course, some data base files have an annual lease cost, and the printed sources are revised from time to time.)

The cost per match for some sources may actually be greater than the cost of contacting a manufacturer directly. However, it allows NIOSH to obtain ingredient data much sooner.

Costs of trade name matching using specialized sources could possibly be reduced if products could be categorized according to use. For example, if dyes, pigments, inks, paints, etc. could be classified as coloring agents, only these products could be searched in the Colour Index, eliminating the need to search for all 40 000–100 000 trade names in this source. Also, many additional specialized data sources could be used in generic resolution if the number of trade names to be matched could be limited in this way.

The cost per match for the aggregate data sources may actually be somewhat higher than estimated, due to overlapping coverage among sources. On the other hand, overlap could be beneficial, in that it allows reported product ingredients to be compared and either verified or called into question, in cases of discrepancies.

None of the sources evaluated were found to be totally reliable when compared with NIOSH ingredient data. Practically none provided ingredient percentage information. Depending on the level of accuracy required by NIOSH, manufacturers may eventually have to be contacted directly. However, ingredient data from these secondary sources could still prove useful for interim chemical exposure estimates.

Since most data sources, including the generalized sources, have limited coverage, a combination of sources would have to be used to maximize the number of matches with NIOSH trade names.

All data sources are less current than manufacturer data. (Numerous manufacturers have stated that their product formulations change frequently.)

Use of additional data sources would allow NIOSH to expand its data base to include trade name products not identified during the industrial survey.

The feasibility assessment methodology used in this study might well be of interest to individual firms, public health personnel, and/or labor unions concerned with exposure monitoring of workers. In addition to identifying the ingredients of specific trade name products, the methodology described herein can be adapted to identification of the probable ingredients of certain types of products within various generic use categories such as adhesive and sealant compounds, paints and coatings, and detergents.

User-Oriented Approach to a Computerized Organic Reaction Catalog

BERI J. COHEN

Technicon Instruments Corporation, Tarrytown, New York 10591

Received December 3, 1981

A computerized system for creating and searching an organic reaction catalog is described. Starting materials and products are represented as sets of parameters defining their structural features, written in a notation form. The system deals efficiently with stereochemistry and has extensive capabilities for substructure searching, including searching for closely related functional groups. Its application to small computers is discussed.

INTRODUCTION

The systematic documentation of organic reactions for a computerized retrieval of synthetic information presents a major challenge for organic chemists and information scientists alike.[1-5] The objective is to be able to represent efficiently chemical structures in a retrievable form, so that substructures will be retrieved as well, upon request, in a meaningful way. Two general approaches were described in the literature. One uses key words of partial structural features of the reactants. It is exemplified by the commercially available CRDS system from Derwent Publications[6] and the GREMAS system from IDC.[7] In the first, key words are either trivial names used in organic synthesis jargon or codes of structural fragments according to the "Ringcode".[8] Their limitations as to substructure search in case of complex structures were already discussed.[4,6] The GREMAS system also uses special coding and has considerable versatility in substructure searching. Unfortunately, its high cost makes it unaffordable to most potential users. The other approach is to represent an atom-by-atom structure of the reactant molecules by means of a topological description such as graphs or connection tables. It requires more sophisticated programming and high-performance computers. In recent years many improvements were introduced such as automatic detection of reaction sites,[9,10] use of screens for a preliminary search,[11] and automatic conversion of WLN formulae to connection tables.[12] An in-house system was recently reported from a major company which uses the key word approach,[13] showing that there is no unique, satisfying answer to the problem.

Today, most organic chemists are expected to have access to a computer service. The present work was aimed, therefore, at investigating the possibility of using the key word approach for creating and searching an organic reaction data base in the particular areas of interest of one or several research groups, updating and searching being done by the users themselves. In designing the system, several objectives were set: (a) representation of molecular structures that eliminates the use of dictionaries with a minimum number of rules; (b) easy coding of reaction data and interpretation of computer output after a search has been performed; (c) flexibility and versatility in searches of substructures. We believe that these goals are successfully met by the system described below.

