(4) Gasteiger, J.; Marsili, M.; Hutchings, M. G.; Saller, H.; Low, P.; Rose, P.; Rafeiner, K. Models for the Representation of Knowledge about Chemical Reactions. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 467-476.
(5) Asbrink, L.; Fridh, C.; Lindholm, E. HAM/3, a Semiempirical MO Theory. The SCF Method. *Chem. Phys. Lett.* **1977**, *52*, 63-68. This paper reports a similar approach for the calculation of electronic energies. This approach uses the semiempirical MO theory and represents a good precedent for our method.
(6) (a) Nalewajski, R. F. Recursive Combination Rules for Molecular Hardnesses and Electronegativities. *J. Phys. Chem.* **1989**, *93*, 2658-2666. (b) Nalewajski, R. F.; Korchowiec, J.; Zhou, Z. Molecular Hardness and Softness Parameters and Their Use in Chemistry. *Int. J. Quantum Chem., Quantum Chem. Symp.* **1988**, *22*, 349-366 and references cited therein.
(7) (a) Komorowski, L. Electronegativity and Hardness in the Chemical Approximation. *Chem. Phys.* **1987**, *114*, 55-71. (b) Komorowski, L. Chemical Hardness and L. Pauling Scale of Electronegativity. *Z. Naturforsch.* **1987**, *42a*, 767-773. (c) Komorowski, L. Empirical Evaluation of Chemical Hardness. *Chem. Phys. Lett.* **1987**, *134*, 536-540 and references cited therein.
(8) Baumer, L.; Sala, G.; Sello, G. A. New Method for the Calculation of Atomic and Local Hardness. *J. Comput. Chem.* **1990**, *11*, 694-699.
(9) (a) Pearson, R. G. Absolute Hardness: Companion Parameter to Absolute Electronegativity. *J. Am. Chem. Soc.* **1983**, *105*, 7512-7516. (b) Pearson, R. G. Absolute Electronegativity and Absolute Hardness of Lewis Acids and Bases. *J. Am. Chem. Soc.* **1985**, *107*, 6801-6806. (c) Yang, W.; Lee, C.; Ghosh, S. K. Molecular Softness as the Average of Atomic Softness: Companion Principle to the Geometric Mean Principle for Electronegativity Equalization. *J. Phys. Chem.* **1985**, *89*, 5412-5144. (d) Pearson, R. G. Absolute Electronegativity and Hardness Correlated with Molecular Orbital Theory. *Proc. Natl. Acad. Sci. USA* **1986**, *83*, 8440-8441. (e) Bohm, M. C.; Schmidt, P. C. Electronegativities and Hardnesses of the Mean Group Elements from Density Functional Theory: Dependence on the Hybridization of the Chemical Bond. *Ber. Bunsen-Ges. Phys. Chem.* **1986**, *90*, 913-919. (f) Orsky, A. R.; Whitehead, M. A. Electronegativity in Density Functional Theory: Diatomic Bond Energies and Hardness Parameters. *Can. J. Chem.* **1987**, *65*, 1970. (g) Pearson, R. G. Chemical Hardness and Bond Dissociation Energies. *J. Am. Chem. Soc.* **1988**, *110*, 7684-7690. (h) Pearson, R. G. Absolute Electronegativity and Hardness: Applications to Organic Chemistry. *J. Org. Chem.* **1989**, *54*, 1423-1430.
(10) (a) Gordy, W. A. New Method of Determining Electronegativity from Other Atomic Properties. *Phys. Rev.* **1946**, *69*, 604-607. (b) Gordy, W.; Thomas, W. J. O. Electronegativities of the Elements. *J. Chem. Phys.* **1956**, *24*, 439-444.
(11) Pauling, L. *The Nature of the Chemical Bond*, 3rd ed.; Cornell University: Ithaca, NY, 1960.
(12) Bader, R. F. W. Atoms in Molecules. *Acc. Chem. Res.* **1985**, *18*, 9 and references cited therein.
(13) Thresholds are energy differences and could have a dimension, but they are used dimensionless to free them from any precise physical meaning; therefore maintaining a relation with the thresholds used in the previous approach.[3]
(14) (a) Edgecombe, K. E.; Boyd, R. J. Molecular Orbital Treatment of Substituent Effects. I. Structures of Some Carbon Acids and Their Conjugated Bases. *Can. J. Chem.* **1983**, *61*, 45. (b) Ab initio results for cations were obtained using the program HONDO8 (included in the MOTECC package) at the minimal basis set (STO3G).
(15) Ampac: Austin Method 1 Package; QCPE Program No. 527.
(16) As our algorithm calculating NPs works only with energy differences, this correction is not required.
(17) The chemical evidence is mostly taken from: March, J. *Advanced Organic Chemistry*, 3rd ed.; Wiley-Interscience: New York, 1985.
(18) Chandrasekhar, J.; Andrade, J. G.; Schleyer, P. von Ragué Efficient and Accurate Calculation of Anion Proton Affinities. *J. Am. Chem. Soc.* **1981**, *103*, 5609-5612.

Generation of Molecular Graphs for QSAR Studies: An Approach Based on Acyclic Fragment Combinations

S. S. TRATCH, O. A. LOMOVA, D. V. SUKHACHEV, V. A. PALYULIN, and N. S. ZEFIROV*
Department of Chemistry, Moscow State University, Moscow 119899, USSR

Received July 15, 1991

A generating algorithm for substituted derivatives of a given structure is elaborated for the purpose of QSAR studies, and the complex substituents are, in turn, constructed from given sets of elementary fragments. Elementary as well as composite fragments are classified into terminal, linear, and branched ones; three types of combining operations are introduced for these fragments. The combining operations make it possible to correctly generate the complete sets of substituents with prescribed values of several numerical characteristics. A detailed description of the generating procedure is presented. The application of graph and permutation group theories shows that only automorphism groups of some elementary branched fragments are required for nonduplicate construction of all substituents. The related analytical enumeration problems are solved on the basis of Burnside's Lemma. Additional selection criteria and internal representations of fragments are also discussed. The WLN codes of fragments are actually used in computer implementation of the suggested algorithm.

1. INTRODUCTION

During the last decade the interest in the problem of computer-aided design of organic structures with prescribed properties has significantly increased. The investigation of quantitative structure–activity relationships (QSAR) has become a popular and often powerful method for predicting many kinds of biological activity.[1,2] In the most typical situation, the specific parent structure together with a certain set of substituents is initially found to be responsible for the desired activity. At the second stage, some structure generator is needed to search for new active derivatives of the parent "skeleton". It is significant that the enlarged set consisting of generated target structures can lead to the refinement of the initial structure–activity model owing to detection of new highly favored or disfavored substituents.

The well-known generation techniques are oriented, however, either to the constructive enumeration of nonisomorphic abstract graphs[3-5] or to the construction of acyclic[6-8] or cyclic[9-16] molecular graphs corresponding to a given gross formula. Such generators seem to be quite inadequate for the QSAR studies because not the gross formula but the parent skeleton must be present in all generated target structures. On the other hand, the broad class of relatively complex substituents (having been, in turn, constructed of atoms and simple groups) seems to be preferable as compared to the fixed set of given substituents. It results in some similarity between desired and "normal" graph generators; in particular, all generators must produce only nonduplicate target structures with the total number of structures being dependent on certain limitations.

In this paper, we describe our efforts to elaborate the fragment generator which makes it possible to sequentially construct all *acyclic* substituents from a given set of "elementary" fragments. The idea of the suggested algorithm is somewhat similar to that used in ref 11, but in our approach the elementary fragments are allowed to be represented not only by atoms but also by arbitrary multiatomic groups. The

0095-2338/92/1632-0130$03.00/0

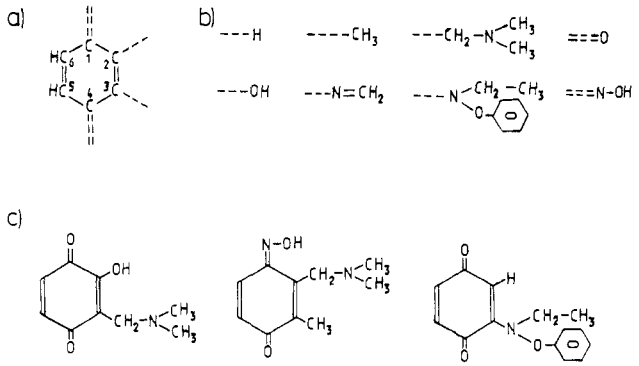

Figure 1. Central fragment (a), substituents (b), and three substituted *p*-quinone derivatives (c).

mathematical statement of generation and enumeration problems as well as other related topics are also discussed throughout this paper.

2. SCOPE OF THE PROBLEM

The target structures for the QSAR studies are thought here to contain one and the same *central fragment*. If the sets of proper *substituents* are predefined and the symmetry of the central fragment is known, then the generation problem for target structures can be formulated as a constructive enumeration problem (see Section 6 for a rigorous mathematical statement). The effective methods for computer-assisted solution of similar problems have been elaborated[17,18] and applied to generating substitution derivatives for the given parent structures (see, e.g., refs 19–21).

Figure 1 represents an illustrative example of the central fragment together with the set of univalent and divalent substituents and three target *p*-quinone derivatives. It should be noted that free valencies must not necessarily be assigned to all positions of the parent skeleton (cf. Figure 1a); this fact usually results in decreased symmetry of the central fragment. Moreover, in some practically useful cases distinct sets of substituents can be associated even with equivalent valencies of the central fragment, and this leads to further reduction of its symmetry. For example, the generation problem for 1,3,5-trisubstituted benzenes with alkyl substituents in position 1, acyl substituents in position 3, and hetero substituents in position 5 needs the completely unsymmetrical central fragment to be considered. The graphs corresponding to such complicated cases are discussed in Section 6.

