this type was forged by Pólya; it is the "Hauptsatz" of his 1938 paper [13]. The use of this fundamental theorem enables the findings of Cayley, Blair and Henze, etc. to be proved much more easily and concisely than by the original methods, and applications to some kinds of cyclic compounds (benzene derivatives, etc.) became possible.

One somewhat unsatisfactory aspect of the above work was the transition from the alkyl radicals to the alkanes. It can be seen from figure 2(b) that the enumeration of alkanes is essentially the enumeration of a type of tree; but in an alkyl radical one of the carbon atoms, namely the one incident with the free bond, is different from the others, so that the enumeration of the alkyl radicals is an enumeration of rooted trees. The transition in question is therefore essentially that of going from the enumeration of rooted trees to the enumeration of unrooted trees, and at the time of Pólya's paper this could be achieved only by somewhat ad hoce methods. A better method of making this transition was published

by Otter [10], and was extended by Harary and Norman [9].

In this paper we shall use all available up-to-date methods in order to enumerate acyclic hydrocarbons with the general formula C_nH_m . We have already remarked that for the alkanes, m=2n+2. If m<2n+2 then there must be double or triple bonds present in the molecule. We shall therefore enumerate these compounds according to the number n of carbon atoms that they contain, the number d of double bonds and the number t of triple bonds. Blair and Henze enumerated the ethylene derivatives (having exactly one double bond) and, with Coffman, the acetylene derivatives (having exactly one triple bond), but the problem of enumerating acyclic hydrocarbons with any numbers of double and triple bonds does not seem to have been studied before. Thus this enumeration represents an extension of results already obtained, and the accuracy of this work can be checked, in part, by comparison with these previously known results.

Section 3. Hydrocarbon radicals.

In this section we shall enumerate the hydrocarbon radicals, which are of three types, according to the nature of the free bond. The radicals of each of the three types will be classified by the number of carbon atoms, and the numbers of double and triple bonds.

We shall let $G_{n,d,t}$ denote the number of monovalent radicals having n carbon atoms, d double bonds and t triple bonds. that we do not need to specify the number of hydrogen atoms, since it is easily verified that this will be 2n + 1 - 2d - 4t. similarly define $H_{n,d,t}$ to be the number of divalent radicals (in the sense that the free bond is double) and $I_{n,d,t}$ the number of trivalent radicals. These types of radicals are depicted diagrammatically in figure 3, and will be called, for brevity, G-radicals, H-radicals and I-radicals. In figure 3 each shaded "balloon" denotes the rest of the radical apart from the free bond and the carbon atom to which it is attached. Note that the free double or triple bond is included in the count for d or t, as the case may be.

For each of these three types of radical we define a counting series, as follows:

(1)
$$G(x,y,z) = \sum_{n} \sum_{d} G_{n,d,t} x^{n} y^{d} z^{t}$$

(1)
$$G(x,y,z) = \sum_{n} \sum_{d} G_{n,d,t} x^{n} y^{d} z^{t}$$

(2) $H(x,y,z) = \sum_{n} \sum_{d} E_{n,d,t} x^{n} y^{d} z^{t}$

(3)
$$I(x,y,z) = \sum_{n=1}^{\infty} \sum_{t=1}^{\infty} I_{n,d,t} x^{n} y^{d} z^{t}$$

If we examine the G-radicals in closer detail we see that they are of three types, as shown in figure 4. G-radicals of type (a) can be enumerated by means of Polya's theorem. We have three positions - the three balloons - each of which can be occupied by a G-radical, or possibly by a hydrogen atom. These three positions can be permuted by any permutation of the symmetric group S_3 . Using Pólya's theorem, with figure-counting series G(x,y,z) and group S_3 , we obtain the configuration counting series

(4) $xZ(S_3; G(x,y,z))$

in the notation of Polya's paper; or rather, we would if the possibility of a G-radical being a single hydrogen atom were included in G(x,y,z). It is therefore convenient to adopt this convention, and regard the radical "H-" as a G-radical, with n=0. Expression (4) is then valid. Note the multiplier "x" in (4), which is needed to account for the carbon atom to which the three radicals are attached.

In figure 4(b) we have two dissimilar positions into one of which a G-radical must be placed, while an H-radical must be placed in the other. Hence radicals of this type are enumerated by $x \ G(x,y,z).H(x,y,z).$

In figure 4(c) there is only one position, in which must be placed an I-radical. Hence these radicals are enumerated by

(6) x I(x,y,z).