STRUCTURE OF THE CATALOG

Each reaction in the catalog consists of a set of parameters describing the relevant structural features of the starting material, a similar set for the product, a parameter describing the general type of the reaction, and a short text providing additional essential data such as reaction conditions and references. The text can be reduced to a single number referencing a physical storage system such as a card file or microfilm. Apart from this text, all the other parameters are

0095-2338/82/1622-0195$01.25/0

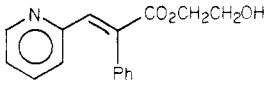
searchable by the handling program.

One starting material and one product are coded per reaction. If two or more starting materials contribute to the structure of the product or the reaction leads to more than one major product, that reaction will be coded more than once, to account for various reactant combinations.

A. Representation of Molecular Structures. Structural Parameters. A line notation system for an efficient representation of structures of reactant molecules was considered. A notation for reactants participating in organic reactions must not necessarily follow the guidelines for a notation for compounds. One reason is that the number of useful synthetic reactions is probably in the range of tens of thousands whereas millions of compounds have been so far registered. Compactness of the representation can therefore be sacrificed in favor of easy interpretation and versatility in substructure search. A second reason is that only that part of the reactant molecule relevant to the particular reaction in question has to be represented, resulting in the most general applicability of such a reaction to organic synthesis.

The relevant part of a reactant molecule participating in the reaction is described by structural parameters of three priority levels, position parameters, describing the relative locations of these structures, and a stereochemistry parameter, providing details of stereochemical aspects such as absolute configuration, cis/trans isomers, etc.

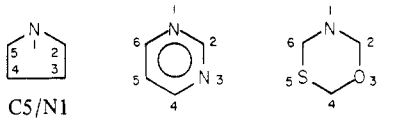
Primary Structural Parameters. These consist of the following. The main structural parameter (MSP), representing the molecular skeleton, may be a saturated carbon chain or a ring system. The MSP is chosen so that a substructure of which will no longer represent meaningfully the reactant structure, relevant to the reaction under consideration. Other primary structural parameters are substituents such as alkyl, aryl, or functional groups located directly on the molecular skeleton. Associated with the MSP is also a numbering system for locating the other substituents. *Secondary structural parameters* are substituents on alkyl, cycloalkyl, or aryl primary substituents. *Tertiary structural parameters* are substituents on secondary alkyl, cycloalkyl, or aryl side chains and substituents on carbon-chain-containing heteroatom functional groups. During output, the MSP is enclosed by quotation marks. For example:



pyridine = MSP
2-alkyl = primary substituent
double bond
phenyl } secondary substituent
ester
hydroxyl = tertiary substituent

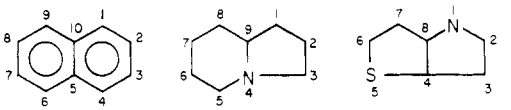
Notation for Structural Parameters. The MSP may be a saturated carbon chain and is then denoted by S. Numbering starts from the first encountered substituent. If there are two or more substituents, numbering will be in the direction that will provide lower position numbers. In case of symmetrical arrangements, numbering will start from the side where the first substituent name is lower in alphabetical order. When the MSP is a cycloalkyl ring, the notation is C followed by the number of carbons in the ring, e.g., C6 for cyclohexane. Numbering is clockwise (for stereochemistry considerations, more on this later), starting from the first substituent. When there are two or more substituents, the lower in alphabetical order of their names will have position 1, the rest should have the lowest possible numbered positions. An aromatic benzene ring is denoted by A6. Numbering is as for cycloalkyl rings. As a matter of convenience, reactions that apply to various sizes of single cyclic rings are recorded once only for the six-membered ring C6. Similarly, reactions that are typical to aromatic systems are recorded once only for benzene, A6.

