Computer-Assisted Topological Analysis and Completion of Chemical Reactions

C. BEHNKE and J. BARGON*

Institute of Physical Chemistry, University of Bonn, Wegelerstrasse 12, D-5300 Bonn 1, West Germany

Received January 29, 1990

The concept of connectivities, which has been used before for the analysis of machine-readable chemical reactions, has been extended by a generic differentiation of the concept of the "reactive site" to allow an exact determination of the reactive centers, bonds, and groups contained therein by using an algorithm which is even tolerant of nonstoichiometric representation of chemical reactions, and which can, therefore, be applied also to the reaction files contained in a variety of the popular commercial databases. This new approach achieves a stoichiometric completion of chemical reactions for the first time, and thereby it provides a key prerequisite for the application of numerous previously available but incompatible programs. The highly discriminating new method outlined here in detail determines connectivities of first (and eventually of higher) order starting from hybridization patterns as chemically meaningful information but keeping track in addition of the abundance of individual chemical elements in the reactions. The integer connectivities are modified utilizing an atom index and real connectivity numbers to assure exact comparison while preventing a loss of part of the maximum common substructure, via a negative value which is assigned to the connectivities where appropriate.

INTRODUCTION

Databases containing chemical information have achieved considerable volume and thereby significance.[1,2] This development has rendered computer-assisted searching more attractive, but at the same time more difficult. Whereas searching in such databases, which contain textual information (e.g., the CAS CA file) or chemical compounds as separate entities (e.g., the CAS REGISTRY file) is rather straightforward,[3] information retrieval systems containing *chemical reactions* in machine-readable form (e.g., ORAC, REACCS, and SYNLIB[1,2]) pose a much more sophisticated problem.

Accordingly, searching in files for textual information can easily be done by locating specific character strings or substrings. In order to find chemical compounds, searching for (chemical) substructures is an appropriate method.[3]

Searching in databases containing chemical reactions is a much more complex endeavor. The challenge of identifying "similar reactions" is hard to satisfy by carrying out substructure searches among the educts and products, because reactions which a chemist would easily recognize as "similar" can occur with a variety of different substrates, which the computer would not be able to identify as such.

A more general formulation of the task, on the other hand, runs the risk of yielding too much information that would have to be further evaluated intellectually thereafter. Instead a computer-assisted search in such an information retrieval system is expected to reduce the amount of useless material and thereby increase the density of the meaningful information. The ultimate goal would be the creation of a meaningful classification scheme for chemical reactions that is suitable for searching and comparing "similar" reactions. A significant step in this direction would be the development of parameters describing the reactivity of the reactive centers, the bonds, and the groups. Therefore, we have come up with a procedure which allows an exact identification of the reactive atoms and bonds in a reactive site via a generic differentiation. It is important for the application to reactions contained in commercial databases that the program must be capable of functioning properly even with incomplete, i.e., nonstoichiometric representations of machine-readable reaction schemes. This approach, namely, neglecting all reactants other than the interesting substrate, is the very method practiced by chemists, and it is also used in common storage concepts of chemical reaction schemes. An important goal of this work is to come up with a format that can handle reaction files in such a form as is common practice in the laboratory.

This is the very reason for the application of a topological approach, which has basically been well-known for a long time. We have not yet been able, however, to build a uniform method for all chemical reactions with this method, but it is possible to create a system that can handle nonstoichiometric reactions.

A procedure described recently by Funatsu et al.[4] for the determination of reactive centers in chemical reactions starts with the same concept, namely, a topological approach, but continues the analysis in an intrinsically different way.

In this study, we use the same basic algorithm for all tasks of the program, even for our essentially new approach to an automatic completion of nonstoichiometric reactions.

The generation of complete reactions is the common prerequisite for the application of many well-known, powerful programs, whether for synthesis planning or for the examination of reaction mechanisms. As a typical example in this context may serve the well-established approach of Ugi and Dugundji,[5] which is based on the mathematical model of constitutional chemistry, describing chemical reactions via a matrix transformation.[6] This concept requires a FIEM (Family of Isomeric Ensembles of Molecules), namely, a stoichiometric reaction. This is also the reason for our application of the method of the maximal common substructure, which is the best method for the special need in handling nonstoichiometric, sometimes dramatically truncated reaction equations. As already mentioned, this is a main goal of this work. In terms of graph theory, the PMCD (Principle of Minimum Chemical Distance)[6b,c,f] instead is a principle of the maximum set of maximum common subgraphs in educt and product. But again, some expedient features of the PMCD, e.g., the atom-by-atom correlation of the educts and products of a chemical reaction, require a FIEM. The variety of possible applications of such a method and the additional potential for other concepts emphasize the significance of and the need for a reaction completion scheme, which to our knowledge has not yet been published.

GENERAL ALGORITHM FOR THE ANALYSIS AND COMPLETION OF CHEMICAL REACTIONS

In principle, our program is based on the application of a highly modified Morgan algorithm,[7] which is applied here simultaneously to both sides of a reaction. This method of recognizing *inter-* rather than *intra*molecular equivalences was used for the first time in 1978 by Lynch and Willett[8] for *The*

0095-2338/90/1630-0228$02.50/0

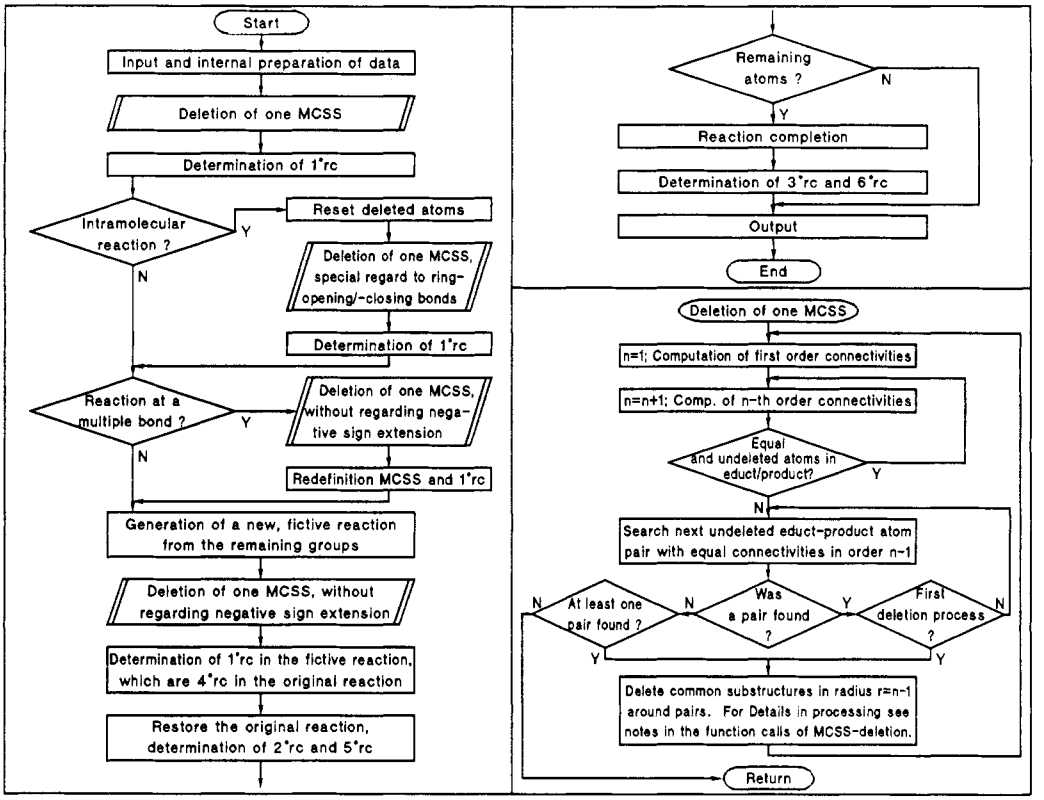

Figure 1. Principle program structure.