Combining these results, and adding a term "1" for the hydrogen radical, we obtain the following equation.

(7) $G(x,y,z) = 1 + x[Z(S_{3};G(x,y,z))+G(x,y,z)H(x,y,z)+I(x,y,z)].$

Looking more closely at H-radicals we see two types, as in figure 5(a) and 5(b). In figure 5(a) we have two interchangeable positions in which to place a G-radical. Polyals theorem then gives

(8) $xy Z(S_2; G(x,y,z))$

as the counting series for these, the factor xy being included to account for the free double bond and the incident carbon atom.

In figure 5(b) we have a single position in which to put an H-radical. This gives us the counting series xyH(x,y,z). From this, and (8), we derive

(9)
$$H(x,y,z) = xyZ(S_2; G(x,y,z)) + xyH(x,y,z)$$

(10)
$$H(x,y,z) = \frac{xy}{1-xy} Z(S; G(x,y,z))$$

I-radicals are of one type only (figure 5(c)), and hence are given by the counting series xzG(x,y,z). Thus

(11)
$$I(x,y,z) = xz G(x,y,z).$$

We now substitute for H and I, from (10) and (11) in equation (7), and obtain

(12)
$$G(x,y,z) = 1 + x[Z(S_3;G(x,y,z)) + \frac{xy}{1-xy}G(x,y,z).Z(S_2;G(x,y,z)) + xz G(x,y,z)]$$
.

Equation (12) enables G(x,y,z) to be calculated recursively up to terms in a given power of x. The calculation is extremely tedious by hand, but not too difficult on a computer. Results for $n \le 10$ are given in Table I.

As a check on these results, we note that if we put y = z = 0 we eliminate all radicals having double or triple bonds, and are left with the alkyl radicals alone. If we write r(x) for G(x,0,0) and perform the indicated substitution in Z(S), equation (12) becomes $r(x) = 1 + \frac{1}{6}x \left[r^3(x) + 3r(x)r^4(x) + 2r(x^3)\right]$ which is the functional equation given by Polya for the alkyl radicals (see [13], page 150).

Section 4. Hydrocarbons with a distinguished carbon atom.

As stepping-stones toward our goal of enumerating hydrocarbons we need two subsidiary results. The first of these is the enumeration of hydrocarbons in which one carbon atom has been distinguished from the others. We shall denote this distinguished carbon atom by C*.

We readily see that these compounds are of four types, shown in figure 6. Each of these types can be enumerated using Polya's theorem. We find that those in figure 6(a) are enumerated by

(14) $\times Z(S_{4}; G(x,y,z));$

those in figure 6(b) by

(15) $x Z(S_2; G(x,y,z)).H(x,y,z);$

those in figure 6(c) by

(16) $\times Z(S_2; H(x,y,z));$

and those in figure 6(d) by

(17) x G(x,y,z).I(x,y,z).

Adding these four results we obtain the counting series for the hydrocarbons with a distinguished carbon atom. We shall denote it by K(x,y,z). Hence we have

(18)
$$K(x,y,z) = xZ(S_{4};G(x,y,z)) + xZ(S_{2};G(x,y,z)).H(x,y,z) + xZ(S_{2};H(x,y,z)) + xG(x,y,z).I(x,y,z)$$

from which K(x,y,z) can be computed, since everything else is known.

Section 5. Hydrocarbons with a distinguished carbon-carbon bond.

Our second subsidiary result is the enumeration of hydrocarbons in which one particular carbon-carbon bond has been distinguished from the rest. If this bond is single, then our compound looks like figure 7, and is clearly the result of joining two monovalent radicals. Since the distinguished bond is to join two carbon atoms, neither of these radicals can be just a hydrogen atom. We therefore have a Polya type enumeration problem, with two interchangeable positions, into each of which must be put a radical enumerated by $G_1(x,yz) = G(x,y,z) - 1$. These hydrocarbons are therefore enumerated by

(19)
$$Z(S_2; G(x,y,z)) = \frac{1}{z} [G^2(x,y,z) + G(x^2,y^2,z^2)].$$

For reasons that will become apparent, we shall be more interested in those hydrocarbon of this kind for which the two radicals are <u>dissimilar</u>. The counting series for these can be found either by subtracting from (19) the counting series for hydrocarbons for which these radicals are identical, or by using a modification of Pólya's theorem (for which see [13], page 172). In either case the counting series turns out to be