Heterocyclic Ring Systems. Their notations are derived from the corresponding carbon ring, followed by a slash and the heteroatom symbol with its position such that the position numbers will be the lowest possible. If different heteroatoms are present in the ring, they will be numbered so that the lower in alphabetical order will have position 1, e.g.:



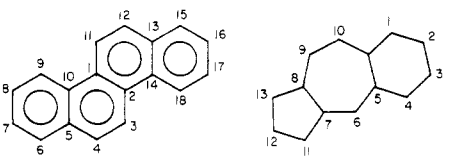
C5/N1 A6/N1N3 C6/N1O3S5

Fused Two-Ring Systems. Their notation is composed from the appropriate notations of the individual carbon rings. The larger one is drawn on the left, or if they are of the same size, the less heavily substituted ring is drawn on the left. A benzene ring precedes a fused cyclohexane. Numbering starts from the ring on the right at the upper position next to the fusion point and continues clockwise. Heteroatoms are denoted as explained for single rings. If more than one possibility exists, lower position numbers are first sought and then the alphabetical order of heteroatoms, e.g.:



A6A6 C6C5/N4 C5C5/N1S5

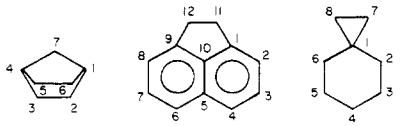
Linearly Fused Rings. By this is meant that each two rings have only one common bond. The two largest fused rings are chosen first and denoted as explained above. Each new ring is denoted by C or A followed by the number of atoms in it and the link positions in parentheses. With each new ring added, numbering of the newly introduced atoms succeeds the last numbered atom in the system, starting from the lower-numbered link. When all rings are of the same size, new rings are added on the right-hand side. Heteroatoms are denoted as explained above. In case of confusion, their lowest numbered positions are sought, e.g.:



A6A6A6(1-2)A6(13-14) C7C6C5(7-8)

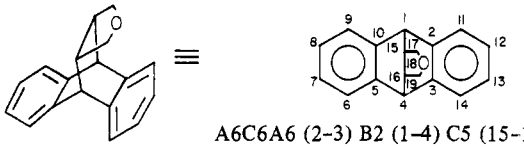
Bridged Systems. A chain of atoms bridging a cyclic system is denoted by B, followed by the number of atoms in the chain and the link positions in parentheses. The new atoms introduced by the bridge are numbered successively from the lower numbered link. Heteroatoms in the bridge and other substituents can thus be located.

Spiro Rings. A spiro ring is denoted by "<" followed by the number of atoms in the ring and the grafting position in parentheses. Numbering succeeds the last numbered atom in the ring system. If there are heteroatoms or substituents in on the ring, numbering will proceed in the direction that will provide the lowest possible indices, e.g.:



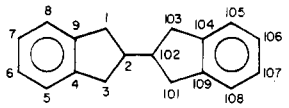
C6B1(1-4) A6A6B2(1-9) C6<3(1)

Complex Ring System. The following rules allow the construction of complex ring systems. A two-dimensional drawing of the system is made. A partial structure consisting of linearly fused rings is constructed first, each new ring adding to the numbering sequence. Bridges are added next. Spiro rings follow and, finally, heteroatoms. When confusion arises, the rule is to make the system with the lowest possible numbered positions. Although fairly elaborate structures can be represented in this way, it is rarely encountered that a synthetic reaction will involve a simultaneous change in more than five rings. E.g.:

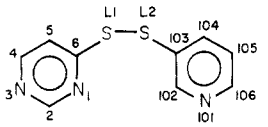


A6C6A6 (2-3) B2 (1-4) C5 (15-16)/O18

Two Ring Systems Linked by a Single Chain. Each ring system of the two is first given the appropriate notation, including numbering and heteroatom assignment. Then, a determination is made as to which one will be considered the major ring system. The major system is the one that has more rings, larger rings, or more heteroatoms or is more heavily substituted. It will be left unchanged. The minor ring system will then have all its positions numbered with the digit 1 preceding each numbered position. The linking chain is denoted by L, followed by the number of atoms in it and the positions of connection with the two systems. Numbering along the link starts at the major ring system side, allowing location of heteroatoms and substituents. The positions are numbered L1, L2, etc. When heteroatoms are denoted at the link, the "L's" are omitted. In the full notation for the system, the notation for the major system is written first, than a hyphen, then the link notation, another hyphen, and finally the notation for the minor ring system, e.g.:



A6C5 -L0 (2-102) -A6C5



A6/N1N3-L2 (6-103)/S1S2-A6/N101

B. Alternative Numbering System for Ring Structures. The Ring Index numbering system for ring structures may be used as an alternative to the one described so far. The advantage is the familiarity of organic chemists with that system. However, the Ring Index is, in fact, a dictionary, the use of which would be eliminated if the set of rules for constructing and numbering ring systems described in section A is adopted. If the Ring Index numbering system is to be used, the following considerations must be made.