Automatic Detection of Chemical Reaction Sites. There, these authors tried to find out the *reaction site*, i.e., the very substructure containing the reactive centers and bonds, but not the specific atoms and bonds explicitly. Funatsu et al.[4] start out by identifying the reaction site (which was carried out similar to the procedure of Lynch and Willett), and then they apply a back-tracking algorithm for the more exact determination of the reactive centers. Here we use uniformly the concept of connectivities and common substructures for the exact determination of up to six classes of reactive centers, and we also apply exactly the same routines to achieve the reaction completion. The principle structure of the program, demonstrated in Figure 1, will be discussed using the example given in Figure 2.[9] Further details will be illustrated in the subsequent sections.

Depending on the format of the input file, the internal representation of the reaction is being built up starting from the input data.

As already mentioned, we use the concept of connectivities in order to find intermolecular equivalences. The *first order connectivity* represents the type of an atom and its hybridization pattern (for details and enhancements with respect to older methods dealing with connectivities, see Discussion of Components). Thus, in Figure 2 the atoms 1–12 in the educt (denoted E1–E12 in the following) are all identical, similarly the atoms 1–12 in the product (P1–P12) are identical among themselves and they are also equal to E1–E12. Furthermore, E1 and P1 have the same connectivity as E2–12 and P2–12, since hydrogen atoms are included implicitly in the first-order connectivity set. All atoms mentioned above are sp^3-hybridized carbon atoms with four single bonds.

It should be emphasized that we also tried to regard hydrogen atoms, which were explicitly given in the reaction equation because they play an important role, e.g., acid hydrogen atoms which undergo a substitution reaction. But it is obvious that any differentiation between such explicit hydrogen atoms and the others leads to different connectivities for topological equivalent atoms in a molecule, as will be outlined again in the following.

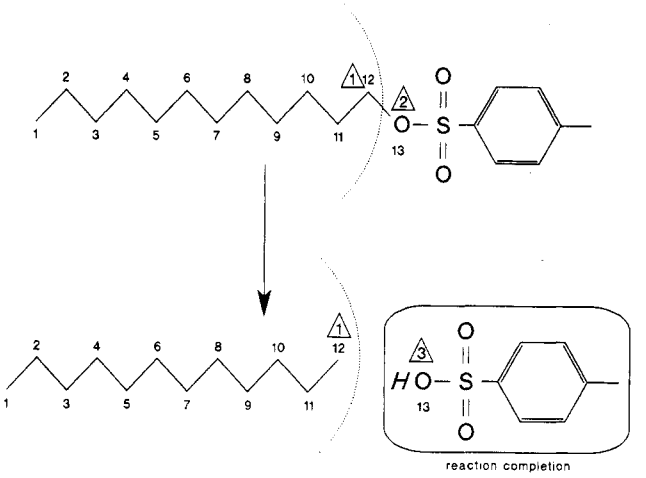

Figure 2. Example for the analysis and completion of a substitution reaction involving reacting implicit hydrogen atoms.

A *connectivity of a higher order* is subsequently computed from the connectivity of the specific atom itself adding the sum of the connectivities of the directly connected atoms. Looking at the atoms E1–E12 in Figure 2 as an example, the connectivity of E1 (neighbor atom: one carbon) differs from that of E2–E11 (each has two adjacent carbons). Furthermore, E12 has another connectivity value now (one carbon and one oxygen are bound). Therefore, figuratively speaking: In the first order, we only "see" the considered atom itself and its hybridization pattern, whereas in the *n*th order the radius of discrimination includes the atoms which are up to $n-1$ bonds removed from the reference atom, i.e., we see a substructure within the radius of $n-1$ bonds around the atom considered. Thereafter we compute connectivities of higher order for all atoms both in the educt and the product until no equal connectivities can be detected anymore on either side of the reaction, i.e., now the radius of discrimination reaches all the

way to the reactive bond starting from the most distant atom. Because the structures are all different (beginning at the reactive bond), no equal connectivities can occur in higher orders. In our example, in the 12th order the radius of discrimination extends all the way from E1/P1 to E12/P12, respectively. The substructures within the radius of 12−1 = 11 bonds around E1/P1 are identical; therefore, the 12th order connectivities of E1/P1 are also identical. We may also look at this example from the other side of the carbon chain, where the connectivities of E12/P12 are different in the second order and up. Accordingly, the difference in the structures causes a difference in the connectivity of E2/P2 in the 12th order. Now, E1/P1 is the last pair of atoms with the same connectivity value in educt and product. In the 13th order there are now no more equal connectivities on the two sides of the reaction. Therefore, we now go back to the last order, search pairs of equal connectivities on the opposite sides, and find the pair E1−P1 in the n = 12th order. This means that E1/P1 are the centers of a common substructure within the radius of n−1 = 11 bonds, which is symbolized in Figure 2 by the dotted lines.

At this point, the necessity of equality of all the hydrogen atoms becomes obvious since different notations of a reaction equation in respect to the H-atoms should not affect the result. One possibility would be recognizing only explicit hydrogens. Assuming P12 in our example fully written out as a CH_3 group and the educt as shown, E12 would have two bonds in contrast to P12 with its four bonds, resulting in the recognition of a too little common substructure within a radius of only 10 bonds around E1/P1. We would also find a wrong definition of the carbons E12/P12 as or as parts of the reactive groups. A similar situation occurs with the reaction in Figure 2, written as shown with no explicit hydrogen atoms: E12 with two bonds would not correspond to P12 with one bond, the same problem we have with any other reaction concerning implied hydrogens. In the case of recognizing explicit and implied hydrogens in a different way, the program would fail in every situation where the representation is not the same on the educt and the product side. Typical examples are H-atoms indicating stereochemistry, written only on one side, or the two hydrogens of a reacting methylene group, shown both on one side of the reaction and only the new substituent on the other side. The most effective way to overcome the problem of different possible notations of a reaction equation, using explicit and/or implicit hydrogens, is to consider them all equal and implicit.

Coming back to our example, deleting all atoms accessible within a radius of 11 bonds, starting at E1/P1, means to delete the maximum common substructure (MCSS) of educt and product since there are no more common atoms.

Now, the reactive bond in the educt is E12−E13, exactly the bond which connects the MCSS and the undeleted, reactive group, i.e., the tosyl group. In the product, the corresponding bond is the one between P12 and an implicit hydrogen.

As a reactive center of first order (in the following abbreviated as 1°rc), we define the atom at the reactive bond which is part of the MCSS. In our example these 1°rc's are E12 and P12, respectively. The reactive centers of second order (2°rc's) are the atoms at the other side of the reactive bond which belong to the reactive group: here this is E13 (since hydrogens are implicit, no atom is marked as 2°rc on the product side). In Figure 2 and in the following figures we will mark a reactive center of nth order with a triangle containing the number n. Simple numbers refer to atom numbers, which will be used in the following with the prefix E for educt atoms or P for product atoms, respectively.

In the case of an undeleted reactive group connected by the reactive bond to the deleted MCSS, the 1°rc and 2°rc can easily be found at the end of the reactive bond. This method would fail, however, on the product side of the reaction depicted in Figure 2 as an example. Because of some other special cases which are unsuited for this method of identifying reactive centers, we use other algorithms to determine the 1°rc. Especially helpful for this purpose is the fact that 1°rc's can be found among those atoms which are canceled not until the last rounds of deletion processes. In order to delete a common substructure of radius n, also n rounds of canceling processes are necessary: the starting atom itself, then its neighbors, and so on. Since we start at the most distant atom from the reactive center, this one must be deleted in the last round.