The specific feature of our approach consists of the fact that substituents are also treated as consisting of relatively small *elementary fragments* (EFs). For this reason, a separate enumeration problem for substituents arises; the corresponding generating algorithm is a central point of this paper.

The decomposition of two univalent and one divalent substituents (*N*,*N*-dimethylaminomethyl, methyleneimino, and oxime groups, cf. Figure 1b) into EFs is shown in Figure 2a. One can easily see that the EFs of several levels may be observed in the decomposition process: the elementary fragments of the first level always contain the free valency of the substituent together with one or more valencies leading to EFs of the second level, etc. For this reason, the most convenient way to represent EFs consists of subdividing the free valencies into two subclasses. The first subclass contains a single free valency leading to the EF of the preceding level (or to the central fragment), while the second subclass consists of one or more valencies leading to the EFs of the next level. The free valencies of the first subclass are designated by *out-arrows* (starting at the atom), while the valencies of the second subclass are designated by *in-arrows* (ending at the atom). The recomposition of any substituent obviously consists of replacement of one in-arrow and one out-arrow of the same

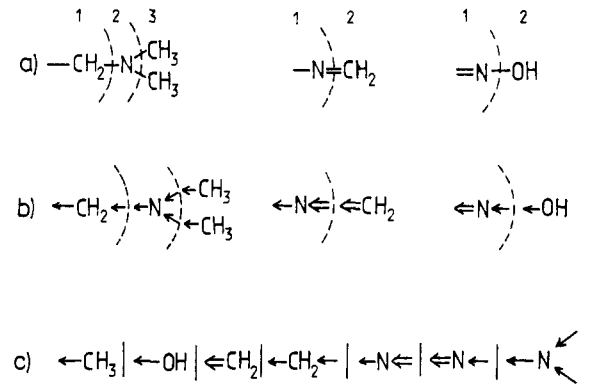

Figure 2. EFs as building blocks of composite fragments.

multiplicity by the chemical bond (cf. Figure 2, parts a and b).

Two significant notes should be made. First of all, two or more EFs can be associated with one and the same unsymmetrical chemical group due to different localizations of in- and out-arrows. Two EFs (←CH_2-O← and ←O-CH_2←) corresponding to an oxymethylene group can serve as an example; another example (⇐N← and ←N⇐ elementary fragments) can be found in Figure 2b. The complete set of EFs being required to generate the three substituents of Figure 2a, is represented in Figure 2c.

Secondly, the sets of EFs are not thought to contain only simple chemical groups that consist of one non-hydrogen atom. These sets are also allowed to contain EFs which are unbranched (e.g., ←CH_2-O←) or branched chains, cycles (e.g., ←C_6H_5 or ←C_6H_4←), and even polycyclic structures. An additional *intersection problem* arises, however, if the complex multiatomic fragments are included in the set of EFs. This problem originates from the fact that the combination of two or more "small" EFs can be identical to the "large" EF; several different combinations of multiatomic EFs can also lead to identical results. The simple examples are the following: ←CO← and ←OH EFs produce a ←COOH EF; the combination of ←CH_2CH_2← and ←O← EFs leads to the same result as the combination of ←CH_2← and ←CH_2O← EFs. Strictly speaking, the mathematical model and generation algorithm of Sections 4–6 do not correctly solve the intersection problem. The multiatomic fragments are allowed, however, to be included in the sets of given EFs. In this case, the identical results are successfully recognized by an additional tool of our approach; the WLN codes (see Section 7) are used for that purpose.

Although the elementary fragments can contain cycles, only *acyclic combinations* of EFs are considered here. This means that no additional cycles are allowed to be formed when the pairs of in- and out-arrows are replaced by bonds. For this reason, all the fragments having been generated from EFs are denoted here as *acyclic composite fragments* (ACFs). The composition rules for EFs and ACFs are discussed in the next section.

The above considerations make it possible to formulate the generation problem for target structures as consisting of two subproblems:

(i) construction of the complete sets of substituents corresponding to nonequivalent positions of a given central fragment

(ii) construction of the target structures by assigning permissible substituents to all positions of the central fragment

3. COMPOSITION RULES FOR EFS AND ACFS

Prior to exactly defining the operations over EFs and ACFs, a simple classification scheme and the numerical characteristics

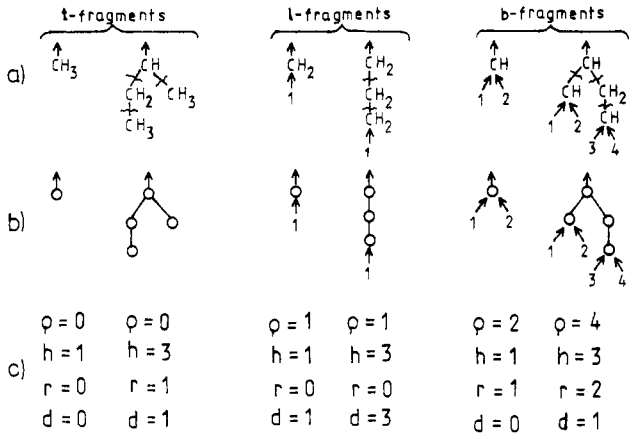

Figure 3. Examples, pictorial representations, and numerical characteristics of t-, l-, and b-fragments.

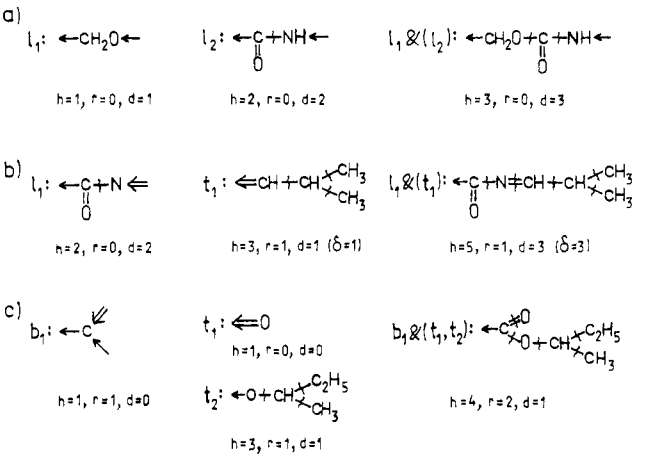

Figure 4. Illustrative examples of $l_1\&(l_2)$, $l_1\&(t_1)$, and $b_1\&(t_1, t_2, ..., t_\rho)$ combining operations.

of the fragments must be discussed. The suggested classification is based on the count of in-arrows in EFs and ACFs. Thus, the fragments which contain no in-arrows are named *terminal* and denoted as t-fragments. Similarly, EFs and ACFs with exactly one or two or more in-arrows are called *linear* (l-) or *branched* (b-) fragments. The t-, l-, and b-fragments always contain a single out-arrow. Examples of elementary and composite fragments are presented in Figure 3a. For completeness, the *central* fragment (with no out-arrows and one or more in-arrows) is thought to be a specific kind of fragment and is denoted as c-fragment.

The further classification of fragments which takes into account the multiplicities of out- and in-arrows will be also used when needed. Thus, the sets of fragments $t^{(i)}$ and $b^{(i)}$ (i = 1, 2, or 3 denotes the multiplicity of an out-arrow) and sets of fragments $l^{(ij)}$ (i = 1, 2, or 3 and j = 1, 2, or 3 denote multiplicities of out- and in-arrows, respectively) are used in the next section.

For each elementary or composite fragment, a specific pictorial representation can be drawn. In these representations (shown in Figure 3b) all arrows and all chemical bonds linking the pairs of EFs are conserved, although the "internal" structures of all EFs disappear. An examination of Figure 3b shows that all elementary fragments always contain a single node; in contrast, composite t-, l-, and b-fragments necessarily contain two or more nodes. Obviously, the composite l-fragments are always unbranched, composite b-fragments are always branched, and composite t-fragments are either branched or unbranched.

Several numerical characteristics of EFs and ACFs need to be introduced for the following discussion; these characteristics may be directly observed by the inspection of pictorial representations. The first of them, the *degree* ρ of a fragment is defined as a number of in-arrows in a given EF or ACF (ρ = 0 for terminal fragments, ρ = 1 for linear fragments, $\rho \geq$ 2 for branched fragments; the multiplicities of arrows are not taken into account). We note here that the following discussion needs all in-arrows of EFs and ACFs to be arbitrarily numbered; the possible numberings are shown in Figure 3, parts a and b.

The discussion of other numerical characteristics needs a node with the single out-arrow, in a pictorial representation, to be regarded as the root (or first level) node; the corresponding EF may also be considered as the root or first level EF. The *height* h of any EF or ACF is defined as the total number of nodes in the longest path connecting the root of a pictorial representation with a node of the maximal level (obviously, h = 1 for elementary t-, l-, and b-fragments). Similarly, the *rank* r of a fragment is defined as a maximal number of branching nodes in the paths; all paths connecting the root with terminal nodes must be taken into account. Evidently, r = 0 for unbranched t- and all l-fragments. Finally, the *disperse* d of any fragment is defined as the maximal number of adjacent nodes of valency 2 in a pictorial representation. The value of d evidently depends upon the length of the longest subchain consisting only of l-fragments. Figure 3c summarizes all the above characteristics for the fragments of Figure 3a. We once more stress that internal structures of EFs must not be taken into account (thus, h = 1, r = 0, and d = 1 for a $\leftarrow CH_2O\leftarrow$ group if it is considered as an elementary l-fragment).

The substituents may evidently be regarded as elementary or composite t-fragments with the allowed values of h, r, and d. The accepted restrictions for r and d are discussed in the next section. The generation of composite t-fragments out of given EFs requires the composition rules for various t-, l-, and b-fragments to be exactly formulated. The specific feature of the generating algorithm consists of the fact that it needs only three of the many formally possible combinations of fragments to be considered. These combinations are formulated here as a result of a specific *operation* which is designated by &.