(a) Only aromatic (six-membered) and fully saturated rings are supported by the notation system. The Ring Index contains entries that are partially unsaturated (mostly heterocyclic) ring systems. These are not allowed.

(b) Ring junctions in fused aromatic systems are not numbered at all in the Ring Index system. When these become partially or fully saturated, junction positions have the number of the preceding ring atom plus a letter, e.g., positions 4a, 4b, 8a, and 9a in the fluorene system. Confusion may arise with the notation for positions of secondary substituents (see section D). It is much more convenient to have ring positions as simple numbers.

(c) For stereochemistry considerations (see section E), ring systems must be drawn as they appear in the Ring Index. If their mirror images are drawn, erroneous stereochemistry assignments will result.

C. Substituents and Functional Groups. Once the MSP has been defined and a numbering pattern established for the molecular structure, other substituents are written in the order of their numbered positions and priority level. If two or more substituents are located on the same position, their alphabetical order will prevail.

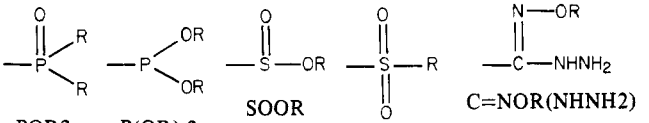
The notation for simple substituents and functional groups follows their trivial formulas, i.e., I, OH, CF3, CN, COOR, SO2OR, NHCOOR, etc. This is in contrast with most notation systems that have meaningless symbols for functional groups. No need for a predetermined set of "legal" functional groups is therefore necessary. The present notation also results in a quick interpretation of the computer printout after a search is complete because of the familiarity of the substituent names. In particular, it enables the use of a unique feature of searching for closely related functional groups, as will be described later.

Some convenient notations for frequently encountered atoms and functional groups follow: C, methyl (only when reacting differently than a longer alkyl chain); R, alkyl; A, phenyl or aryl; <n, a carbon ring with *n* atoms; E, double bond (for "ene"); T, triple bond; O, carbonyl; O< epoxide; X, bromine or chlorine; D, deuterium or tritium; Z, silicone; M, metal.

Heteroatom functional groups that contain aryl groups are written as those containing alkyl groups by substituting A for R: SR and SA for thioalkyl and thioaryl; NHCOR and NHCOA for amides of aliphatic or aromatic acids.

Heteroatom primary functional groups in which the heteroatom is part of a ring skeleton are preceded by ">": >NR = cyclic tertiary amine, >SO = cyclic sulfoxide, >ZA2 = cyclic diaryl silane.

Complex heteroatom functional groups are written in the order of atoms and groups as long as these have a one-letter notation, including a heteroatom–oxygen or heteroatom–sulfur double bond. If an atom is substituted with any combination of hydrogen, alkyl, or aryl groups, their order follows H > A > R. If a complex heteroatom functional group has substituents having a more than one letter notation, the first written are double bond substituents indicated by "=" (except a heteroatom–oxygen or heteroatom–sulfur double bond), the rest are enclosed in parentheses, following an alphabetical order. Equivalent groups are written once followed by their number. Valences not explicitly written are deduced for some important cases. Thus, O means =O; NH2 represents $-NH_2$, but NH stands for =NH; CO is a ketene, =C=O, etc., e.g.:



POR2 P(OR) 2 SOOR SO2R C=NOR(NHNH2)

D. Position Parameters. Each substituent is followed by its numbered position on the MSP or on a primary or secondary substituent. The number is enclosed in parentheses.

The position parameter for a primary substituent is simply its numbered position set by constructing the MSP. Primary substituents on the link in a system of two rings linked by a single chain have L preceding the numbered position.