One new approach, introduced in this study, is the stoichiometric completion of the reaction. In the case of Figure 2, the tosyl group of the educt should appear as "waste" on the product side. The philosophy here is to apply again the same algorithms which we already used before to determine the MCSS, but this time in order to achieve the completion of the reaction. This is done in three steps: At first a new, fictive reaction is being built up out of the undeleted reactive groups. In our example, this means that a tosyl group reacts to nothing. More exactly, the educt side of the fictive reaction is toluenesulfonic acid, since we replace the bond to the deleted MCSS by an implicit hydrogen. Secondly, the same procedure of computing connectivities until no more correspondencies can be found (on the two sides of the reaction) is applied to the fictive reaction, including the deletion of common substructures. Since there are no structures which can be matched, it is obvious that the tosyl group of the educt must be copied to the product side to yield a stoichiometric reaction. The resulting oxygen P13, which is the 2°rc on the educt side, is defined as a 3°rc in the tosyl group, and this is added by the program to the product side. Since the reactive bond E12−E13 is broken in the reaction, the oxygen atom would still have a free valence. It is possible to compensate for this either via a negative charge or with an implicit hydrogen added to P13. We decided to saturate the free valence by an implicit hydrogen, which was already introduced in generating the fictive reaction. Exactly these components of the fictive reaction are used now for the reaction completion. The implicit hydrogen atom introduced by the program is highlighted in italic in the reaction completion in Figure 2. Here and in the following, only these hydrogen atoms added to non-hydrogen reaction completions are shown in the figures (in italics), whereas normally all other hydrogen atoms already given in the input file as well as those of the reaction completion, which correspond to given hydrogens are not shown. This simplified representation is preferred and used in the following since it provides for a shorthand that communicates the essential features of this procedure in a suitable form.

So far, the analysis of the reaction and the reaction completion concerned only non-hydrogen atoms with implied hydrogens. But even a demand of specifying reactive hydrogen atoms and of balancing the stoichiometry up to an equal number of hydrogens on educt and product side can be satisfied. Each 1°rc in the substrate without an attached 2°rc indicates the reaction of an implicit hydrogen; 3°rc's everytime correspond to 2°rc's on the other side of the reaction, the possibilities to saturate the free valence of a 3°rc in the reaction completion were just discussed.

If the reaction in Figure 2 would be a normal hydrolysis, imaging the product to be dodecyl alcohol with an oxygen at the carbon atom P12, this oxygen copied by the reaction completion as H_2O to the educt side would even balance the stoichiometry in respect to the number of hydrogens. By comparing the number of implicit hydrogens at all the reactive centers, we find in the reaction equation as shown in Figure 2 a deficit of two hydrogen atoms on the educt side. This result is indicated in the output file in a separated information block (see Data Input under Discussion of Components) because of

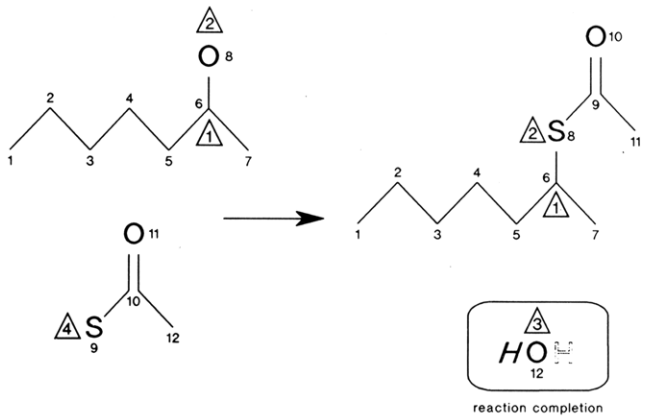

Figure 3. Example for a reaction equation with two given components on one side. The implicit H-atom added by the program is shown in italic, the other (outline letter) was already given on the educt side and is shown only for optical reasons, demonstrating the reaction completion is water.

two reasons. First, we want to continue with the principle to consider all hydrogens implicit. A reaction completion of two explicit hydrogens on the educt side would break this convention. Second, the form in which these hydrogen atoms had to be added cannot be determined in a definite way. In this reaction, it is not a catalytic hydrogenation (where hydrogen really as H_2 would be added) but a reaction with sodium borohydride. But this is only indicated as textual comment in the reaction file. Also other circumstances, like the choice of solvents participating in some way in the reaction, prevent the possibility of a chemically meaningful reaction completion of explicit hydrogen atoms. Another example will be given later.

DETAILS AND DISCUSSION OF EXAMPLES

The approach used here is essentially similar to the method pioneered by Lynch and Willett[8] to detect reaction *sites*. These reaction *sites* mean greater substructures containing more than the reactive centers and bonds, which are not exactly determined. Our procedure differs from their scheme by adding the following:

- An exact marking of up to six classes of reactive centers and an exact determination of the reactive bonds
- A new reaction completion scheme
- Additional enhancements to achieve a meaningful analysis of more complex reactions, allowing the exact determination of reactive centers

In order to deal with the most important fact, namely, that in almost all cases the MCSS of educt and product cannot be deleted in one step, we start by deleting *one* common substructure only, leaving all other corresponding groups untouched. Thereafter, the whole procedure of computing connectivities beginning with the first order is repeated until no further correspondencies can be found anymore. This step yields the starting points to delete the next common substructure in the radius which is given by the number of the last order. It has to be emphasized, however, that this simple method, as described by Lynch and Willett, works only in special cases, namely, where no *inter*molecular or even *intra*molecular ambiguities exist.

The reaction given in Figure 3[10] may serve as an example for the most common case, namely, a two-component reaction. Initially we compute the connectivities up to the seventh order: By now the influence of the difference between E8 (oxygen) and P8 (sulfur) has reached E1/P1 and causes different connectivities there. Therefore, in the first step we delete a

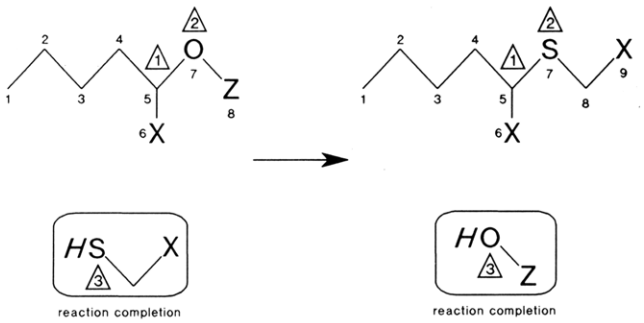

Figure 4. Example for a reaction where the method of the spreading negative sign overcomes intramolecular ambiguities.

common substructure of $r = 6$ starting at E1/P1, containing E1...E6 and P1...P6. The methyl groups E7 and P7, parts of the MCSS, are still undeleted. But now, the highest order in which E7/P7 would be equal is the second one, since in the third order the difference of E8/P8 will cause different connectivities for the methyl groups. Looking at the carbons E12/P11, we find that these atoms have equal connectivities up to the third order. The fourth order connectivities are different, however, because E9/P8 have a different environment. Following the principal algorithm, we would have to delete the thioacetic acid on the educt side (E9...E12) and the corresponding ester group on the product side (P8...P11). However, this very result is not desired here, because this deletion would cause a loss of the reactive groups which have to be retained since they are essential for building up the new, fictive reaction that we want to examine in order to achieve a stoichiometric reaction completion. Problems of this type could still be resolved by demanding that, in subsequent deletion processes after the first one, we may only use starting points within the same molecule which we modified in the first deletion process.

Let us now examine the much more difficult, even if rare, case of a fictive reaction as shown in Figure 4: An alkoxy group is replaced here by a better nucleophile, namely, a thioalcoholate—and this, the reactive group, may contain a substructure X which also occurs in the MCSS of the molecules. After the first deletion process (with $r = 5$, starting at E1/P1, deleting E1...E5 and P1...P5), the second order of new computed connectivities is the highest order in which equal values on the two sides of the reaction can be found. It should be mentioned, that, of course, after the first deletion process we only look for equal connectivities among the still remaining atoms. In case of Figure 4, E6 is equal to P6 or P9 in the second order (remember, E5 and P5/P8 are equal in first order because of implicit hydrogens). Depending on the internal numbering of the atoms or sorting methods used in the program, not only a deletion with $r = 2$ starting at E6/P6 would be possible, but also the wrong pair E6/P9 with $r = 2$ could be chosen, leading to a false result: the reactive group would be deleted versus a part of the MCSS. Note that we compute the connectivities always regarding the *whole* structure, but only the still undeleted atoms are tested relative to equal connectivities. Using only the remaining atoms after each deletion process for the analysis of the reaction would mean an unnecessary loss of information. Actually, another case equivalent to that shown in Figure 4 would be if the reactive groups contained equal substructures.