The combining operation for the fragment x of degree $\rho >$ 0 (thus, x is either an l- or b-fragment) and the ordered sequence $(y_1, y_2, ..., y_\rho)$ with y_i, i = 1, 2, ..., ρ, being identical or nonidentical t-, l-, or b-fragments, is denoted as $x\&(y_1, y_2, ..., y_\rho)$. This operation is always defined if the multiplicity of every in-arrow (numbered by i) of the fragment x coincides with the multiplicity of the single out-arrow of the corresponding fragment y_i, i = 1, 2, ... ρ. Six simple combinations $l_1\&(l_2)$, $l_1\&(t_1)$, $l_1\&(b_1)$, $b_1\&(t_1, t_2, ..., t_\rho)$, $b_1\&(l_1, l_2, ..., l_\rho)$, $b_1\&(b_1, b_2, ..., b_\rho)$ [and many mixed combinations $b_1\&(y_1, y_2, ..., y_\rho)$ with $y_1, y_2, ..., y_\rho$ being the fragments of different types] are formally possible. Only the first, second, and fourth of the simple combinations are used in the suggested generation procedure; these combinations are discussed in detail.

$l_1\&(l_2)$ Combinations. The combining operation for two elementary or composite fragments, l_1,l_2, produces the single composite fragment $l = l_1\&(l_2)$; see Figure 4a for an example. The out- and in-arrows of the resulting fragment l are originated by an out-arrow of l_1 and by an in-arrow of l_2, respectively. An in-arrow of l_1 and out-arrow of l_2 are replaced by a chemical bond of proper multiplicity. It is evident that the height and disperse values are summed up (eqs 1a,b) while the rank values are equated to 0 in the resulting l-fragments (eq 1c).

$$h(l) = h(l_1) + h(l_2) \quad (1a)$$

$$d(l) = d(l_1) + d(l_2) \quad (1b)$$

$$r(l) = r(l_1) = r(l_2) = 0 \quad (1c)$$

$l_1\&(t_1)$ Combinations. The combining operation for elementary or composite fragments l_1 and t_1 produces the single composite fragment $t = l_1\&(t_1)$ with the out-arrow originated by that of l_1. The new chemical bond arises owing to replacement of an in-arrow of l_1 and of an out-arrow of t_1. An illustrative example is shown in Figure 4b. The height, disperse, and rank values of the resulting terminal fragments are defined by eqs 2a–c; the value of δ in eq 2b denotes the length

$$h(t) = h(l_1) + h(t_1) \quad (2a)$$

$$d(t) = \max[d(l_1) + \delta, d(t_1)] \quad (2b)$$

$$r(t) = r(t_1) \quad (2c)$$

of a subchain consisting of linear EFs located at the root of the composite t-fragment. We note that t-fragments with $\delta \neq 0$ (exemplified by a fragment of Figure 4b) are never combined with l-fragments in an algorithmic implementation of $l_1\&(t_1)$ operation.

$b_1\&(t_1, t_2, ..., t_\rho)$ Combinations. In this case $\rho = \rho(b_1)$ new chemical bonds are produced by replacement of ρ pairs of arrows (namely, of an ith in-arrow of b_1 and of a single out-arrow of t_i, $i = 1, 2, ..., \rho$). Each pair of arrows must be of the same multiplicity; see Figure 4c for an example. The resulting terminal fragment $t = b_1\&(t_1, t_2, ..., t_\rho)$ always contains a single out-arrow (originated by b_1); we note that only elementary fragments b_1 are actually involved in the algorithmic implementation of this combining operation. For this reason, eqs 3a–c can be used for calculating numerical characteristics of the resulting t-fragment.

$$h(t) = 1 + \max[h(t_1), h(t_2), ..., h(t_\rho)] \quad (3a)$$

$$d(t) = \max[d(t_1), d(t_2), ..., d(t_\rho)] \quad (3b)$$

$$r(t) = 1 + \max[r(t_1), r(t_2), ..., r(t_\rho)] \quad (3c)$$

4. GENERATION PROCEDURE FOR SUBSTITUENTS

In the preceding sections only the compositions of individual fragments have been considered; the ordered sequences were also thought to consist of preselected, possibly identical, fragments. The production of complex substituents (i.e., terminal ACFs) needs, however, the sets of various fragments to be formed in many steps of the generation procedure. In order to illustrate the operations with fragment sets, the corresponding analytical enumeration problems will be initially solved.

Let $T^{(1)} = \{t^{(1)}\}$, $|T^{(1)}| = p_1$, $T^{(2)} = \{t^{(2)}\}$, $|T^{(2)}| = p_2$, and $T^{(3)} = \{t^{(3)}\}$, $|T^{(3)}| = p_3$ be separate sets consisting of elementary or composite t-fragments with multiplicities of out-arrows being equal to 1, 2, or 3, respectively. Similarly, $B^{(1)} = \{b^{(1)}\}$, $|B^{(1)}| = q_1$, $B^{(2)} = \{b^{(2)}\}$, $|B^{(2)}| = q_2$, and $B^{(3)} = \{b^{(3)}\}$, $|B^{(3)}| = q_3$ are the sets of given b-fragments with multiplicities of out-arrows equal to 1, 2, or 3; the total number of in-arrows of the jth fragment $b_j^{(i)} \in B^{(i)}$, $i = 1, 2$, or 3, $j = 1, 2, ..., q_i$, is denoted here by $\rho(ij)$. Finally, nine sets $L^{(ij)} = \{l^{(ij)}\}$, $|L^{(ij)}| = n_{ij}$, $i = 1, 2$, or 3, $j = 1, 2$, or 3, are thought to be given. The set $L^{(ij)}$ consists of n_{ij} fragments with multiplicity of out-arrows equal to i ($i = 1, 2, 3$) and with multiplicity of in-arrows equal to j ($j = 1, 2, 3$). The schematic representations of the sets $T^{(i)}$, $B^{(i)}$, and $L^{(ij)}$ are depicted in Figure 5. We note that the empty sets $L^{(23)}$, $L^{(32)}$, and $L^{(33)}$ ($n_{23} = n_{32} = n_{33} = 0$) are usually formed if only simple monoatomic fragments are given.

Now we try to estimate the number x of composite l-fragments and the numbers y and z of composite t-fragments [constructed by means of $l_1\&(t_1)$ and $b_1\&(t_1, t_2, ..., t_\rho)$ combining operations, respectively]. Simple combinatorial considerations (cf. Figure 5) show that all combinations of the fragments from $L^{(ik)}$ with the fragments from $L^{(kj)}$ (i, j, and k are equal to 1, 2, or 3) are allowed and lead to $n_{i1}n_{1j} + n_{i2}n_{2j} + n_{i3}n_{3j}$ composite l-fragments with multiplicity of out-arrows and of in-arrows equal to i and j, respectively. Thus, the total number of "two-component" linear ACFs may be calculated by eq 4a.

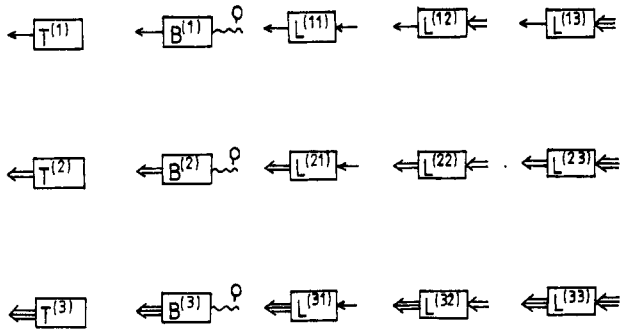

Figure 5. Pictorial representations of fragment sets.

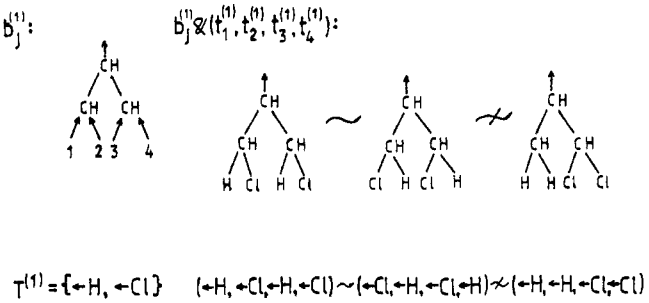

Figure 6. Examples of equivalent and nonequivalent $b_j^{(1)}\&(t_1^{(1)}, t_2^{(1)}, t_3^{(1)}, t_4^{(1)})$ combinations.

$$x = \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3} n_{ik}n_{kj} \quad (4a)$$

$$y = \sum_{i=1}^{3}\sum_{j=1}^{3} n_{ij}p_j \quad (4b)$$

$$z = \sum_{i=1}^{3}\sum_{j=1}^{3} q_i 1 / |H_{ij}| [\sum_{k=1}^{|H_{ij}|} \prod_{l=1}^{3} p_l^{\rho(h_{ijk})}] \quad (4c)$$

Similarly, all combinations of l-fragments from $L^{(ij)}$ with t-fragments from $T^{(j)}$ ($i = 1, 2, 3$; $j = 1, 2, 3$) are allowed and lead to $n_{i1}p_1 + n_{i2}p_2 + n_{i3}p_3$ composite t-fragments with multiplicity of out-arrows equal to i. Thus, the total number of "two-component" terminal ACFs may be calculated by eq 4b.

The results obtained show that the sets of l-fragments and the sets of t-fragments produced by $l_1\&(t_1)$ operation require a very simple procedure to be generated; the application of eqs 4a and 4b makes it also possible to sequentially count the number of similar fragments consisting of three or more EFs. The situation is much more complex, however, with the last combining operation $b_1\&(t_1, t_2, ..., t_\rho)$.