Secondary substituents are located on an alkyl side chain (R), a carbon ring (<n), or a phenyl ring (A), which are primary substituents. Their position parameter is the primary substituent position followed by O (letter, not zero), A, B, C, etc. (equivalent to the greek letters commonly used) to mark their location on the primary substituent. The O (origin)

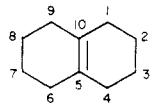
position is used with R or <n secondary substituents for double bonds and epoxy groups that link the atom at the branching point with the atom at position A.

Tertiary substituents are substituents on secondary R, <n, or A substituents or on a functional group carrying an R or A substituent. Their position parameter is denoted by T followed by the position of the secondary substituent and their location on it as for the secondary substituents. When more than one secondary substituent or a functional group are present at the same position, prime signs are used to make the distinction for the tertiary substituent.

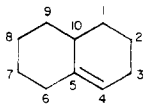
When only one primary substituent is present and its position is irrelevant to the structural formula, the position parameter is "1".

The pyridine derivative given as an example for the substituent priority assignment will have the notation "A6/N1" R(2) E(2A) A(2B) COOR(2B) OH(T2B').

Double bonds and epoxy groups on ring junctions are a special case. Usually their position is denoted by a single number, assuming that the double bond is between that numbered atom and the next one. On ring junctions, confusion may arise, and the positions of the two atoms between which the double bond exists are denoted for those cases when the linked atoms do not have consecutive numbers. The same holds for epoxy groups, e.g.:



"C6C6" E(5-10)

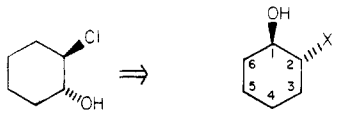


"C6C6" E(4)

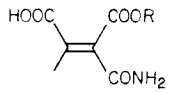
E. Stereochemistry Parameter. Since the way a ring system is drawn and numbered defines a unique orientation, assigning specific stereochemistry to atoms and bonds is a simple matter. The letters R and S following the position number define absolute configurations by the accepted terminology. Since ring systems are drawn flat, substituents will stick above or below the ring plane. An A or B following the substituent position will then represent the actual stereochemistry of that substituent. In ring junctions, A and B will refer to the substituent (including hydrogen), which is *not* part of the ring skeleton. If the junction is of four rings, absolute configuration notations R and S are used.

Configurations of substituents on double bonds will be referred to as C and T for cis and trans in the usual manner, when this bond has only two substituents. If more substituents exist, C and T will refer to the first lower alphabetically ordered substituents on each side of the bond. Configurations around an epoxy group follow the same rules. If no stereochemical features exist, the stereochemistry parameter is a hyphen.

The stereochemistry parameter is placed in brackets during output, e.g.:



"C6" OH (1) X(2) [1A-2B]



"S" CONH2 (1) COOR (1) E(1) COOH (2) [1T]

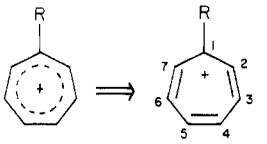
F. Reaction Type Parameter (RTP). This parameter is included in the reaction representation to speed up the search process, since it is the first parameter matched with the user-specified reaction type parameter. Moreover, it serves as a guideline for the organization of the catalog into reaction files, all having the same reaction type, as will be explained in the File Organization section. The following categories of reaction types are used: NC, no change in molecular skeleton, functional group interconversions only; C+, carbon atoms are added to the starting material by forming a new C–C bond; C–, loss of carbon atoms occurs in the starting material by cleaving a C–C bond; RF, ring formation; RO, ring opening; RE, ring expansion; RC, ring contraction; RR, rearrangement.

Any reaction that cannot be simply classified under a single category is denoted RR as well.

During output, the RTP is placed between the starting material and product parameters, preceded and followed by "≫" signs.

G. Special Cases. Charged Species. Charges and counterions are ignored whenever they are obvious, such as in quaternary ammonium salts, sulfur or phosphorus ylides, etc. Also, if a noncharged species can be obtained by removal or addition of a proton, it will be the one used for actual representation, such as an acid rather than its salt, etc. In all other cases, a charge is treated as a substituent, denoted as + or –, followed by its position parameter.