The uniform solution for problems of this kind is simple but very effective, as becomes evident after the following consideration: Which properties have the reagent thioacetic acid and the reactive ester group in Figure 3 as well as the reactive groups in Figure 4 in common? They are either completely separated from the MCSS (first reactant on the educt side in Figure 3) or they are connected to the MCSS by "the

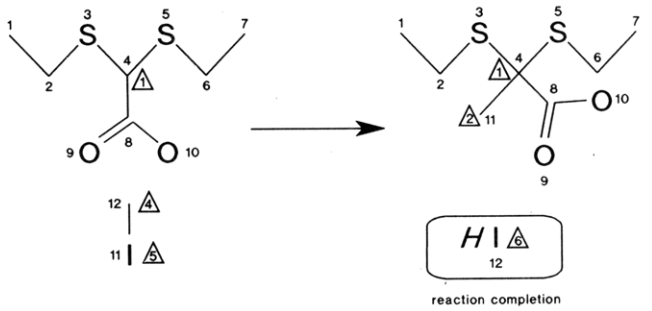

Figure 5. Example for a reaction where two common substructures could be deleted at once at the beginning, without disturbing the further analysis.

difference" of the structures. Here "the difference" means oxygen or sulfur in comparable positions in the molecules. Accordingly, atoms in the reactive group are at least one bond further away from the common substructure, which has been deleted in the first step, than atoms of the MCSS, which are directly connected to this common substructure.

In the deletion process we perform the deletion by assigning a negative sign to the ident of the canceled atoms. (Here the "ident" means a number representing an atom; for details see Discussion of Components.) In the next deletion process, the negative sign of the ident causes a negative sign of the first-order connectivity. Since the higher order connectivities consist of the connectivity of the atom under consideration (and thereby its ident) and the sum of connectivities of the neighbor atoms, we assign a negative sign to the higher order connectivity of an atom if a neighbor atom has a negative connectivity value. The consequence of this procedure is that the process of computing connectivities after the first deletion process causes negative signs, which spread from the first common substructure by one consecutive radius for each order. In the second and all higher deletion processes, we now accept the condition "connectivities are equal" only if the absolute values are the same *and* the two connectivities are both negative! In our examples, atoms of the thioacetic acid as the reagent in Figure 3 cannot have negative connectivities in the determination of the MCSS, since the first deletion starts within the heptanol-2, and no connection is given to the acid—i.e., no "bridge" by which the minus sign could reach it in the way described. In Figure 4, E6 and P9 have the same absolute value of their connectivities up to the second order in the second deletion process, but only E6 (and P6!) has a negative value, defining beyond doubt E6/P6 as the starting points for the second deletion with the radius $r = 2$. The same situation would occur with common substructures in the reactive groups—i.e., the spreading of the negative sign is always one atom behind relative to the possibility of finding equal absolute values of connectivities.

In other cases with symmetric molecules as substrates it is possible that multiple pairs of atoms can be found which are suitable as starting points for a deletion process of the same radius. Let us consider for example the reaction scheme in Figure 5:[11] It would be possible to take E1/P1 *and* E7/P7 as starting points for two deletions with $r = 4$ in the first step, removing the whole sulfur-containing chain at once. But this method does not work correctly with other reactions. In Figure 6,[12] the radius of the substructure to the left and the right from the reactive bond in the product (P8–P9) is the same ($r = 6$ atoms); therefore, we would find two pairs (E23/P14 and E1/P1) to start the deletion with $r = 6$, which causes substructures in both educts to be deleted. No differentiation between substrate and reagent will occur and no reaction completion can be done. Therefore, in the first deletion process, we use only *one* pair of corresponding atoms as starting points, regardless of the number of possible pairs. Then, in

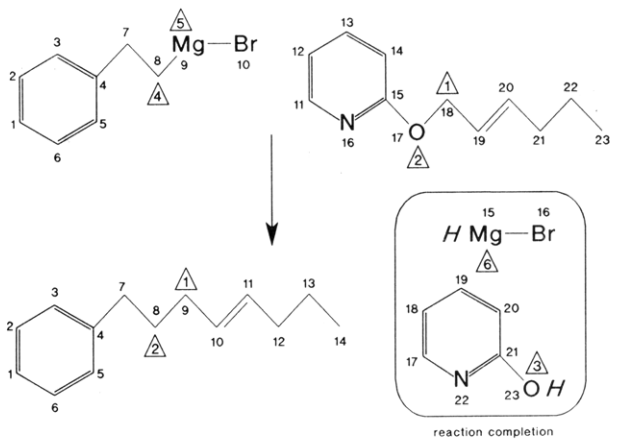

Figure 6. In contrast to Figure 5, deleting two possible common substructures at once would prohibit the correct differentiation between substrate and reagent, leading to a wrong analysis and reaction completion.

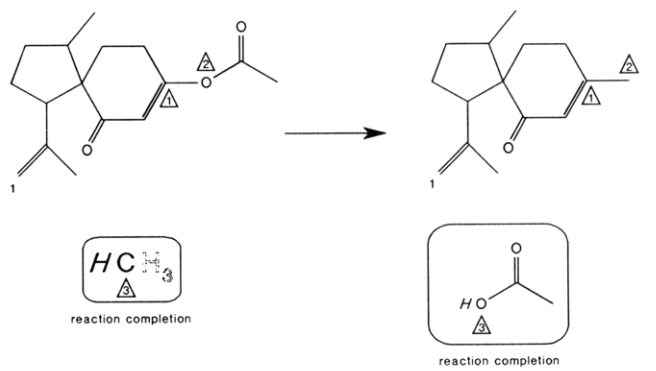

Figure 7. Example for a "full reaction completion". The three H-atoms already given (in outline letters) are only for optical purpose, showing methane as reaction completion.

subsequent deletions, the method of the spreading negative sign suppresses the deletion of common substructures in more than one molecule on each side or in one molecule on two sides of the reactive bond.

The reaction in Figure 6 will also serve as another example for the impossibility of completing a reaction with explicit hydrogen atoms. As in Figure 2, by counting all implicit hydrogens at the reactive centers we find a deficit of two hydrogen atoms on the educt side. Also this Grignard reaction is not a hydrogenation, the two missing hydrogens could be added in the best way as one molecule of water (aqueous workup giving MgBrOH), but this is not indicated in the reaction file. The reaction completion is done mainly in one of two ways, which we call either *the reaction completion in part* or *the full reaction completion.* In this context, we will first explain the definition of the already used (Figure 5) reactive centers of fourth, fifth, and sixth order (4°rc, 5°rc, 6°rc). Whereas the 1°rc (in the MCSS) and the 2°rc (in the reactive group) are at the ends of the reactive bond of the substrate, and the 3°rc corresponds to the 2°rc in a reaction completion, the 4°rc to 6°rc have the same meaning for a separated reagent molecule since the program can handle multicomponent reactions. As already mentioned in the previous section for the 1°rc in the substrate, a 4°rc without an attached 5°rc indicates a reacting implicit hydrogen atom in the reagent molecule. The 4°rc and 5°rc mark the reactive bond in a separated molecule which does not contain the MCSS, like E11 and E12 in Figure 5. The 4°rc becomes the 2°rc in the product; the 5°rc will be completed, if necessary, as a 6°rc on the product side (P12 in Figure 5).

However, this example is already a "special case" of a reaction completion in part. In more general terms, the process

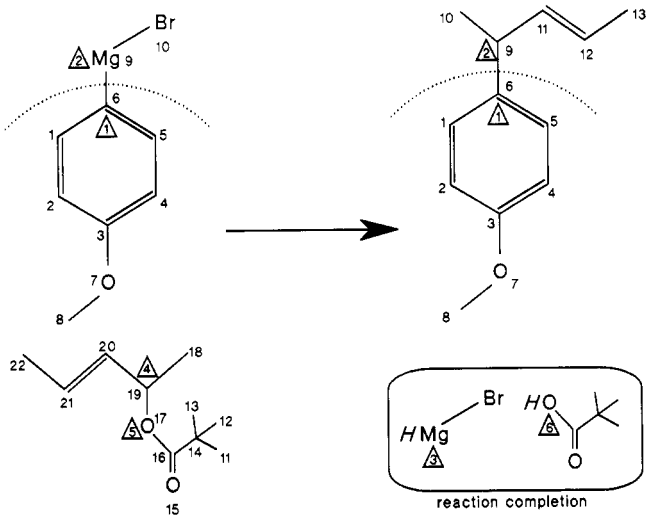

Figure 8. Example for a "reaction completion in part".

of the reaction completion should be explained using the reaction given in Figure 7:[13] After deletion (starting at E1/P1 with $r = 7$), the 1°rc and 2°rc will be determined as shown. For this reaction completion, the new, fictive reaction generated by the program would be "acetic acid to methane". This result occurs because the free valencies generated via breaking the reactive bonds become saturated by implicit hydrogen atoms. As mentioned, we could also use negative charges resulting in "acetate to methyl", which is only a matter of definition.