The number of possible combinations of a single (jth) fragment $b_j^{(i)} \in B^{(i)}$ with t-fragments of the sets $T^{(1)}$, $T^{(2)}$, and $T^{(3)}$ will be first estimated. The total number of ordered sequences $[t_1, t_2, ..., t_{\rho(ij)}]$ is evidently equal to $p_1^{a_1}p_2^{a_2}p_3^{a_3}$ (a_k, $k = 1, 2$, or 3, denotes the number of in-arrows with multiplicity k in a b-fragment under discussion). These sequences can be either equivalent or nonequivalent, and this depends upon the symmetry of $b_j^{(i)}$. A very simple example is presented in Figure 6. This example demonstrates that the ordered sequences of terminal EFs [(←H,←Cl,←H,←Cl) and (←Cl,←H,←Cl,←H)] must be regarded as equivalent because they lead to just the same composite t-fragment; these sequences are nonequivalent, however, to the sequence (←H,←H,←Cl,←Cl). Thus, a more rigorous treatment is needed in this case.

The permutation group considerations, based on the known Burnside's Lemma, make it possible to solve this problem. In Section 5, the number of nonequivalent sequences (and, hence,

the number of resulting t-fragments corresponding to a single fragment $b_j^{(i)}$) is proved to be equal to $1/|H|\sum_{k=1}^{k=|H|} p_1^{c_1(h)} p_2^{c_2(h)} p_3^{c_3(h)}$ [H = {h} is the permutation group characterizing the symmetry of $b_j^{(i)}$; $c_1(h)$, $c_2(h)$, and $c_3(h)$ are the numbers of cycles corresponding to in-arrows of multiplicity 1, 2, and 3, respectively]. The generalization of this formula for all q_i fragments $b_j^{(i)}$ of the same set and for all sets $B^{(i)}$, i = 1, 2, or 3, leads to the quite complicated eq 4c which makes it possible to count the desired number of "[$\rho(ij)$ + 1]-component" t-fragments.

The algorithm for the sequential generation of all possible substituents may be formulated as follows. The starting sets $T^{(i)}$, $B^{(i)}$, $L^{(ij)}$, i = 1, 2, or 3, j = 1, 2, or 3, are thought to be given. All l-fragments with the disperse value d, not exceeding the given value $d_{max} = M$, are initially generated and accumulated in the sets $L^{(ij)}$, i = 1, 2, or 3, j = 1, 2, or 3 (in steps 1–4 of the detailed description). All unbranched (r = 0) t-fragments are then generated and accumulated in the sets $T^{(i)}$, i = 1, 2, or 3 (in steps 5 and 6 of the detailed description). The generation of branched ($r > 0$) t-fragments consists of combining b-fragments of degree ρ from $B^{(i)}$, i = 1, 2, or 3, with the appropriate nonequivalent sequences (t_1, t_2, ..., t_ρ) of previously constructed t-fragments from $T^{(i)}$, i = 1, 2, or 3 (in steps 8 and 9), with subsequent "enlarging" of resulting t-fragments by l-fragments from $L^{(ij)}$, i = 1, 2, or 3, j = 1, 2, or 3 (in steps 10 and 11). The obtained t-fragments with the fixed non-zero value of rank (r = 1, r = 2, etc.) are sequentially accumulated in the sets $T^{(i)}$, i = 1, 2, or 3, and this process is repeated until all t-fragments with the rank value r, not exceeding the given value $r_{max} = N$, are constructed. The main feature of the suggested algorithm consists in the fact that only the symmetries of elementary b-fragments of $B^{(i)}$, i = 1, 2, or 3 (but not the symmetries of intermediate t-fragments), need to be known.

Stepwise Description of the Algorithm.

S1. [s denotes the current disperse value for l-fragments being constructed in steps 2 and 3]
s = 2;

S2. [9 auxiliary sets $X^{(ij)}$ consist of l-fragments with $d = s$ and with multiplicities of out- and in-arrows equal to i and j, respectively]
$\forall\, i \in \{1, 2, 3\}, \forall\, j \in \{1, 2, 3\}$ $X^{(ij)} = \{l^{(i1)} \& (l^{(1j)})\} \cup \{l^{(i2)} \& (l^{(2j)})\} \cup \{l^{(i3)} \& (l^{(3j)})\}$ with $l^{(ik)} \in L^{(ik)}$, $d(l^{(ik)}) = 1$ and $l^{(kj)} \in L^{(kj)}$, $d(l^{(kj)}) = s - 1$, k = 1, 2, 3;

S3. [l-fragments with $1 \le d \le s$ are collected in the sets $L^{(ij)}$]
$\forall\, i \in \{1, 2, 3\}, \forall\, j \in \{1, 2, 3\}$ $L^{(ij)} = L^{(ij)} \cup X^{(ij)}$;

S4. [all l-fragments with $1 \ge d \le M$ are constructed if $s = M$]
if $s < M$ then $s = s + 1$ and go to step 2;

S5. [3 auxiliary sets $Y^{(i)}$ consist of unbranched t-fragments with multiplicities of out-arrows equal to i]
$\forall\, i \in \{1, 2, 3\}$ $Y^{(i)} = \{l^{(i1)} \& (t^{(1)})\} \cup \{l^{(i2)} \& (t^{(2)})\} \cup \{l^{(i3)} \& (t^{(3)}\}$ with $l^{(ij)} \in L^{(ij)}$, $t^{(j)} \in T^{(j)}$, j = 1, 2, 3;

S6. [all elementary and unbranched composite t-fragments are collected in the sets $T^{(i)}$]
$\forall\, i \in \{1, 2, 3\}$ $T^{(i)} = T^{(i)} \cup Y^{(i)}$;

S7. [s denotes the current rank value for t-fragments being constructed in steps 8 and 9 and in steps 10 and 11]
s = 1;

S8. [composite t-fragments with $r = s$ form sets $Z_j^{(i)}$; each set $Z_j^{(i)}$ corresponds to jth b-fragment of $B^{(i)}$, and consists of t-fragments with multiplicity of out-arrows equal to i; m_k denotes the multiplicity of kth in-arrow of $b_j^{(i)}$]
$\forall\, i \in \{1, 2, 3\}, \forall\, j \in \{1, 2, ..., q_i\}$ $Z_j^{(i)} = \{b_j^{(i)} \& (t_1, t_2, ..., t_\rho)$ such as $t_k \in T^{(m_k)}$, k = 1, 2, ..., $\rho = \rho(ij)\}$; the ordered sequences (t_1, t_2, ..., t_ρ) must be nonequivalent and contain one or more fragments t_k, for which $r(t_k) = s - 1$;

S9. [t-fragments with $0 \le r \le s$ are collected in the sets $T^{(i)}$]
$\forall\, i \in \{1, 2, 3\}$ $T^{(i)} = T^{(i)} \cup Z_1^{(i)} \cup Z_2^{(i)} \ldots \cup Z_{q_i}^{(i)}$

S10. [3 sets $Y^{(i)}$ consist of "enlarged" t-fragments with $r = s$ and with multiplicities of out-arrows equal to i]
$\forall\, i \in \{1, 2, 3\}$ $Y^{(i)} = \cup_{j=1}^{j=q_i} (\{l^{(i1)} \& (z_j^{(1)})\} \cup \{l^{(i2)} \& (z_j^{(2)})\} \cup \{l^{(i3)} \& (z_j^{(3)})\})$ with $z_j^{(i)} \in Z_j^{(i)}$, $l^{(ik)} \in L^{(ik)}$, k = 1, 2, 3;

S11. ["enlarged" t-fragments with the rank values equal to s are collected in the sets $T^{(i)}$]
$\forall\, i \in \{1, 2, 3\}$ $T^{(i)} = T^{(i)} \cup Y^{(i)}$;

S12. [all t-fragments with $d \le M$, $r \le N$ are constructed if $s = N$]
if $s < N$ then $s = s + 1$ and go to step 8, otherwise stop.

It is easy to see that the above algorithm generates the terminal ACFs with the maximal height value $h_{max} = (M + 1)(N + 1)$; this means, however, that not all such fragments but only those with $d \le M$, $r \le N$ are actually constructed.

The number of generated substituents is evidently dependent on the number of elementary EFs and on the values of M and N. This number increases drastically with increased cardinalities of the sets $T^{(i)}$, $B^{(i)}$, $L^{(ij)}$, i = 1, 2, or 3, j = 1, 2, or 3 (see eqs 4a–c), even for small M and N. To avoid the combinatorial explosion, additional *selection criteria* need to be elaborated. The selection criteria are also thought to exclude the generation of those substituents which are unrealistic from a chemist's point of view or do not satisfy the chosen model of activity. Two of many possible types of such criteria will be briefly discussed in Section 7.

5. APPLICATION OF GRAPH AND PERMUTATION GROUP THEORIES

One significant topic of the generating algorithm was not completely explained in the preceding section. A more rigorous mathematical model is needed to recognize whether two ordered sequences of t-fragments (see Figure 6) are equivalent or not. For an unambiguous description of this model the language of graph[22,23] and permutation group[24,25] theories must be used.