(20)
$$\frac{1}{2} \left[G_{1}^{2}(x,y,z) - G_{1}(x^{2},y^{2},z^{2}) \right].$$

If the distinguished bond is double, then a similar argument gives us

(21)
$$\frac{1}{2y} \left[H^{2}(x,y,z) - H(x^{2},y^{2},z^{2}) \right]$$

for those compounds for which the two radicals are dissimilar. The corresponding result for a triple bond is

(22)
$$\frac{1}{2z} \left[I^{2}(x,y,z) - I(x^{2},y^{2},z^{2}) \right] .$$

The division by y and z in (21) and (22) results from the fact that in putting the two radicals together we convert what were previously $\underline{\mathsf{two}}$ bonds into one.

Section 6. Acyclic hydrocarbons.

We shall now tie together these apparently disconnected results, and obtain the counting series for general acyclic hydrocarbons that we are seeking. To do this we need a theorem, given in [9], which we now briefly describe.

Let T be a tree, and u and v two of its nodes. If there is an automorphism of T which maps u onto v we say that u and v are equivalent. This relation is an equivalence relation, and the nodes of T are therefore partitioned into equivalence classes. Let p* be the number of these classes.

An automorphism of T induces, in an obvious way, a mapping of the set of edges of T onto itself, and we can define equivalence between edges a and b if a maps onto b in some such mapping. The edge set of T is thus partitioned into a number, q* say, of equivalence classes. The theorem is then that

(23)
$$p* - (q* - s) = 1,$$

where s=1 if T has a symmetric edge, and s=0 otherwise. An edge uv is symmetric if there is an automorphism which maps u onto v and v onto u. Clearly the "half-trees" at the ends of a symmetric edge must be identical, and a tree can have at most one symmetric edge.

 $\label{eq:continuous} \mbox{If we sum (23) over all trees with a given number p of nodes} \\ \mbox{we obtain}$

$$(24) \qquad \sum_{p^*} - (\sum_{q^*} - \sum_{s}) = \sum_{s} 1.$$

Now p* is the number of distinct ways (to within automorphism) of choosing a node of T, and is therefore the number of trees with a distinguished node that can be obtained from T. Thus $\sum p*$ is the total number of distinct trees rooted at a node. By similar

reasoning $\sum q *$ is the total number of distinct trees having one edge distinguished from the others.

The term $\sum s$ in (24) is the number of trees having a symmetric edge, and the term $\sum 1$ on the right-hand side is simply the total number of trees. Hence this latter number can be computed if $\sum p*$, $\sum q*$ and $\sum s$ are known.

For simplicity this argument has been given in terms of ordinary graph-theoretical trees, but it holds just as well for the hydrocarbons that wehave been discussing. For any given n, d and t the term $\sum p^*$ gives the number of hydrocarbons with a distinguished carbon atom, and this is the coefficient of $x^ny^dz^t$ in K(x,y,z) of equation (19). The term $\sum q^* - \sum s$ will be the number of hydrocarbons with a distinguished carbon-carbon bond minus the number of these with a symmetric bond. We have anticipated this subtraction in finding (20), (21) and (22). The sum of these three expressions will be a counting series in which the coefficient of $x^ny^dz^t$ is $\sum q^* - \sum s$.

Finally, let us denote by L(x,y,z) the counting series in which the coefficient of $x^ny^lz^l$ is the number of hydrocarbons that we are seeking, having n carbon atoms, d double bonds, and t triple bonds. This coefficient is thus the $\sum 1$ of (24). Putting together the results of this section we have

(25)
$$L(x,y,z) = K(x,y,z) - \frac{1}{2}[G_1^2(x,y,z) - G_1(x^2,y^2,z^2)] - \frac{1}{2}[H^2(x,y,z) - H(x^2,y^2,z^2)] - \frac{1}{2}[I^2(x,y,z) - I(x^2,y^2,z^2)].$$

The coefficients in L(x,y,z) have been calculated for n \leq 10 and are exhibited in Table II.

Section 7. Deductions and verification.