Now, however, in this "new reaction", the methane on the product side could be deleted versus the methyl group of the acetic acid—and using the remaining fragment for the reaction completion, this would give a wrong result. The same problem occurs if only parts of the reactive groups (which build the new, fictive reaction) can be deleted versus reactive groups or separate reagents. Therefore, even if parts are deleted, we use the reactive groups as they were before the deletion process in the new reaction in order to perform the reaction completion. In Figure 7, the reaction completion is done as shown with all reactive groups.

A Grignard reaction, shown in Figure 8,[14] represents a typical example of one key exception: the benzene ring with the methoxy group (below the dotted lines) is found as the MCSS, i.e., the 1°rc and 2°rc are determined as shown. The new, fictive reaction is "Mg – Br plus the pentyl ester to 2-methylbutene-2". Now, the whole structure on the product side can be found in the reactants—but unlike in Figure 7, here the correspondent group on the left side of the reaction is a separate reagent, not a reactive group bond to the MCSS. In this case here, it is meaningful to complete the reaction only with the undeleted parts of the fictive reaction. The result is shown in Figure 8. The situation in respect to the number of implicit hydrogens is the same as in the reaction of Figure 6.

SPECIAL REACTION CLASSES

So far, all examples given were concerned with reactions, which, in the most general sense, could be described as *inter*molecular substitutions. In these cases a clear definition of the MCSS, reactive group, and bond exists. Differences like substitution of implicit hydrogens or representations of reactions with or without explicitly given reagents do not change the principle of this approach.

***Intra* molecular reactions** behave exceptionally as a group. Here, a substitution can occur via a part of the same molecule. In order to deal with such cases, we decided in the course of this study to take the following approach: At first, a normal MCSS search is done, which usually deletes too many atoms. Then, with the MCSS, we touch at least one reactive center.

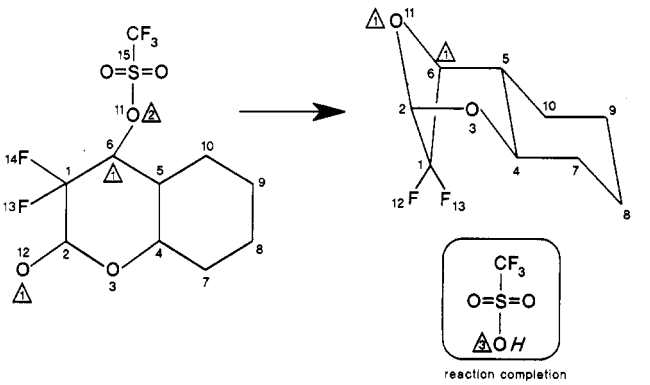

Figure 9. Example for an intramolecular reaction.

Therefore, going one atom further away from the MCSS toward the reactive group, we can determine if there is a change in the number of rings between educt and product. Of course, we only look at this "MCSS plus one radius of additional atoms" because other molecules or the reactive groups can also contain rings. Once an *intra*molecular reaction has been detected, we start all over with the deletion process again. This time, in the last deletion radius a comparison is performed on the educt and the product side with the goal, to detect a difference of the number of bonds between the atoms to be deleted in the last deletion radius. If a different number of bonds are being found, then no deletion is initiated because the 1°rc's are among these atoms. Other atoms which are not reactive centers will be deleted in following deletion processes using other starting points. Most importantly with this method, we can detect the reactive bond in *intra*molecular reactions. The definition of a reactive bond, i.e., the meaning of 1°rc and 2°rc, changes somewhat in these reactions, but the difference to *inter*molecular reactions is not significant since the reaction type is different, and by then we have obtained the information that we are dealing with this special kind of reaction in this specific case.

In *intra*molecular reactions, the bond which opens or closes a ring has *two* 1°rc's at its ends. But also 2°rc's are possible here: one 1°rc and a 2°rc define a second type of reactive bond which connects a leaving (e.g., during a ring-closure reaction) or an incoming (ring-opening) reactive group, for example, in an *inter*molecular reaction, as discussed before.

In some cases of intramolecular reactions, which will not be discussed here in detail, the program detects on each side of the reaction two 1°rc's, and consequently the ring-opening or ring-closing bond, but in addition a leaving group with 2°rc, for example. These reactions can be regarded as being analyzed correctly at once.

Frequently, a type of intramolecular reactions occurs where the first analysis (which is already customized for these reactions, however) yields a result with two 1°rc's on one side and only one 1°rc on the other side of the reaction. Using a more elaborate procedure, even such examples can be handled as outlined for the reaction shown in Figure 9.[15]

After the deletion processes (E8/P8 $r = 6$, E13/P12 $r = 4$, E14/P13 $r = 4$), the two oxygen atoms in the educt, E11 and E12, are found as 1°rc's; on the product side, however, it is the oxygen P11. This apparent discrepancy is caused by the symmetric deletion process with the starting points E8/P8 and $r = 6$, since the atom which closes the ring (E12/P11) is an element of the same type as the reactive atom of the leaving group (E11). It should be emphasized that P11 can be found as 1°rc even though no reactive group is connected to it, and yet we find it as the only atom that is deleted only in the last radii of several deletion processes. P1, which is as far away as P11 from the starting point P8, is deleted a second time when processing the fluorine atoms (deletion starts at

E13,14/P12,13 with $r = 4$), and there the carbon is reached in $r = 2$ of a total of four rounds. Whereas it is obvious that P11 is one of the two 1°rc's in the product, we have to choose either E11 or E12 as the equivalent atom in the educt. This information provided, we can then identify the other 1°rc (namely E2 or E6 and P2 or P6, respectively). The following section provides further details.

Within this algorithm, we perform a temporary truncation of the reaction, omitting both the reactive centers and groups on both sides of the reaction. This is done by setting the idents of these atoms to zero, saving the previous values to restore the original reaction later. The same is done with the entries in the bond matrix, which contain the bond orders of bonds between reactive centers/groups and the MCSS. Subsequently, the "rc-candidates" of the remaining MCSS's can be correlated by computing connectivities as usual. In this context, we will refer to all atoms of the MCSS as "rc-candidates", which are bound to the 1°rc's already determined. Among these is the other 1°rc for the very side of the reaction where we first found *one* 1°rc. On the other side, one 1°rc of the two has to be canceled since otherwise the result would be wrong. The right one, therefore, is a rc-candidate. The computation of connectivities is done until one of two possible terminating conditions occurs. In the first one, all rc-candidates on either side have different connectivities on this side of the reaction. In this case a correlation of the rc-candidates is possible. Otherwise, the computation of connectivities is terminated when the order reaches a number greater than the maximum number of atoms on any side of the reaction. Atoms which have equal connectivities until now can be considered as equal in the sense of being candidates for the 1°rc's. In the case of our example in Figure 9, E2,6 and P2,6 are our rc-candidates. In the second order we obtain the correlated pairs E2/P2 and E6/P6. Now we have to determine which of the two 1°rc's on the educt side is the right one. If these two would be different elements, the answer could be given immediately: We could cancel the wrong 1°rc and declare the rc-candidate bound to it as the new, second 1°rc. But here, both 1°rc's found initially are oxygen. Now we turn the situation around and perform a truncation of the previously restored original reaction, now omitting the MCSS, including all bonds between the MCSS and the remaining atoms, and correlate the reactive centers by computing connectivities as described above. This results in E12/P11 being the only possible pair, and it is shown that E12 is the right 1°rc on the educt side among the two atoms found originally. Therefore, E11 is canceled as rc, and the bound rc-candidate E6 is the second 1°rc. By having the information about the correlation E6/P6, P6 can be easily defined as the second 1°rc for the other side. The last step is to find the bond E6–E11 as a reactive bond from the MCSS to the reactive group, identifying E11 as a 2°rc. The following reaction completion takes place as shown in Figure 9.