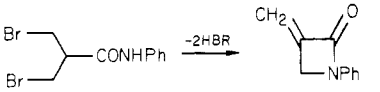
Resonance. In resonating structures (except six-membered aromatic rings), the canonical form chosen for representation will be the one providing lower-numbered positions for the double bonds. If a charge is involved, the form where the charge has a lower-numbered position is preferred, e.g.:



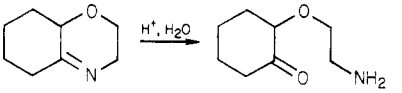
"C7" +(1) R(1) E(2) E(4) E(6)

Complexes. Complexes are difficult to represent in any notation system due to the vague nature of the link between species. Some organometallic and other complexes can be represented in our notation by assigning an "unspecified" position (X) for the attached substituent.

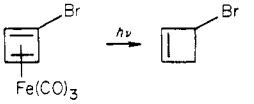
The full notations for four reactions exemplifying the principles outlined so far follow:



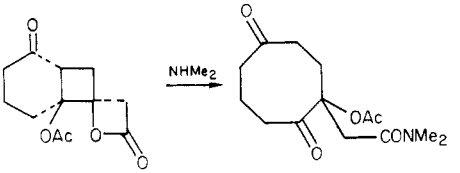
"S" X(1) CONHA(2) X(3) [–] >>RF>>
"C4"/N1 >NA(1) O(2) R(3) E(3O) [–]



"C6C6/N1O4" E(10) [–] >>RO>>
"C6" O(1) OR(2) NH2 (T2B) [–]



"C4" E(1)X(1)E(3) M(CO)3(X) [–] >>NC>> "C4" E(1) X(3) [–]



"C6C4<4(1)/O9" O(4) OCOR(8) O(10) [1R-3B-8B] >>RR>>
"C8" O(1) OCOR(2) R(2) O(5) CONR2(2A) [–]

FILE ORGANIZATION

The exact way the encoded reaction data is arranged, stored, and retrieved depends to a certain extent on the type of computer system used. During the past few years, microcomputers became increasingly affordable for chemical laboratories, where they are used for data analysis and for monitoring instrumentation. Recent developments include faster and more powerful microprocessors, availability of compilers for a variety of high-level languages, and mass storage on hard-disk systems. This field continues to develop at a fast rate, and the type of a data base for organic reactions presented in this paper seems perfectly suitable for the small computer system. However, most disk operating systems for microcomputers support only simple data structures. The actual system outlined in the rest of this paper was designed for application on an average microcomputer.

Each encoded reaction constitutes a record within a reaction file. When constructing the file, the reactions are chosen so that all will have the same RTP and so that there will be as little variance as possible in the reactant MSP's. Ideally they should have all three parameters the same. Thus, a file having the starting material MSP "S", product MSP "S", and the RTP "NC" will contain all functional group interconversions on aliphatic chains. A file where all reaction records have those three parameters as "A6" >>NC>> "A6" is, similarly, a collection of all substitutions and functional group interchanges in aromatic chemistry. The two files where all reaction records have "S" >>RF>> "C5" and "S" >>RF>> "C6" as common MSP's and RTP's also represent two large groups of cyclization reactions where five- and six-membered rings are formed. Grouping organic reactions into files having the same basic reactant structures and reaction type constitutes essentially an alternative approach to the classification of organic reactions, with the advantage of being a basis for a retrieval system.

Further subdivisions and hierarchies in the records within a file can be incorporated. This may be done on a basis of reactant structure complexity (numbers of parameters in each priority level, presence of heteroatoms in substituents, etc.) or reaction complexity (e.g., the extent of difference between the starting material and the product).

With less frequently encountered reactant skeletons it may no longer be practical to assign files to small numbers of reactions (this is not true for larger computers), and a limited variance in the MSP's is allowed. Since there are only seven different RTP's, reaction records within those files will always have the same RTP. The variance in MSP's allowed in our system is in the field length of the parameters.

HANDLING PROGRAM

The handling program performs the necessary utility operations such as introduction of new reactions, correcting or replacing reactions already stored, and preparation of backup file copies. Naturally, most important is its searching part, designed for maximum flexibility in substructure searching.