First of all, the intuitive notions of t-, l-, and b-fragments must be substituted for exactly defined G_t, G_l, and G_b *fragment graphs* (FGs). The molecular graph $G = G(V,E)$ is usually defined as consisting of the set $V = \{v_i\}$ of its vertices (corresponding to atoms) and the set $E = \{e_j\}$ of its edges (corresponding to chemical bonds). Multiple edges are also allowed in G; for this reason, molecular graphs are multigraphs in the more exact terminology. Similarly, the *fragment graph* $G_x = G_x(V \cup V' \cup V'', E \cup E' \cup E'')$, x = t, l, or b, consists of the union of the vertex sets V, V′, and V″ and of the union of the edge sets E, E′, E″; multiple edges are also allowed. The sets V and E correspond to atoms and chemical bonds, respectively. The vertices $v \in V'$, $|V'| \ge 0$ (*in-vertices*), and $v \in V''$, $|V''| = 1$ (*out-vertices*), correspond to the ends of in- and out-arrows. These vertices are necessarily differentiated from other vertices by means of specific labels (black and white squares in Figures 7–10). Finally, the sets E′ ($|E'| = |V'|$) and E″ ($|E''| = |V''|$) consist of ordinary, double, and triple edges which connect "additional" vertices of V′ or V″ with "normal" vertices of the set V. The FGs corresponding to C_6H_5 (t), 1,4-C_6H_4 (l), and 1,3,5-C_6H_3 (b) "aromatic" fragments are depicted in Figure 7; parts a–c. It is evident that the set V′ is necessarily empty in the case of G_t graphs (Figure 7a), consists of a single vertex in the case of G_l graphs (Figure 7b),

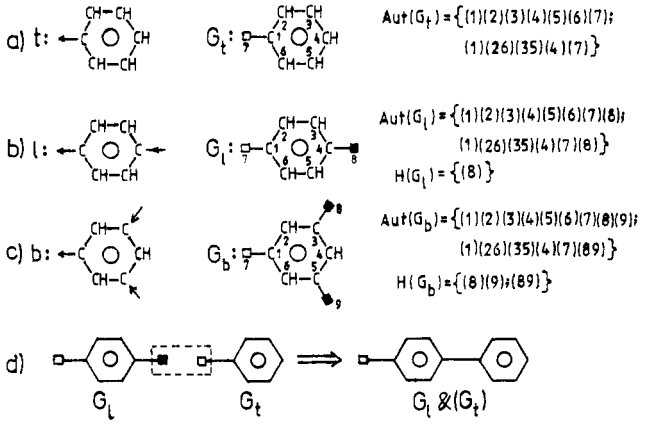

Figure 7. Examples of G_t, G_l, G_b, and composite $G_l\&(G_t)$ fragment graphs.

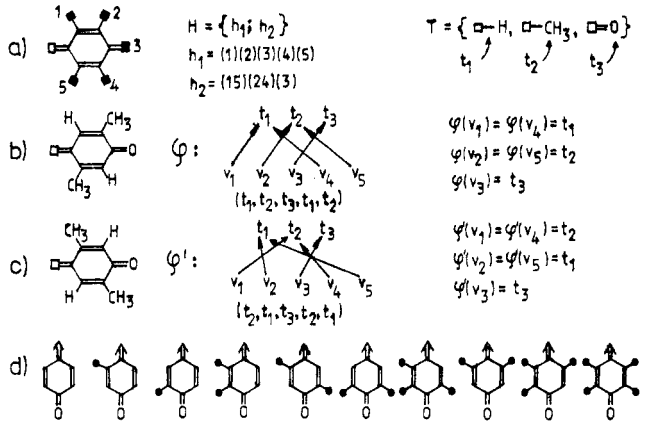

Figure 8. Two equivalent mappings (b and c) corresponding to a fixed G_b graph (a); the resulting t-fragments (d).

and consists of two or more vertices in the case of G_b graphs (Figure 7c).

For the construction of composite FGs, the pairs of in- and out-vertices must be obviously removed while the corresponding edges (of the sets E′ and E″) must be united to form new chemical bonds of appropriate multiplicity. An illustration for $G_l\&(G_t)$ composite fragment is shown in Figure 7d.

Unlike the symmetry of any fragment, the symmetry of the corresponding FG can be rigorously characterized by its *automorphism group* $\mathrm{Aut}(G_x)$, x = t, l, or b. This group consists of the permutations (of a symmetric group S_m, $m = |V| + |V'| + |V''|$) which preserve all graph adjacencies as well as all vertex labels. The automorphism groups of the example graphs G_t, G_l, and G_b are presented in Figure 7, parts a–c, respectively. We mention here that the groups $\mathrm{Aut}(G_t)$, $\mathrm{Aut}(G_l)$, and $\mathrm{Aut}(G_b)$ need not be isomorphic to the groups of "normal" chemical graphs which can be obtained from FGs by removal of in- and out-vertices together with their edges.

It is easy to see that the groups $\mathrm{Aut}(G_x)$, x = t, l, or b, always contain additional information that is useless for the construction of composite FGs; for that purpose, only the permutations of in-vertices (black squares in Figure 7) must be known. Thus, the *restricted actions* of automorphism groups onto the sets V′ must be considered; in most cases these groups are, however, action (but not permutation) groups. This means that "duplicate" permutations can be present in the restricted automorphism groups. The permutation groups obtained by deleting all duplicates will be denoted by $H = \{h_k\}$. It is evident that the group H cannot be constructed for any graph G_t (because the set V′ is empty in this case) and is an identity group for any graph G_l ($|V'| = 1$, cf. Figure 7b). For this reason, only the groups of the graphs G_b are discussed here.

The composite fragment graphs $G_b\&(G_{t_1}, G_{t_2}, ..., G_{t_\rho})$ will be now considered. First of all, it is evident that each composition of G_b (corresponding to the b-fragment of degree ρ) with ρ graphs G_{t_i}, $i = 1, 2, ..., \rho$, may be characterized by a certain *mapping* $\varphi = V' \Rightarrow T$ from the vertex set V′ ($|V'| = \rho$) of G_b into the given set $T = T^{(1)} \cup T^{(2)} \cup T^{(3)}$, $|T| = p_1 + p_2 + p_3$. The arrow schemes and functional notations for two equivalent mappings φ and φ' [corresponding to ordered sequences (←H,←CH_3,⇐O,←H,←CH_3) and (←CH_3,←H,⇐O,←CH_3,←H)] are represented in Figure 8b,c. One can easily see that the mappings φ, φ' can be transformed one into another if the permutation h_2 of Figure 8a acts on the set V′ (lower ends of the arrows starting at v_1,v_5 and v_2,v_4 must be permuted in this example).

Let $\Phi = \{\varphi: V' \Rightarrow T\}$ be a set consisting of all allowed mappings; the cardinality of this set, $|\Phi|$, evidently depends on cardinalities p_i of the sets $T^{(i)}$ (i = 1, 2, 3), and on numbers a_1, a_2, a_3 of univalent, divalent, and trivalent vertices in V′ (see eq 5a). It is well-known[25] that if the action of the permutation group (H) on the domain set (V′) of mappings is defined, then the induced group (Γ) acting on the set of mappings (Φ) can be constructed. The order of the new permutation group $\Gamma = \{\gamma_k\}$ is typically identical to that of the group H but the degree of Γ is equal to $|\Phi|$. The group Γ introduces the equivalency relation on the set Φ; in other words, it partitions Φ into several equivalency classes which are usually denoted as *orbits* or *transitive sets*. Each orbit consists of the mappings φ which can be interconverted by one or several permutations $\gamma_k \in \Gamma$; the notations $\varphi' = \gamma_k\varphi$, $\varphi'(v) = \varphi[h_k^{-1}(v)]$ mean that for each permutation γ_k (corresponding to $h_k \in H$) any mapping φ is converted to the equivalent mapping φ', and the image of any $v \in V'$ in φ' coincides with the image of $h_k^{-1}(v)$ in φ.

The above consideration shows that the construction of nonequivalent sequences (t_1, t_2, ..., t_ρ) in step 8 of the algorithm (see Section 4) actually needs the arbitrary representatives for all orbits of the group Γ on Φ to be constructed. The corresponding analytical enumeration problem is reduced to simply counting the orbits of Γ. This problem can be easily solved by means of Burnside's Lemma[22,24,25] (if the number of orbits must be calculated) or Pólya's Enumeration Theorem[26,27] (if the enumeration series must be derived). In this paper, we prefer the simple calculations based on Burnside's Lemma; this Lemma has also been applied by two of us to other enumeration problems of organic chemistry (i.e., to enumeration of chiral derivatives of organic structures,[28] symbolic equations corresponding to linear and cyclic topologies,[28–30] 2D and 3D chain configurations).[28,31,32]

Burnside's Lemma makes it possible to calculate the total number n of orbits for an arbitrary permutation group $\prod = \{\pi_i\}$ acting on the set $X = \{x_j\}$. For this purpose, the numbers $\chi(\pi_i)$ of elements $x_j \in X$ which are converted into themselves must be known for every permutation π_i (see eq 5b).

$$|\Phi| = \prod_{i=1}^{i=3} p_i^{a_i} \tag{5a}$$

$$n = 1/|\prod| \sum_{i=1}^{i=|\prod|} \chi(\pi_i) \tag{5b}$$

$$n = 1/|H| \sum_{k=1}^{k=|H|} \prod_{i=1}^{i=3} p_i^{c_i(h_k)} \tag{5c}$$

For the discussed problem, the group $\prod$ is an induced group Γ, and X is a set Φ of allowed mappings. In this case the values $\chi(\gamma_k)$ can be calculated by means of simple combinatorial considerations. One can easily see that any mapping φ is converted into itself by the kth permutation γ_k of Γ if and only if the images of all vertices $v \in V'$, *belonging to the same cycle of corresponding permutation* h_k, are identical. This means that the cyclic structures of permutations $h_k \in H$ (but not the permutations $\gamma_k \in \Gamma$) must be known. Let $c_1(h_k)$, $c_2(h_k)$, and $c_3(h_k)$ be the total numbers of those cycles in h_k which consist of univalent, divalent, and trivalent vertices of V′, respectively. In this case, the images of the vertices belonging

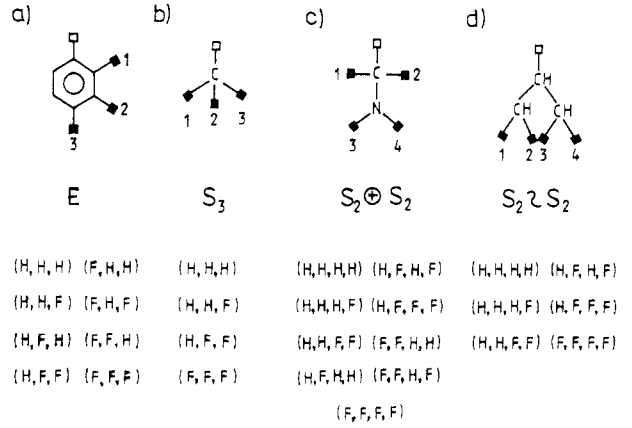

Figure 9. Canonical representatives of the orbits corresponding to several b&(t_1, t_2, ..., t_ρ) composite fragments (T = {←H, ←F}, ←H < ←F).

to $c_l(h_k)$ cycles (l = 1, 2, or 3) may be obviously chosen in $|T^{(l)}|^{c_l(h_k)} = p_l^{c_l(h_k)}$ ways. Thus, $\chi(\gamma_k) = \prod_{l=1}^{l=3} p_l^{c_l(h_k)}$ and the total number of orbits n can be calculated by eq 5c.