It is of interest to see briefly how (25) links up with previously known results. The coefficient of y in L(x,y,0) will enumerate compounds having just one double bond, i.e. the alkyl derivatives of ethylene. From Table II we extract the information that it is

(26) $x^{2} + x^{2} + 3x^{4} + 5x^{4} + 13x^{6} + 27x^{7} + 66x^{8} + 153x^{9} + 377x^{10} + \dots$ have

The coefficients in (26) agree with the numbers of ethylene derivatives as given by Blair and Henze [6] who computed them by a completely different method.

Similarly the coefficient of z in L(x,0,z) will enumerate the alkyl derivatives of acetylene (having just one triple bond). From Table II we see that this is $(27) \quad x^2 + x^3 + 2x^7 + 3x^5 + 7x^6 + 14x^7 + 32x^8 + 72x^9 + 171x^7 + \dots$ in which the coefficients agree with the computations of Coffman, Blair and Henze [8].

Another interesting result, not previously obtained, is that for the number of hydrocarbons with a given number n of carbon atoms, irrespective of the number of double and triple bonds. This number is the coefficient of x in L(x,1,1) and, from Table II, we have $(29) \quad L(x,1,1) = x + 3x^{2} + 4x^{2} + 12x^{2} + 27x^{2} + 84x^{6} + 247x^{7} + 826x^{7}$

 $11 + 2777x^9 + 9868x^6 + ...$

74 2986

XX

Section 8. Prospects.

One naturally asks whether results such as these could be extended further, or generalized. One could consider, for example, removing the restriction to acyclic compounds, but if one does this one is beset by a host of difficulties. The problem is then essentially that of enumerating graphs or multigraphs with a given partition, and although methods have been published for doing this (see [7], [11]) they are all extremely cumbersome and, in general, quite impracticable. The prospect in this direction is not very bright, though some special cases may be amenable to treatment.

One can, instead, retain the restriction to acyclic compounds but allow atoms other than carbon or hydrogen. Oxygen is the obvious next candidate, and I have, in fact, carried out an enumeration, similar to that given here, for acyclic compounds of carbon, oxygen and hydrogen, having given numbers of these atoms, and given numbers of single, double and triple bonds. The results are complicated, tedious and uninspiring, and, what is more, they suffer from a grave defect.

The molecules enumerated by L(x,y,z) are all more or less chemically feasible, and we can therefore cherish the illusion that what we have done in this paper has at least a tenuous connection with real life. When we include oxygen atoms, all such illusions are shattered. For example, chains of any number of oxygen atoms, of the form

are theoretically possible, whereas chains of more than three oxygen atoms are not chemically feasible. This difficulty can be overcome

(at the expense of making the results even messier) but there are other snags. Take, for instance, the hydroxyl radical "-O-H". There is no obvious reason why there should not be two such radicals attached to one and the same carbon atom; yet such a configuration is not chemically possible. Considerations such as these make it virtually impossible to derive a series which will count only those carbon-oxygen-hydrogen compounds that are chemically significant; yet any enumeration that includes impossible compounds is pretty much a waste of time. There seems to be little scope in this direction either.

One related problem that is not entirely pointless is that of enumerating molecules taking into account something of the way that they are situated in space. If this is done to the extent of recognising differences between left-handedness and right-handedness we have the problem of enumerating stereo-isomers. This problem was discussed by Polya for the alkanes, and I have extended it to general hydrocarbons, much as in this paper. The results of this enumeration, and others, are to be included in a chapter of a book [1] at present in course of preparation.

References

- Balaban, A.T. Chemical Applications of Graph Theory.
 Academic Press. (To appear).
- Cayley, A. On the mathematical theory of isomers.
 Phil. Mag. 47(1874) 444 446.
- 3. ----- On the analytical forms called trees, with applications to the theory of chemical combinations.

 Report of the British Association for the Advancement of Science. (1875) 257 305.
- 4. Blair, C.M., Henze, H.R. The number of structurally isomeric alcohols of the methanol series.

 J. Amer. Chem. Soc. 53(1931) 3042 3046.
- The number of isomeric hydrocarbons of the methane series.J. Amer. Chem. Soc. 53(1931) 3077 3085.
- isomeric hydrocarbons of the ethylene series.

 J. Amer. Chem. Soc. 55(1933) 680 686.
- 7. Cadogan, C.C. On multigraphs with a given partition.
 Bull. Australian Math. Soc. 3(1970) 125 137.

- 8. Coffman, D.D., Blair, C.M., Henze, H.R. The number of structurally isomeric hydrocarbons of the acetylene series.