It is again possible that no correlation of the rc-candidates is practical. An intramolecular formation of an ether by elimination of water from two equivalent hydroxyl groups could serve as such an example. In this case we arbitrary select an atom.

The second main group of reactions that have to be handled in a special way is all **reactions which change multiple bonds**. At first, we will restrict ourselves to additions and eliminations in the most general sense. Because of the procedure used to compute the connectivities of first order (which utilizes the hybridization pattern, see next section), atoms which have a different number or kind of multiple bonds are not recognized as equal and, therefore, not as a part of the MCSS.

In the example show in Figure 10,[16] the first deletion process starts at E1/P1 with a radius $r = 4$, because E7/P7 have

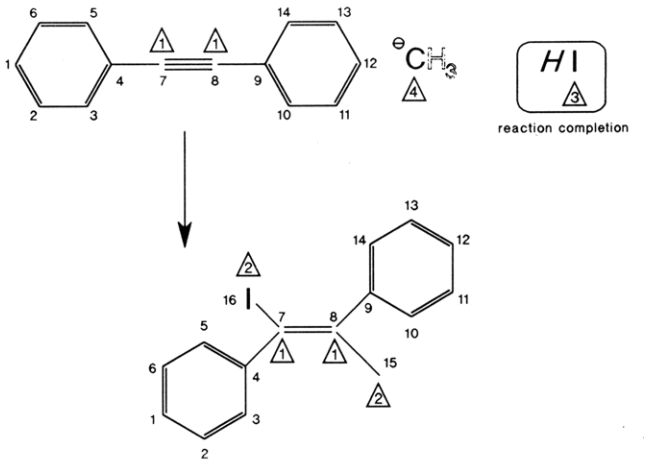

Figure 10. Example for a reaction at a multiple bond. H-atoms in outline letters are again only for optical purpose, showing a given methyl anion as reagent.

different connectivities in the first order. The goal is to delete the common substructure (the other phenyl group) on the other side of the multiple bond first, then to recognize E7/E8 and P7/P8 as the 1°rc's which mark the changed multiple bond, and finally to find the other reactive centers. This is achieved in the following way: Since we know that this is a reaction at a multiple bond (from the hybridization pattern at the reactive centers found initially), we now neglect the method of the spreading of the negative sign of connectivities. Thus, if the example is a reaction converting a multiple bond into a single bond, normal deletion processes are performed until the MCSS is found, excluding the multiple bond and the corresponding atoms on the other side. Since these atoms are not deleted, reactive groups bound to these atoms also remain unchanged. In our example, there is only one additional deletion process, namely E12/P12 with $r = 4$, which deletes the second phenyl group. Now a transformation of rc's takes place: The "old 1°rc's" go back in the MCSS (E4,9 and P4,9), and the "old 2°rc's" become the "new 1°rc's" (E7,8 and P7,8). The rest of the analysis (reactive groups, 2°rc's and the reaction completion with 3°rc and 4°rc) is carried out in the same way as in a "normal" reaction.

As already mentioned in the introduction, the method is not sensitive to nonstoichiometric reactions, but it does not represent a universal algorithm to deal with chemistry as a whole. This is evident from the mandatory special treatment of cases like the intramolecular and the multiple-bond reactions whereby the last one represents a rather common group of reactions.

Excluding a few exotic cases which cannot be grouped into a category of "special cases", as defined above, there exist two types of reaction equations which cannot be handled universally with this approach, even though some of these reactions are analyzed correctly.

Since the basic principle is the detection of a MCSS, using connectivities for the representation of topological equivalences, all **rearrangements** with considerable changes of the molecular skeleton give unsatisfactory results. But this could be the subject of further enhancements.

Another problem is **multiple reactions**, i.e., more than one chemical reaction written in one reaction equation. Unfortunately for the application of this algorithm here, it is a widespread practice to write down whole reaction sequences as a single reaction, denoting many consecutive transformations as textual comments above the arrow symbolizing the reaction. Especially if the individual reactions comprise reaction types which have to be handled differently by the program, there is little chance to achieve a correct analysis overall. Effectively

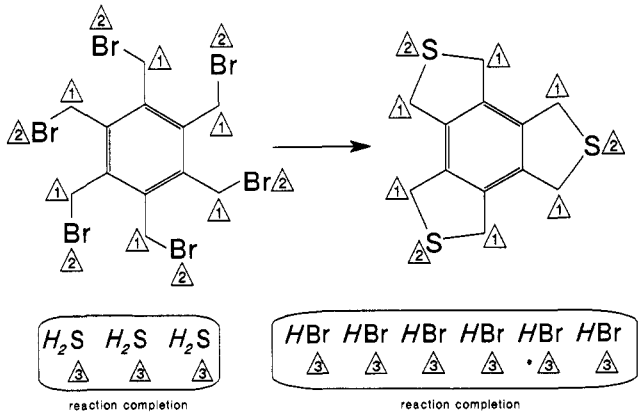

Figure 11. Example for a reaction equation which comprises several reactions of the same type.

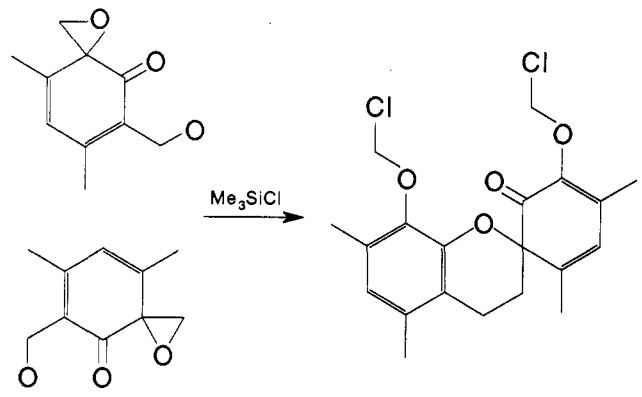

Figure 12. Example for a reaction equation comprising many different steps, including rearrangements. No chemically meaningful result can be obtained.

such reactions are essentially equally difficult to deal with as complex rearrangements.

However, examples of multiple reaction steps in one equation, which can be handled well in general, are reactions of the same type carried out more than once in a molecule. As an extreme example, Figure 11[17] illustrates a case where six substitutions occur, and accordingly, almost the whole educt and product consist of reactive centers. Since all transformations are of the same type, the determination of the reactive centers and also the reaction completion is done well.

If, however, in the example depicted in Figure 12[18] almost all atoms react in a certain way or change their bond pattern, no chemically meaningful result is obtained.

One possible concept to split up a complex reaction into single steps is the approach to generate complex reaction networks,[6g] which, in this context, could be very useful, but has not been utilized here thus far.

DISCUSSION OF COMPONENTS

In this section, some general technical aspects and some concerning the subroutines shown in Figure 1 will be outlined.

The program is written in FORTRAN-77 and has about 3000 lines without comments. With the necessary hardware-dependent modifications, it has been run under DOS on a PC-AT, under VM/CMS on IBM 3081/4381, and under VMS on DEC-VAX systems. The execution times, depending on the size of the molecules and the number of deletion processes (see intramolecular and reactions at multiple bonds), are in the range of seconds on a PC-AT. Typically, it takes 10 s or less for an average reaction; however, some examples with up to 100 atoms and multiple bonds can require a little more than 20 s computing time on a 12-MHz PC-AT. An important fact is that no algorithms are involved which use computing time exponentially depending on the number of atoms.

Data Input. The input routines are intended to be capable of accepting two kinds of data formats. These are either RXN files, used by MDL[19] programs like ChemBase, MACCS, and REACCS,[1,2,20] or SMD files,[21] as developed by a consortium of seven German and Swiss chemical companies to ensure a convenient transfer of chemical data between different hardware and software systems. In both cases, the same information concerning number and kind of atoms and bonds (the connection tables and atom vectors) is extracted from the input files and converted to a unique internal representation of the reaction, independent upon the input format, whether it was a RXN or a SMD file, respectively. Other input data are stored to be printed out unchanged at the end of the program. Especially valuable is the flexibility of the SMD format, because it allows a convenient insertion of new information blocks, e.g., for the reaction centers, for the kind of reaction, the reaction completion, and other data. Otherwise, new information gained during the execution of the program must be stored in unused or unimportant data positions, e.g., isotope information. However, such provisions may change the data format and can cause problems with other routines.