The application of eq 5c is exemplified for the G_b fragment graph of Figure 8a; for simplicity, the set T consists of three terminal FGs corresponding to hydrogen, methyl, and oxo groups (that is, $p_1 = 2$, $p_2 = 1$, and $p_3 = 0$). The restricted automorphism group H of G_b consists of two permutations. In an identity permutation h_1, four cycles correspond to univalent in-vertices of V′ and one cycle corresponds to divalent in-vertex of V′; that is, $c_1(h_1) = 4$, $c_2(h_1) = 1$, and $c_3(h_1) = 0$. Similarly, for the permutation h_2 = (15)(24)(3) we obtain $c_1(h_2) = 2$, $c_2(h_2) = 1$, and $c_3(h_2) = 0$. The application of eq 5a shows that the number of orbits is equal to $^1/_2(2^4 \cdot 1^1 \cdot 0^0 + 2^2 \cdot 1^1 \cdot 0^0) = 10$ (0^0 is equated to 1 by definition). The corresponding t-fragments are shown in Figure 8d with methyl groups being indicated by filled circles.

Let us turn to the generating problem of composite b&(t_1, t_2, ..., t_ρ) fragments. The solution of this problem needs a rule which makes it possible to choose the unique representative for every orbit of the group Γ acting on Φ. Two mappings, φ and φ', can be easily compared by means of their linear codes (identical to the ordered sequences of t-fragments) if the linear order on the set T has been initially introduced. This means that for each pair t_1, $t_2 \in$ T one of the following inequalities holds: $t_1 < t_2$ or $t_1 > t_2$. The last requirement is automatically satisfied because all fragments are represented in computer memory by their WLN codes (see Section 7), which can always be compared lexicographically. Thus, in every orbit of Γ the unique mapping φ *with the smallest code* can always be detected, and this mapping is thought to be the *canonical representative* of the whole orbit. For example, two mappings, φ and φ', of Figure 8, parts b and c form an orbit; if $t_1 < t_2 < t_3$, then φ is the canonical representative [because the code (t_1, t_2, t_3, t_1, t_2) is lexicographically smaller than the code (t_2, t_1, t_3, t_2, t_1)].

The above consideration shows that the codes of all generated mappings [i.e., the ordered sequences (t_1, t_2, ..., t_ρ)] must be compared with their images that are obtained by action of permutations h_k of the group H corresponding to fixed b-fragment. The mappings with the codes being converted into smaller codes are equivalent to the mappings generated before and, hence, must be deleted. This general method can be applied in all cases but there exist three situations which do not need the groups H to be explicitly known. These situations are briefly described.

(1) If the group H of the b-fragment is an identity group (E), then each mapping φ = V′ ⟹ T forms an orbit; all ordered sequences are nonequivalent in this case (cf. Figure 9a).

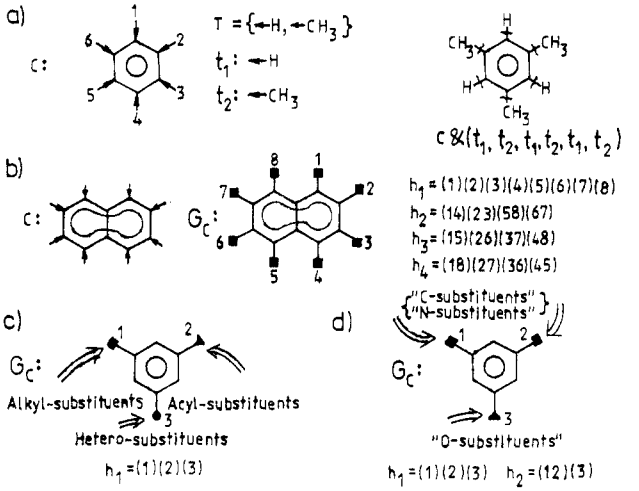

Figure 10. Example of c&(t_1, t_2, ..., t_ρ) combining operation (a); graphs G_c and their restricted automorphism groups (b–d).

(2) If the group H is a symmetric group S_n (n is equal to the degree ρ of b-fragment), then the canonical representatives of all orbits are easily recognized. All components of the ordered sequences must be arranged in nondescending order in this case (see Figure 9b for an example).

(3) If the group H is a direct sum of two or more symmetric groups S_{m_i}, each of them acting on its own subset V′$_i$ ⊂ V′ consisting of m_i vertices (∪ V′$_i$ = V′; V′$_i$ ∩ V′$_j$ = ∅ if $i \neq$ j), then the components corresponding to each subset must be arranged in nondescending order. An example is shown in Figure 9c. In this example, H is the direct sum $S_2 \oplus S_2$; the first S_2 group acts on the subset {v_1,v_2} and the second one on the subset {v_3,v_4}. For this reason, in the complete set of canonical representatives (see Figure 9c), the first component of all ordered sequences is less than or equal to the second component, and the third component is less than or equal to the fourth component.

In practice, only a small part of elementary b-fragments need the restricted automorphism groups to be stored in computer memory. The cyclic b-fragments (similar to that of Figure 8a) and the complex acyclic b-fragments (Figure 9d) with the groups H isomorphic to wreath[26] or generalized wreath[33] products can serve as examples. In principle, the special highly effective procedures (like that of ref 18) are preferable in these cases. In the present-state program implementation of the generating algorithm, the direct comparing of ordered sequences with their images is used. This simplification does not lead, however, to less effective procedures because the orders of the groups H are typically small. For example, the groups H corresponding to b-fragments of Figures 8a and 9d consist of two and eight permutations, respectively.

6. GENERATING PROBLEM FOR TARGET STRUCTURES

Let the central fragment c with ρ in-arrows be given, and the set of substituents T = T$^{(1)}$ ∪ T$^{(2)}$ ∪ T$^{(3)}$ be constructed as described above. Then the combining operation c&(t_1, t_2, ..., t_ρ) means that each of the ρ numbered in-arrows of the c-fragment together with an out-arrow of corresponding t-fragment of the set T produces a new chemical bond; ρ pairs of in- and out-arrows must consist of arrows of the same multiplicity. This operation is evidently very similar to the b&(t_1, t_2, ..., t_ρ) combining operation but leads to final target structures. A simple example is shown in Figure 10a: the mesitylene structure is originated by "benzene-skeleton" c-fragment of degree 6 and the ordered sequence (←H, ←CH_3,

←H, ←CH_3, ←H, ←CH_3) which corresponds to the set T = $T^{(1)}$, consisting of two elementary t-fragments.

The generation of all possible c&(t_1, t_2, ..., t_ρ) combinations evidently needs nonequivalent (with respect to the symmetry group of c-fragment) ordered sequences (t_1, t_2, ..., t_ρ) to be constructed. For that purpose, the fragment graphs G_c = (V ∪ V′, E ∪ E′) are introduced. These graphs are very similar to the FGs of b-fragments but contain no out-vertices (the sets V″ and E″ are empty); an example of the "naphthalene-skeleton" graph is shown in Figure 10b. The symmetries of c-fragment graphs can be characterized by their automorphism groups Aut(G_c) and by restricted automorphism groups H = {h_k}. For the example graph of Figure 10b, the group H = {h_1, h_2, h_3, h_4} of degree 8 is isomorphic to the group Aut(G_c) of degree 18 (|V| + |V′| = 10 + 8 = 18).

The mathematical statement of the enumeration problem for c&(t_1, t_2, ..., t_ρ) combinations is identical to that described in the preceding section; this problem may be considered as typical "labeling of objects having symmetry".[17] Thus, the set Φ consists of all possible mappings φ from the set V′ of in-vertices into the set T of substituents; the cardinality of Φ can be also calculated by eq 5a. The induced group Γ with permutations γ_k corresponding to h_k evidently acts on the set Φ, and the orbits of this action uniquely correspond to nonequivalent ordered sequences (t_1, t_2, ..., t_ρ). For this reason, Burnside's Lemma and eq 5c can be successfully used to evaluate the number of nonequivalent c&(t_1, t_2, ..., t_ρ) combinations. For example, the number of naphthalene derivatives (Figure 10b) with m substituents ($p_1 = m$, $p_2 = 0$, $p_3 = 0$) is equal to $^1/_4(m^8 + 3m^4)$; this number is equal to 76 if $m = 2$ (cf. ref 19) and to 16576 if $m = 4$. The latter result demonstrates that additional selection criteria are required if "polyvalent" c-fragments or large sets of substituents are involved in the constructive enumeration problem.

The generation procedure for c&(t_1, t_2, ..., t_ρ) combinations is very similar to that used for b&(t_1, t_2, ..., t_ρ) composite fragments; three special situations (cf. Section 5) can also take place in this case. Thus, the comparing of ordered sequences (t_1, t_2, ..., t_ρ) with their images (obtained by action of permutations $h_k \in$ H) is not really performed if the group H of the graph G_c is an identity group E, or symmetric group S_n (n is equal to the degree ρ of c-fragment), or direct sum of several symmetric groups S_{m_i} ($\sum m_i = n$).