The catalog can be searched according to starting materials, products, or both, which is equivalent to answering the following questions, constituting the essence of organic synthesis planning: (a) List all the reactions in the catalog that a specified starting material undergoes. (b) List all the reactions that lead to a specified product. (c) List all the reactions in which a specified starting material is converted to a specified product.

Search is conducted in stages, and a record is rejected if at any stage the match is negative. When data is input into the reaction files, the program prepares and updates a "directory" file which is first looked into to determine which files of the catalog will be searched, according to the user-specified MSP's and RTP. Within a reaction file, each record is checked for the MSP's again in case of variance (hashing functions can replace the directory file if there is no variance in the MSP's throughout the whole data base). Next, the numbers of substituent parameters specified by the user in each priority level is compared to those in the reaction record. Further searching into this reaction will not continue if the user specified more substituents than are present in the record.

The user may wish to perform an exact match of structural and position parameters or a partial match. During an exact match, a one-to-one comparison is performed between the user-specified structural and position parameters and those stored in the file. If this fails to provide an output, the user has the option of requesting a partial match. Here, reactions that have the user-specified parameters as a subset of all the stored structural parameters will be retrieved. Because the user-specified structural parameters are fewer in number than those in the file, they will have in most cases different position parameters than those in the reaction file. A special subroutine then establishes the relationship between the substituents, regardless of their absolute position numbers, and will retrieve a reaction stored in the catalog that has a partial set of structural parameters having the same relationship between themselves as the user has specified. Some fruitful synthetic ideas may arise from such a search, since the extra functional groups present in the computer-retrieved compounds may be easy to get rid of or may even be used further in the next synthetic steps.

Other types of substructure search under partial match include the following options: (a) search for a specified substituent at any position, (b) search for any substituent at a specified position, and (c) search for any substituent at any position in a specified priority level.

Options a–c can be carried out for any number of substituents/positions. Case c can be illustrated as a response to a request of the type "list all reactions in the catalog leading to a specified molecular structure containing a functionalized side chain at any position". That extra functionality may be needed for copolymerization, linking to a fluorescent marker, etc.

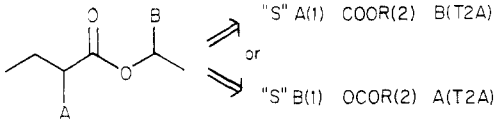
A unique option which is possible with our notation system for functional groups is a search for closely related or "families" of functional groups that may be interconverted easily. Both under exact or partial match, the user may specify a functional group by one or more of the first letters composing its notation followed by an asterisk. The program will match only those letters in the substituent field in the record even if the notation is longer and retrieve a reaction if such a match is positive. Thus, if the user looks for an OH group at a certain position and specifies O* as a query, a retrieval will occur with OH, OR, O, OA, OCOR, OCOA, OCOOR, O<, etc., all convertible to a hydroxyl. In principle, this type of search can go as "deep" into the substituent notation as required.

EXPERIENCE WITH THE SYSTEM

The first 400 numbered reactions in the 1980 volume of Theilmeier's "Synthetic Methods" were used as a sample of organic reactions from various fields of synthesis. Some of them contain multiple reactions that brought the total number of coded reactions to 530. All could be satisfactorily represented by the above-described notation. We found that after a short training, coding of reactions and interpretation of the computer output upon searching is fast and easy. The determination of the relevant parts of the reactants participating in the reactions may, however, depend on personal viewpoint as to the role of remote substituents in directing the reaction course. Since the computerized search serves as a guideline for synthesis planning, we feel that providing the chemist with as many examples as possible to choose from is best.

Therefore, we determined as a relevant reactant structure the one obtained by stripping the starting material and product molecules of alkyl side chains and all aryl and functional group substituents not changed by the reaction and that are two or more saturated carbon atoms away from the reaction site. In those reactions where one of two or more identical functional groups is selectively changed, however, all these functional groups are recorded.