Connectivities of First Order. It is intended to achieve a high differentiation already in low orders. Since it is advisable to use moderately low numbers (see below under Connectivities of Higher Orders), we do not simply use the atomic number of an element as its first-order connectivity. Instead, the connectivity information is being built up of the hybridization pattern as a relevant chemical property and of an ident I, which depends on the probability of occurrence of the element in organic chemical reactions and on the possible hybridization patterns. For example, carbon as the most frequent element has the ident $I = 1$. The hybridization pattern is compiled as the sum of the squares of the bond orders B of all the b bonds at the atom and is added to the ident. The number of implicit hydrogens h count as single bonds in the computation of the first order connectivity of atom number x, ${}^1C_x$:

$$ {}^1C_x = I_x + \sum_{i=1}^{b} B_i^2 + h \qquad (1) $$

Thus, an acetylenic carbon can reach $1 + (1^2 + 3^2) = 11$ as its maximal connectivity; therefore, the ident of the next element will be $I = 12$. Only important elements for organic chemical reactions are handled this way; i.e., we use C, O, N, S, Cl, Br, F, I, Mg, Li, Si, P, Na, K, Cu, Zn and B explicitly and in this order, all other elements have the same ident. Also, for Cl, Br, and I the distance to the next ident is chosen as 2, because valences other than −1 are too seldom for using a greater distance, keeping possible numbers free for chloric acid or other halogen compounds with valences of 3, 5, or 7. Of course, this list and the idents can easily be modified if necessary. Note, that in the computation $\sum_{i=1}^{b} B_i^2$ one double bond and four single bonds add up to the same connectivity value of 4, even though we have experienced that an explicit distinction (e.g., by using the value 7 which is otherwise unused) causes more problems than it solves. In general, a higher discrimination can be reached in the next order by the addition of four or only one neighbor atom.

One of two other, new modifications to the concept of connectivities should be mentioned here: The *index* (i.e., the internal number) of any atom is embedded in the lower digits of the integer connectivity. In the FORTRAN program the maximal number of atoms on each side of the reaction is given via the PARAMETER statement as a number nmax. The connectivity value of atom x (in general, of nth order) ${}^nCV_x$ used during the execution of the program is then derived from the original connectivity ${}^nC_x$ in eq 1 in the following way:

$$ {}^nCV_x = {}^nC_x\text{nmax} + \text{index} \qquad (2) $$

By using the MOD function in FORTRAN, the atom index,

which is frequently utilized within this program, can be extracted at once from the just calculated connectivity value. Furthermore, the original connectivity is available by integer division, namely $C = CV/nmax$. Sorting and comparing the connectivity values CV is performed with special routines since an additional representation of connectivities is used. A discussion of additional details will be included in the next section.

Connectivities of Higher Orders. The nth connectivity of an atom x is defined as five times its connectivity of the $(n-1)$th order plus the $(n-1)$th order connectivities of all *nadj* adjacent atoms:

$$^nC_x = 5^{n-1}C_x + \sum_{i=1}^{nadj} {}^{n-1}C_{x_i} \quad (3)$$

The higher weight of the reference atom gives better discrimination in high orders;[22] this method is usually applied in algorithms dealing with connectivities. Embedding of the atom index, yielding ^nCV_x, is done the same way as described above.

A nasty problem is the fast ascending magnitude of resulting numbers: A 4-byte integer, for example, can already overflow in about the fifth or sixth order, if connectivities of atoms with a high ident number and many adjacent atoms are computed. To assure a correct computation of connectivites even in high orders, we use a twofold strategy: At first, the connectivities are reduced in size before they reach a magnitude which could lead to an overflow of the 4-byte integer numbers used here. But simply "cutting off" the last digits would effect the embedded index and could make some different connectivity values equal. Therefore, we first abstract the index, reduce the number by an other representation (here we use *aaabbbccc* → *aaa* + *bbb* + *ccc*), and then reembed the atom index. However, even this method could lead to equal ^nCV_x of chemical different atoms. Therefore, the second modification of the concept of connectivities mentioned above is introduced: All connectivities are computed twice in the program, initially as the already described 4-byte integers and secondly as 8-byte real numbers. Testing the connectivities for equality is then performed by a statement function, which only return "equal" if the integer connectivities (without index) are *exactly* the same and the real numbers differ by no more than 10^{-12} (because in very high orders, significant digits can be lost in real numbers). In this way, the problem of numeric overflow or occasional equal connectivities of chemically different atoms is resolved, a problem which has never been discussed before.

Testing for Equal Connectivities. This subroutine is the recommended procedure to find equal connectivities or, if there are no more, to determine the deletion radius and the starting points for the deletion process. The necessary sorting algorithm is embedded in the program, since no normal sorting utility would handle the two connectivities computed in parallel in addition to the embedded index.

The remaining steps have been discussed in general in the last two sections.

RESULTS AND CONCLUSIONS

We have shown that the concept of connectivities, introduced by Morgan[7] and used before by Lynch and Willett[8] as well as by Funatsu et al.[4] for the analysis of machine-readable chemical reactions, can effectively be modified to achieve an exact determination of reactive centers, bonds, and groups.

The main features of this approach are

- The consequent application of only one concept for the analysis of the reaction and the exact determination of certain atoms and bonds as reactive centers and bonds, and thus not only the identification of a reaction site which contains these atoms and bonds.
- The tolerance of nonstoichiometric reactions by this approach, a prerequisite for the application of the program to reaction equations obtained from commercial databases.
- The new approach to the concept of completing a reaction.

Additional features are

- The highly discriminating method to build the connectivities of first order, using hybridization patterns as chemically meaningful information, in addition to utilizing the frequency of the occurrence of individual elements in chemical reactions.
- The use of integer connectivities modified via an atom index and real connectivity numbers to assure an exact comparison.
- The loss prevention of parts of the MCSS versus reactive groups by using the spreading of the negative sign of the connectivity values, which in an ideal way uses the concentric spreading of information by computing connectivities.

Problems may occur mainly with intramolecular reactions and reactions which are in principle single-step but modify the molecular skeleton of the substrate highly, as is the case in rearrangements. Another general problem poses multiple reactions, compounded within one reaction equation, especially when the individual steps represent different types of transformations in the sense of this algorithm.

We have tested this program with 100 arbitrarily selected reactions, which could be classified in the most general sense as single reactions, even though cases as opening a cyclic amide followed by a subsequent hydrolysis were included. In 82 cases, the program identified the reactive centers (up to six classes) correctly and exactly as well as the reactive bonds and the groups. In addition, it performed correct reaction completions. In the case of the remaining 18 reactions, there were six equations where reactions at a multiple bond occurred not within the substrate itself (i.e., within the molecule containing the MCSS) but in a separated reagent molecule. This problem is under current investigation, and efforts are now under way to resolve this shortcoming. Another four cases failed due to some structural peculiarities of the reactants: Aliphatic rings without substituents cause equal connectivities in orders up to the deletion radius, leading to ambiguous starting points for the deletion process. The remaining eight reactions failed because of different reasons, typically because they contained rearrangements.

Another 75 reactions comprised multiple reactions with more than one individual reaction at one time in one equation, although in 12 cases a satisfactory analysis was possible.

Even though this topological approach does not represent a universal algorithm to deal with all possible chemical reactions, its powerful features of an exact identification of the reactive centers and the bonds as well as the reaction completion can be used for single-step reactions which are analyzed correctly. Many attributes of a reaction (e.g., intramolecular reaction or reaction changing multiple bonds) are determined during the execution of the program, since these informations are necessary in order to apply the appropriate modifications to the algorithm. The same attributes can be used to predict a case in which the program would probably fail, e.g., the combination of an intramolecular reaction with additional modifications of multiple bonds. The already mentioned cases, which failed because of ambiguous starting points for the deletion process in large aliphatic rings, most often give results with a different number of deleted atoms in the detected MCSS, which is also a criterion for a wrong analysis. The satisfactory interpretation of all detected properties of a re-

action is an additional subject of further investigations.