The problem under discussion displays, however, a specific feature which needs to be additionally considered. In Section 2, we have mentioned that separate sets of substituents can sometimes be associated with equivalent "positions" of a given c-fragment. In such cases, the in-vertices of G_c, to which individual sets of substituents correspond, must be differentiated by specific labels (squares, triangles, circles; see Figure 10, parts c and d). This artificial labeling necessarily modifies the restricted automorphism group and, hence, makes it possible to accurately solve the generation problem. The groups H of example graphs of Figure 10c,d are identity and $S_2 \oplus S_1$ groups, respectively; these permutation groups (unlike the S_3 group of the original c-fragment of degree 3) really ensure the correct construction of benzene derivatives with substituents belonging to 3 (see Figure 10c) or 2 (see Figure 10d) separate sets. We note here that the individual sets of substituents are generated separately in the computerized version of the suggested algorithm; the individual sets of elementary fragments and proper selection criteria may be used for each set of substituents.

7. COMPUTER IMPLEMENTATION OF THE ALGORITHM

In Sections 4 and 5 we have demonstrated that the suggested algorithm makes it possible to produce only nonidentical

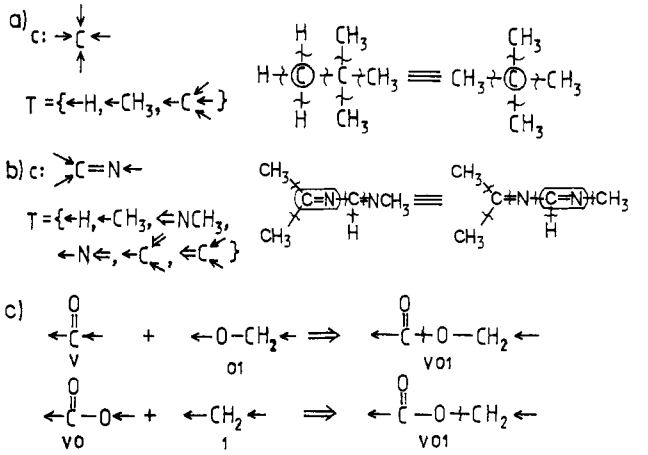

Figure 11. Fragment intersections resulting in duplicate structures (a and b); the recognition of identical composite l-fragments by means of their WLN codes (c).

substituents *if all EFs consist of single atoms*. This is not the case, however, if multiatomic fragments are allowed. In Section 2, the combinations of two "small" EFs leading to a "larger" EF (←CO← + ←OH = ←COOH) or to identical results (←CH_2CH_2← + ←O← = ←CH_2← + ←CH_2O←) were mentioned. Generally speaking, this situation can take place if two or more EFs intersect; in other words, if their structures contain just the same substructure (CO, OH, and CH_2 groups, in the above examples). Now we shall demonstrate that the intersection problem must be also taken into account in the generating procedures for target structures *if the given central fragment and some of elementary (or composite) fragments contain identical substructures*. Typical examples are depicted in Figure 11a,b. It is easy to see that duplicate structures arise if the central fragment is essentially identical to one of fragments (to b-fragment in Figure 11a) or is a part of a composite fragment (constructed from ←C= and ⇐NCH_3 in Figure 11b).

The examples considered show that the computerized version of the algorithm needs to contain a special tool making it possible to recognize the identical results (composite l-fragments, t-fragments, or target structures) having been produced due to fragment intersections. In terms of graph theory, this means that *isomorphic graphs* corresponding to composite l-fragments, t-fragments, or final structures must be recognized as early as possible in the generating procedure. The graph isomorphism problem was and still is very popular among professional mathematicians (we refer to an old review on "isomorphism disease"[34] and to one of prominent current results[35]), but no substantially "good", polynomial algorithms are known till now *for arbitrary graphs*. Although polynomial algorithms were elaborated for more restricted classes of graphs (see ref 35 for details), chemists prefer the *heuristical canonization* procedures to be used for molecular graphs; the canonical representations can be directly compared, thus making it possible to immediately recognize whether the graphs are isomorphic or not.

The Wiswesser Line Notation[36] (WLN) is thought to produce canonical codes of molecular graphs and, hence, is commonly used for managing chemical structure information. The WLN codes are allowed to be concatenated and inserted; this fact makes it possible to successfully use them in appropriate generating procedures for substituents and final structures. On the other hand, the canonicity of WLN codes allows their use in recognizing those fragments (or target structures) which have been produced due to fragment intersections. In a simple example of Figure 11c, the fast lexicographic comparison of two resulting codes immediately shows that the differently constructed composite l-fragments are actually identical.

For the above reasons, in the computerized version of the generating algorithm most of operations are performed with WLN codes. In particular, the input data (namely, the list of given EFs) as well as the output data (the list of generated substituents and final structures) are presented in the form of WLN codes. The codes of the resulting target structures are finally converted into standard adjacency matrices of corresponding molecular graphs.

The other important feature of the computerized procedure consists in the presence of several *selection criteria* which make it possible to avoid combinatorial explosion due to an enormously large number of generated substituents. The first type of these criteria is associated with the restrictions which indicate the pairs of EFs being forbidden to be combined. In other words, these restrictions do not allow construction of composite fragments containing undesirable (e.g., unstable) chemical bonds. The disallowed formation of ←O-OH and ←P=C⁼ composite fragments (←O←, ←OH, and ←P⇐, ⇐C⁼ fragments being combined) can serve as simple examples. In principle, similar limitations can be used in the generating procedure for the final structures.

The second type of selection criteria is associated with the maximal allowed numbers of concrete EFs. Thus, for each (ith) l-fragment the positive integer $\alpha(l_i)$ must be given, and its value shows that the number of elementary fragments l_i in any of constructed composite l-fragments will not exceed $\alpha(l_i)$. Similarly, the positive value of $\beta(b_j)$ for jth b-fragment indicates that only the substituents which contain no more than $\beta(b_j)$ b-fragments (with the sequential number equal to j) will be generated. The analogous restrictions for t-fragments can be also incorporated into the program if needed. We note that the values of $\alpha(l_i)$, $\beta(b_j)$, etc. as well as maximal disperse and rank values (M and N, see above) are really introduced in a dialogue mode. The special dialogue menu of the computer program makes it also possible to select the desired c-, t-, l-, and b-fragments from the standard set, or to introduce new elementary fragments if it is needed.

The computer implementation of the suggested generating algorithm has been written by two of us (O.A.L. and D.V.S.) in PASCAL. Numerous runs on an IBM PC AT computer have shown that the resulting program ensures the relatively fast (about 50 structures per second) generation and storage of 10^4–10^6 final structures on a hard disk.

8. CONCLUSIONS

The computerized version of the algorithm is thought to be only a part of the more extensive program complex in which the correlation searching routines will be involved. The above-mentioned and similar selection criteria are directed to regulate and control the generation process in order to obtain the best possible correlations for the derivatives of a parent compound. In our opinion, this approach will make it also possible to detect new favored and disfavored combinations of fragments and thus to refine the initial models of activity. The treatment of structure–activity relationships and the results obtained are planned to be outlined in further publications.

ACKNOWLEDGMENT

We are grateful to the referees for their attention to this manuscript and especially to Professor A. T. Balaban for his kind suggestions to improve the English.