Although the notation system described so far defines a molecular structure unambiguously, the opposite is not always true, and certain reactant structures may have more than one representation in the encoded form. For example, an alkylbenzene may be represented as an aromatic ring substituted by an alkyl side chain or as a saturated carbon chain substituted by a phenyl group. MSP's will be different, and substituents on the aromatic ring and on the carbon chain will be assigned different priority levels in each representation. A second common case is illustrated by the following ester structure, having substituents A and B at carbons next to the



ester group on both the alcohol and the acid. It is clear that the two representations are so different that one will not be retrieved if the other was specified as a query, even under partial match.

In dealing with cases of possible multiple representations and the resulting possibility of confusion, one has to remember that our system is a *reaction* catalog and not a catalog of structures. The user is not interested in a literature reference to a certain molecular structure but in a reaction in which this structure participates as a reactant. In the two examples given above, if the phenyl ring was not actually changed by the reaction (e.g., by substitution) but changes did occur at the alkyl chain, less useful synthetic information would be obtained from specifying the aromatic ring as the MSP in a query, since by so doing, the user expects the main reaction to occur at the ring. Similarly, if changes in the reactant structures as a result of the reaction occur only at the acid side of the ester in the second example, A should be coded as the primary substituent and B as a tertiary one. As a general rule, if all reactant functional groups that change during reaction can be coded as primary substituents, there will be no need for an alternative coding. Observing this principle of reaction-oriented coding for the reactant structure reduces the number of cases where multiple coding is necessary to a useful minimum. Thus, only 6% of the reactions in the sample from "Synthetic Methods" had to be coded more than once.

Storage space and search speed were considered too. The handling program was written in BASIC for a Commodore CBM microcomputer with 5.25-in. double-density diskette storage. About 2000–3000 reactions can be stored on one diskette, depending on their complexity and keeping the accompanying text to a minimum. Continuous search of two diskettes in one operation is possible. As for search speed, with various search modes, searching time is about 0.6 s per average retrieved reaction. Less time is needed for rejecting a reaction record, and only a fraction of the data base is searched anyway. For microcomputer systems where compilers are available, speed may be increased by one or two orders of magnitude. Adapting the system for use with mainframe computers has its obvious implications as to processing speed and storage capacity.

ACKNOWLEDGMENT

I thank Dr. Louis A. Carpino for helpful suggestions.

REFERENCES AND NOTES

(1) Hendrickson, J. B. "A Systematic Organization of Synthetic Reactions". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 129–136.
(2) Bersohn, M.; Mackay, K. "Steps Toward the Automatic Compilation of Organic Reactions". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 137–141.
(3) Willett, P. "Computer Techniques for the Indexing of Chemical Reactions". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 156–158.
(4) Silk, J. A. "Present and Future Prospects for Structural Search of the Journal and Patent Literature". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 195–198.
(5) Mosby, M. A.; Kier, L. B. "Methods for Generating a Chemical Reaction Index for Storage and Retrieval". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 217–221.
(6) Bawden, D.; Devon, T. K.; Jackson, F. T.; Wood, S. I.; Lynch, M. F.; Willett, P. "A Qualitative Comparison of WLN Descriptors of Reactions and the Derwent Chemical Documentation Service". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 90–93.
(7) Fugmann, R. In "Chemical Information Systems"; Ellis Horwood: Chichester, England, 1975; p 195–226.
(8) Schier, O.; Nubling, W.; Steidle, W.; Valls, J. "A System for the Documentation of Chemical Reactions". *Angew. Chem., Int. Ed. Engl.* **1970**, *9*, 596–598.
(9) Lynch, M. F.; Willett, P. "The Automatic Detection of Chemical Reaction Sites". *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 154–159.
(10) McGregor, J. J.; Willett, P. "Use of Maximal Common Subgraph Algorithm in the Automatic Identification of the Ostensible Bond Changes Occurring in Chemical Reactions". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 137–140.
(11) Willett, P. "The Effect of Screen Set Size on Retrieval from Chemical Substructure Search System". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 253–255.
(12) Willett, P. "The Evaluation of an Automatically Indexed, Machine Readable Chemical Reaction File". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 93–96.
(13) Ziegler, H. J. "Roche Integrated Reaction System (RIRS). A New Documentation System for Organic Reactions". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 141–149.