ACKNOWLEDGMENT

We thank the Fonds der Chemischen Industrie and the BAYER AG for financial support. C.B. thanks the Studienstiftung des deutschen Volkes for a scholarship.

REFERENCES AND NOTES

(1) Kos, A. J.; Grethe, G. Reaktionsdatenbanken—Werkzeuge für den Synthese-Chemiker. *Nachr. Chem., Tech. Lab.* **1987**, *35*, 586–94.
(2) Zass, E.; Müller, S. Neue Möglichkeiten zur Recherche von organisch-chemischen Reaktionen—Ein Vergleich der ≪in-house≫-Datenbanksysteme REACCS, SYNLIB and ORAC. *Chimia* **1986**, *40*, 38–50.
(3) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. The CAS ONLINE Search System. 1. General System Design and Selection, Generation and Use of Search Screens. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93–102.
(4) Funatsu, K.; Endo, T.; Kotera, N.; Sasaki, S.-I. Automatic Recognition of Reaction Site in Organic Chemical Reactions. *Tetrahedron Comput. Methodol.* **1988**, *1*, 53–69.
(5) Dugundji, J.; Ugi, I. An Algebraic Model of Constitutional Chemistry as a Basis for Chemical Computer Programs. *Top. Curr. Chem.* **1973**, *39*, 19–64.
(6) For example, see the following articles and further references cited there: (a) Ugi, I., et al. Neue Anwendungsgebiete für Computer in der Chemie. *Angew. Chem.* **1979**, *91*, 99–111. (b) Jochum, C.; Gasteiger, J.; Ugi, I. Das Prinzip der minimalen chemischen Distanz. *Angew. Chem.* **1980**, *92*, 503–13. (c) Jochum, C.; Gasteiger, J.; Ugi, I. The Principle of Minimum Chemical Distance and the Principle of Minimum Structure Change. *Z. Naturforsch.* **1982**, *37B*, 1205–15. (d) Brandt, J.; Bauer, J.; Frank, R. M.; von Scholley, A. Classification of Reactions by Electron Shift Patterns. *Chem. Scr.* **1981**, *18*, 53–60. (e) Brandt, J.; von Scholley, A. An Efficient Algorithm for the Computation of the Canonical Numbering of Reaction Matrices. *Comput. Chem.* **1983**, *7*, 51–9. (f) Wochner, M.; Brandt, J.; von Scholley, A.; Ugi, I. Chemical Similarity, Chemical Distance, and its Exact Determination. *Chimia* **1988**, *42*, 217–25. (g) Fontain, E.; Bauer, J.; Ugi, I. Computer assisted Bilateral Generation of Reaction Networks from Educts and Products. *Chem. Lett.* **1987**, 37–40.
(7) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–13.
(8) Lynch, M. F.; Willett, P. The Automatic Detection of Chemical Reaction Sites. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 154–9.
(9) Hutchins, R. O.; Hoke, D.; Keogh, J.; Koharski, D. Sodium Borohydride in Dimethyl Sulfoxide or Sulfolane. Convenient Systems for Selective Reductions of Primary, Secondary and Certain Tertiary Halides and Tosylates. *Tetrahedron Lett.* **1969**, 3495–8.
(10) Hojo, K.; Yoshino, H.; Mukaiyama, T. New Synthetic Reactions Based on 1-Methyl-2-fluoropyridinium Salts. Facile Conversion of Alcohols to Thioalcohols. *Chem. Lett.* **1977**, 133–6.
(11) Bates, G. S. New α-Keto Acid Synthon; Alkylation of the Potassium Dianion of Bis(ethylthio)acetic Acid. *J. Chem. Soc., Chem. Commun.* **1979**, 161–3.
(12) Mukaiyama, T.; Yamaguchi, M.; Narasaka, K. A Regioselective Coupling Reaction of Allyl Pyridyl Ethers with Grignard Reagents. *Chem. Lett.* **1978**, 689–92.
(13) Solas, D.; Wolinsky, J. Total Synthesis of (−)-α-Acoradiene and (−)-α-Cedrene. *J. Org. Chem.* **1983**, *48*, 670–3.
(14) Hiyama, T.; Wakasa, N. Asymmetric Coupling of Asrylmagnesium Bromides with Allylic Esters. *Tetrahedron Lett.* **1985**, *26*, 3259–62.
(15) Fried, J.; Hallinan, E. A.; Szwedo, M. J. Synthesis and Properties of 7,7-DifluoroDerivatives of the 2,6-Dioxa[3.1.1]bicycloheptane Ring System Present in Thromboxane A2. *J. Am. Chem. Soc.* **1984**, *106*, 3871–2.
(16) Huggins, J. M.; Bergman, R. G. Mechanism, Regiochemistry, and Stereochemistry of the Insertion Reaction of Alkynes with Methyl(2,4-pentanedionato)(triphenylphosphine)nickel. A Cis Insertion that Leads to Trans Kinetic Products. *J. Am. Chem. Soc.* **1981**, *103*, 3002–11.
(17) Hart, H.; Sasaoka, M. Exocyclic Benzenes. Synthesis and Properties of Benzo[1,2-*c*:3,4-*c'*:5,6-*c''*]trithiophene, a Tristhiahexaradialene. *J. Am. Chem. Soc.* **1978**, *100*, 4326–7.
(18) Cacioli, P.; Reiss, J. A. The Formation and Some Reactions of a Spirocyclic Chroman Derived from a 1-Oxaspiro[2.5]octa-5,7-dien-4-one. *Aust. J. Chem.* **1984**, *37*, 2599–605.
(19) Molecular Design Limited, 2132 Farallon Drive, San Leandro, CA 94577, or Molecular Design MDL AG, Wallstrasse 8, CH-4002 Basel, Switzerland.
(20) Borkent, J. H.; Onkes, F.; Noordik, J. H. Chemical Reaction Searching Compared in REACCS, SYNLIB, and ORAC. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 148–50.
(21) Bebak, H.; et al. The Standard Molecular Data Format (SMD Format) as an Integration Tool in Computer Chemistry. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 1–5.
(22) Shelley, C. A.; Munk, M. E. Computer Perception of Topological Symmetry. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 110–3.

Smiles. 3. Depict. Graphical Depiction of Chemical Structures

DAVID WEININGER

Daylight Chemical Information Systems, 111 Rue Iberville, No. 610, New Orleans, Louisiana 70130

Received February 21, 1989

The DEPICT program converts SMILES, the linear notation of a chemical structure's molecular graph, into a depiction of molecular structure without user interaction. The resulting two-dimensional output display allows all aspects of SMILES representation of structure to be verified easily, including aromaticity, formal charge, bond order assignment, and hydrogen attachment. DEPICT is particularly well suited for computer-generated structures since it requires no manual or structural input. It is designed for use with SMILES notation, a lexical form of a connection table, and it follows that any other connection table can also be used with DEPICT provided only that its input is first converted to SMILES.

INTRODUCTION

Computer graphics of chemical structures and formulas are most important for the interaction between chemist and computer, particularly in the fields of organic synthesis and substructure searching. What appears to be well within the scope of modern technology is surprisingly difficult, because different conventional methods of presenting structures pictorially do not always follow the same rules. Ambiguities and exceptions are often encountered. Publications on this subject[1,2] generally deal with computer-interactive graphics by electronic input with a stylus on a tablet, or with light pens on graphical CRT. Originally, at Chemical Abstracts Service, graphical data were processed manually. As computer processing of two-dimensional graphs improved, structural representations were standardized and files were established for molecular fragments as well as for complete structures.[1] In some cases it was possible to store essential information for creating a diagram separately from connection tables. But for complex structures in large databases this was impractical, so the problem of displaying structural diagrams from connection tables had to be addressed. A procedure developed by Shelley[3] involved an initial perception of structural features as a data tree. This was followed by generating atoms, ring structures, and their connecting bonds with graph-invariant codes. They were used

0095-2338/90/1630-0237$02.50/0