REFERENCES AND NOTES

(1) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Function*; Wiley-Interscience: New York, 1979.
(2) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure–Activity Analysis*; Chemometrics Series No. 9; Research Studies Press: Letchworth, 1986.
(3) Heap, B. R. The Production of Graphs by Computer. In *Graph Theory and Computing*; Academic Press: New York, 1972; pp 47–62.
(4) Faradgev, I. A. Generation of Nonisomorphic Graphs with Given Partition of Vertex Degrees. In *Algorithmic Investigations in Combinatorics*; Faradgev, I. A., Ed.; Nauka: Moscow, 1978; pp 11–19 (in Russian).
(5) Kvasnička, V.; Pospichal, J. Canonical Indexing and Constructive Enumeration of Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 99–105.
(6) Lederberg, J.; Sutherland, G. L.; Buchanan, B. G.; Feigenbaum, E. A.; Robertson, A. V.; Duffield, A. M.; Djerassi, C. Applications of Artificial Intelligence for Chemical Inference. I. The number of Possible Organic Compounds. Acyclic Structures Containing C, H, O, and N. *J. Am. Chem. Soc.* **1969**, *91*, 2973–2976.
(7) Trinajstič, N.; Jeričevič, Z.; Knop, J. V.; Müller, W. R.; Szymanski, K. Computer Generation of Isomeric Structures. *Pure Appl. Chem.* **1983**, *55*, 379–390.
(8) Bangov, I. P. Computer-Assisted Generation of Molecular Structures from a Gross Formula. 1. Acyclic Saturated Compounds. *Commun. Math. Chem.* **1983**, *14*, 235–246.
(9) Masinter, L. M.; Sridharan, N. S.; Lederberg, J.; Smith, D. H. Applications of Artificial Intelligence for Chemical Inference. 12. Exhaustive Generation of Cyclic and Acyclic Isomers. *J. Am. Chem. Soc.* **1974**, *96*, 7702–7714.
(10) Elyashberg, M. E.; Gribov, L. A.; Serov, V. V. *Molecular Spectral Analysis and Computers*; Nauka: Moscow, 1980; pp 141–168 (in Russian).
(11) Mitrofanov, Ju. P.; Raznikov, V. V.; Shkurov, V. A. A Method of Automatic Designing of Topological Constitutional Formulas of Organic Substances from a Given Chemical Formula. *Zh. Anal. Khim.* **1982**, *37*, 1477–1483 (in Russian).
(12) Molodtsov, S. G.; Piottukh-Peletskiy, V. N. Generation of All Nonisomorphic Chemical Graphs from a Given Set of Structural Fragments. *Vychisl. Sist.* **1984**, *103*, 51–58 (in Russian).
(13) Funatsu, K.; Miyabayashi, N.; Sasaki, S. Further Development of Structure Generation in the Automated Structure Elucidation System CHEMICS. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 18–28.
(14) Christie, B. D.; Munk, M. E. Structure Generation by Reduction: A New Strategy for Computer-Assisted Structure Elucidation. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 87–93.
(15) Bangov, I. P. Computer-Assisted Structure Generation from a Gross Formula. 3. Alleviation of the Combinatorial Problem. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 277–289.
(16) Kerber, A.; Laue, R.; Moser, D. Ein redundanzfreier Strukturgenerator für molekulare Graphen. *Anal. Chim. Acta* **1990**, *235*, 221–228.
(17) Masinter, L. M.; Sridharan, N. S.; Carhart, R. E.; Smith, D. H. Applications of Artificial Intelligence for Chemical Inference. 13. Labeling of Objects Having Symmetry. *J. Am. Chem. Soc.* **1974**, *96*, 7714–7723.
(18) Tratch, S. S.; Podymova, E. V.; Zefirov, N. S. The Generating Algorithm of Nonequivalent Labelings on the Sets with Given Permutation Groups. In *Proceedings of the 2nd USSR Conference on Methods and Programs for Solution of Optimization Problems for Graphs and Networks*; Novosibirsk, 1982; Part I, 210–213 (in Russian).
(19) Smith, D. H. Applications of Artificial Intelligence for Chemical Inference. 15. Constructive Graph Labeling Applied to Chemical Problems. Chlorinated Hydrocarbons. *Anal. Chem.* **1975**, *47*, 1176–1179.
(20) Zefirov, N. S.; Tratch, S. S.; Chizhov, O. S. *Cage and Polycyclic Compounds. Molecular Design on the Base of Isomorphic Substitution Principle*; Itogi Nauki i Tekhniki, Ser. Org. Khim., 3; VINITI Publ.: Moscow, 1979 (in Russian).
(21) Kornilov, M. Yu.; Dyadyusha, G. G.; Zamkovoy, V. I.; Dekhtyar, M. L.; Kachkovskiy, A. D. The Computer-Assisted Search for Heterocycles for Cyanine Dyes. *Khim. Geterotsikl. Soedin.* **1984**, 217–222 (in Russian).
(22) Harary, F. *Graph Theory*; Addison-Wesley: Reading, MA, 1969.
(23) Swamy, M. N. S.; Thulasiraman, K. *Graphs, Networks, and Algorithms*; Wiley: New York, 1981.
(24) Wielandt, H. *Finite Permutation Groups*; Academic Press: New York, 1964.
(25) Klin, M. Ch.; Pöschel, R.; Rosenbaum, K. *Angewandte Algebra für Mathematiker und Informatiker. Einführung in gruppentheoretische-kombinatorische Methoden*; VEB Deutscher Verlag der Wissenschaften: Berlin, 1988.
(26) Pólya, G. Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen. *Acta Math.* **1937**, *68*, 145–254.
(27) Harary, F.; Palmer, E. M.; Robinson, R. W.; Read, R. C. Pólya's Contributions to Chemical Enumeration. In *Chemical Applications of Graph Theory*; Balaban, A. T., Ed.; Academic Press: London, 1976; pp 11–24.
(28) Tratch, S. S.; Zefirov, N. S. The Enumeration and Classification of Orbits for S_2^G Power Group: Application of Burnside's Lemma and Expanded Cycle Indices of Permutation Groups. In *Proceedings of the USSR Conference on Molecular Graphs in Chemical Investigations*; Kalinin, 1990; pp 104–105 (see also pp 106–107; in Russian).
(29) Tratch, S. S.; Gamziani, G. A.; Zefirov, N. S. Problems of Molecular Design and Computers. 10. Enumeration and Generation of Equations Characterizing Ionic, Radicalic, and Redox Processes with Linear Electron Transfer. *Zh. Org. Khim.* **1987**, *23*, 2488–2507 (in Russian).

(30) Zefirov, N. S.; Tratch, S. S. Symbolic Equations and Their Applications to Reaction Design. *Anal. Chim. Acta* **1990**, *235*, 115-134.
(31) Tratch, S. S.; Devdariani, R. O.; Zefirov, N. S. Combinatorial Models and Algorithms in Chemistry. Configuration-topological Analogs of Wiener Index. *Zh. Org. Khim.* **1990**, *26*, 921-932 (in Russian).
(32) Zefirov, N. S.; Kozhushkov, S. I.; Kuznetsova, T. S.; Kokoreva, O. V.; Lukin, K. A.; Ugrak, B. I.; Tratch, S. S. Triangulanes: Stereoisomerism and General Method of Synthesis. *J. Am. Chem. Soc.* **1990**, *112*, 7702-7707.
(33) Zefirov, N. S.; Kaluzhnin, L. A.; Tratch, S. S. The Generalization of the Wreath Product of Permutation Groups and Its Application to Description of Chemical Structure Formulas. In *Investigation in Algebraic Theory of Combinatorial Objects*; Klin, M. H., Faradgev, I. A., Eds.; VNIISI Publ.: Moscow, 1985; pp 175-186 (in Russian).
(34) Read, R. C.; Corneil, D. G. The Graph Isomorphism Disease. *J. Graph Theory* **1977**, 339-363.
(35) Goldberg, M. K. A Nonfactorial Algorithm for Testing Isomorphism of Two Graphs. *Discrete Appl. Math.* **1983**, *6*, 229-236.
(36) Wiswesser, W. J. *A Line-Formula Chemical Notation*; Crowell: New York, 1954.

ALF-A: A Knowledge Acquisition Tool for Troubleshooting of Laboratory Equipment

HENRIK ERIKSSON* and PER LARSES†
Department of Computer and Information Science, Linköping University, S-581 83 Linköping, Sweden

Received September 11, 1991

Expert systems can be used to support troubleshooting of technical equipment, such as laboratory instruments. A critical factor and bottleneck in the development of such knowledge-based systems is the acquisition of relevant knowledge for the task. Computer-based knowledge acquisition (KA) tools can support knowledge elicitation from domain experts by transforming structures elicited from or entered by nonprogrammers into knowledge bases. A KA tool (ALF-A) that supports the development of expert systems for fault diagnosis in DNA sequencing machines has been developed. Different categories of experts can use the tool directly (i.e., without an intermediate knowledge engineer) to enter their knowledge according to a predefined model of troubleshooting.

1. INTRODUCTION

Ensuring a high degree of availability for laboratory instruments (e.g., DNA sequencers) is important for most laboratories. Expert system technology offers an approach to consultation software that can facilitate troubleshooting. Such computer programs can lower the amount of time spent on fault finding as well as fault correction and, by enabling nonspecialists to do much of the work, reduce the burden on highly-qualified equipment specialists.

Expert systems are typically organized around *knowledge bases*, which contain knowledge structures for the problem area in question in a declarative format, and *inference engines* (problem solvers) that draw conclusions from the current input guided by the knowledge base. The performance of these systems is critically dependent on the contents of their knowledge base.

The process of modeling expertise, i.e., *knowledge acquisition* (KA) from domain experts, is a problem and bottleneck in the development of knowledge-based systems. Manual development of a reasonably complete knowledge base for troubleshooting a set of equipment can be a huge task, both in terms of the knowledge engineering effort and the expert time. Computer-based KA tools can be used to support knowledge acquisition from experts.

A KA tool (ALF-A) that facilitates the development of knowledge bases for fault detection in a DNA sequencer is described in this article. One of the salient aspects of this problem area is that the set of equipment required involves both chemistry (e.g., gel), hardware (e.g., laser units), and software (e.g., control programs). Consequently, experts in this area are primarily specialized in one of these aspects and, thus, are proficient only in one group of faults. Even though there exist dedicated shells for troubleshooting [e.g., TEST (*15*) and TestBench (TestBench is a trademark of Carnegie Group, Inc., and Texas Instruments, Inc.) (*3*)] and KA tools for troubleshooting [e.g., TDE (*13*, *14*)], it is difficult to find one that simultaneously meets requirements from all aspects of this type of equipment.

Due to the nature of the equipment, a model that reflects the troubleshooting task (shallow model) was used rather than a model of the chemical/physical processes and the equipment (qualitative model). The conceptual domain model is based on a *symptom-fault tree*, which controls the reasoning strategy. This model is accepted to a large extent by the experts involved in this project. The KA tool developed implements this conceptual domain model and allows experts to enter new symptoms and faults as well as define the relationships among them. Knowledge entered can be stored persistently on file or transformed into target knowledge bases.

This article is organized as follows: Section 2 presents the background and motivation of this project. Approaches to tool support for knowledge acquisition are discussed in Section 3. An overview of the KA tool is provided in Section 4. Section 5 treats the specification and generation of the ALF-A system. Related work is discussed in Section 6. Finally, a summary and conclusions are provided in Section 7.

2. BACKGROUND

One set of equipment that can be used to sequence DNA is the "Automated Laser Fluorescent (A.L.F.) DNA Sequencer" (Automated Laser Fluorescent (A.L.F.) DNA sequencer is a trademark of Pharmacia Biosystems AB) (*9*, *20*). The A.L.F. DNA Sequencer automates detection of fluorescently labeled DNA molecules. The system comprises (a) an electrophoresis and laser fluorescent subsystem, (b) the control and analysis software for PC compatibles, and (c) a sequencing reagent kit containing the chemicals required. Labeled DNA migrates through a gel and intercepts the laser beam that excites the fluorescent labels. The gel consists of 40 electrophoresis lanes. Photodiodes for each of the lanes detect emitted light. The system's capacity is two gels per day, which is equivalent to 8 kb at a rate of 1000 bases/h. After a run is completed, the computer calculates the DNA sequence and assigns ambiguity codes to hard-to-call bases.

* Address correspondence to this author at his present address: Stanford University Medical Center, MSOB X215, Stanford, CA 94305-5479.
† Present address: Enator Kunskapssystem AB, St. Larsgatan 12, S-582 24 Linköping, Sweden.

0095-2338/92/1632-0139$03.00/0