 J. Amer. Chem. Soc. 55(1933) 252 253.
- 9. Harary, F., Norman, R. Dissimilarity characteristic theorems for graphs.
 Proc. Amer. Math. Soc. 11(1960) 332 - 334.
- 10. Otter, R. The number of trees.
 Ann. Math. 49(1948) 583 599.
- 11. Parthasarathy, K. Enumeration of graphs with given partition.

 Canad. J. Math. 20(1968) 40 47.
- Perry, D. The number of structural isomers of certain homologs of methane and methanol.J. Amer. Chem. Soc. 54(1932) 2918 2920.
- 13. Polya, G. Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen.

 Acta Math. 68(1938) 145 254.

Table I. Monovalent acyclic radicals.

*n=9	0	1	t 2	3	4	n	= 10	! 	0	1	2	t 3	4	5
0 1 2 3 d 4 5 6 7	35 153 328 367 253 100 28 4	202	69 157 120 29 3	28 29 4	5		0 1 2 3 d 4 5 6 7 8		75 377 901 1196 964 491 160 36	171 718 1245 1074 498 124 17		88 138 53 4	20 8	1

Table II. Acyclic hydrocarbons.

(b) Hexadienyne-1,5,3. (c) Methyl benzene.

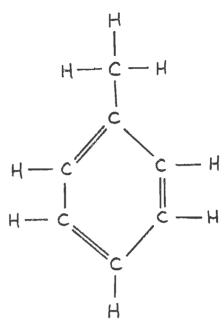


Figure 1.

(a) Isopropyl radical.

(b) 2,3-dimethyl butane.

an alkyl radical Figure 2.

an alkane

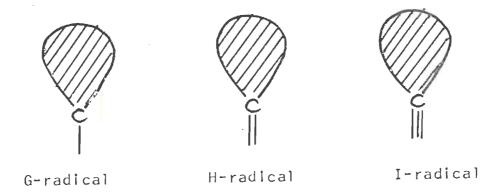


Figure 3. The three kinds of radicals.

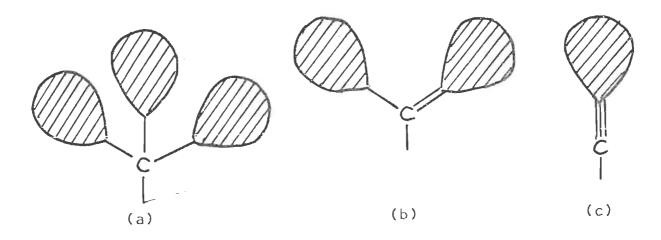


Figure 4. Types of G-radicals.

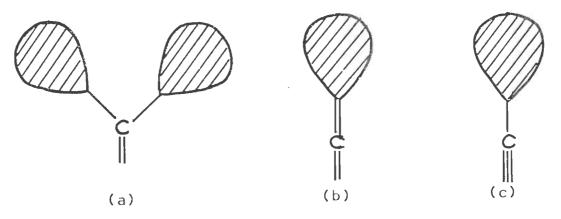


Figure 5. Types of H- and I-radicals.

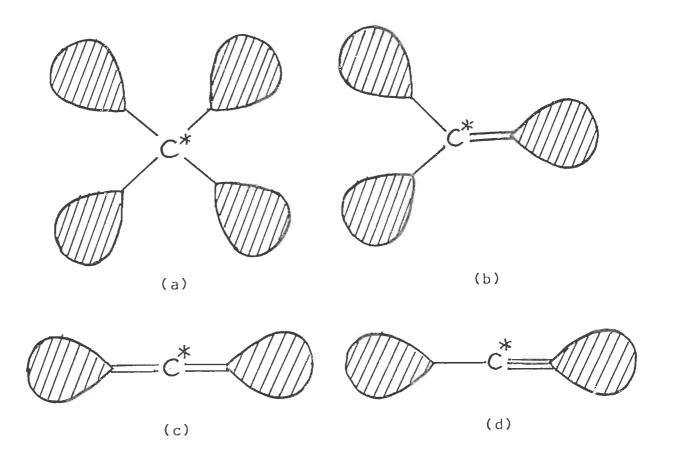


Figure 6.

Types of hydrocarbon with a distinguished carbon atom.

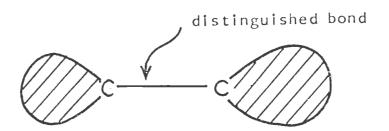


Figure 7.