decades. Perhaps due to U.S. case law, the U.S. collections are the major depositories for microorganisms named in U.K. patents.

The questionnaire also revealed that difficulties have been encountered in obtaining the strains named in patents. The sample of British patents studied showed that some strains are not listed in current culture collection catalogs, or are incorrectly listed. It is theoretically easy to check that restrictions on the dispatch of strains to requestors has been lifted, as most culture collections do not list strains until after the grant of the patent. However, there is a complication in that many "current" catalogs are several years out of date, and of the ones studied only the ATCC has a practiced policy of publishing a new catalog every two years. Hence there is often a delay after grant of the patent before the casual enquirer can discover the status of a strain without writing directly to the culture collection. In addition, the NRRL does not publish a catalog, and all information on strains has to be obtained through communication with the collection.

The picture is further complicated by the fact that an appreciable number of strains have a different name in the patent and in the catalog. This is doubtless partly due to the uncertain state of some areas of microbial taxonomy, and to the not infrequent changes in microbial nomenclature. It is also likely that the employees of commercial organizations will not be experts in all areas of microbial taxonomy. Where culture collections have a numerical list of strains, an accession number has been given in the patent, and the catalog gives some details, such as depositor and patent number; this merely represents a time-consuming nuisance. But not all patents give an accession number, despite the fact[5] that U.K. law requires this. And not all catalogs have numerical lists of strains: the CMI and NCYC catalogs do not have such lists. Further, not all catalogs give enough details to allow the strain to be identified with certainty.

None of the catalogs studied here carry a taxonomic tree of the microorganisms, or details of genera which have undergone changes in name. So there are no clues to help the searcher find strains which are incorrectly listed.

It is therefore not surprising that U.K. law requires that the organism should be described in the patent, as well as deposited in a culture collection. The problems of taxonomy of actinomycetes, especially *Streptomyces* species, coupled with this requirement, results in some excellent descriptions of strains in some antibiotic patents, particularly where the strain is a new isolate. Such patents become a useful source of information on the strain concerned, and it would be interesting to know how many of these strains are later described in the journal literature, and whether any such articles carry as detailed a description as the patent.

In terms of the microorganisms, the sample patents were most useful as sources of information on actinomycetes, and least useful for algae, well-known yeasts, and disease-causing organisms. Information on other bacteria, fungi, and yeasts was intermediate.

The culture collection catalogs can form a useful quick reference source on microorganisms. The current NCIB and ATCC catalogs give considerable details of depositor, patents, and references to the strain in the journal literature, although the NCIB is discontinuing its policy of scanning the journals for references to NCIB strains.[6] Other catalogs, such as the NCTC and NCYC catalogs, give very few details, and are therefore only useful as a first check on whether the strain is available from the collection.

We therefore conclude that patents are an important source of information on microorganisms, a result that is in accord with other work emanating from this Department.[7] However, the unreliability of culture collection catalogs means that a searcher wanting to investigate a particular culture further may have problems finding out if a given microorganism can be obtained or if it has been mentioned in a patent.

REFERENCES AND NOTES

(1) "Budapest Treaty on the International Recognition of the Deposit of Microorganisms for the Purposes of Patent Procedure", *Ind. Property*, 1–19 (1977).
(2) Anon., *Off. J. Eur. Patent Off.*, **1**, 34 (1978).
(3) C. Oppenheim, "The Importance of Industrial Property to the Pharmaceutical Industry", paper presented to 5th AIOPI Conference, 1978.
(4) C. Oppenheim, "Recent Changes in Patent Law and Their Implications for Information Services and Information Scientists", *J. Doc.*, **34**, 217–229 (1978).
(5) "The Patent Rules", Statutory Instrument 1978, No. 216, H.M.S.O., London, 1978.
(6) I. J. Bousfield, private communication.
(7) C. Oppenheim and E. A. Sutherland, "Case Study on Calvalume", *J. Chem. Inf. Comput. Sci.*, **18**, 126–9 (1978).

Substructure Search with Queries of Varying Specificity

ALFRED FELDMAN*
Division of Biometrics, Walter Reed Army Institute of Research, Washington, D.C. 20012

LOUIS HODES
National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20014

Received February 7, 1979

Efficient screening of queries of varying specificity requires a rich endowment of low specificity screens. However, most systems are heavily unbalanced toward high specificity. A more equitable mix can be achieved through use of the authors' method of screen generation.

INTRODUCTION

Systems using substructure search (SSS) depend on screens[1] to reduce the number of atom-by-atom searches. A measure of the power of a screening system is the closeness with which this number of atom-by-atom searches approximates the number of actual matches. In large files relatively few matches, proportionately, are usually required. Therefore, high screenout becomes almost synonomous with power in large files.

Generally, one wants the most powerful screening system for the widest collection of queries. Most screening systems work well on certain classes of queries but fall embarrassingly short on certain other classes. These often can be answered precisely only by virtue of an inordinate number of atom-by-atom searches for a few matches.

* Address correspondence to author at National Cancer Institute, National Institutes of Health, Bethesda, Md. 20014.

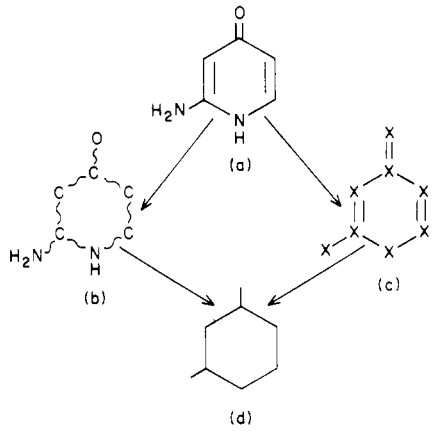


Figure 1. One structure represented under different levels of specificity.

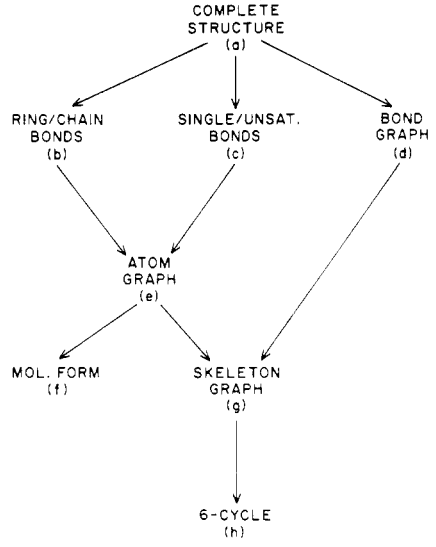


Figure 2. Relationships among different levels of specificity.

The reason is not so much that queries are unpredictable, as that expediency of design sacrifices the ability to process certain types of queries. For example, current screening systems fail to properly utilize specificity levels in structures. The lower specificity screens are often represented incompletely, inconsistently, and/or inadequately. Moreover, this shortcoming may be a necessary consequence of inefficient screening.

In order to demonstrate the pertinence of the foregoing statements it will be helpful to review some familiar concepts and to introduce some terminology.

DEGREES OF SPECIFICITY

Chemical structures will be considered to be at their highest degree of specificity when expressed in their usual manner (Figure 1a) as undirected graphs with labeled nodes (atoms) and labeled edges (bonds). Structures on file are normally stored at this degree of specificity, though there may be small deviations[2] in individual SSS systems. Partially undefined file structures will not be treated here.

Other useful degrees of specificity are the atom graph degree (1b) where bonds are unspecified, and the bond graph degree (1c) where atoms are unspecified. The latter two degrees are not comparable with each other but are both of higher specificity than the skeleton graph degree (1d), and all three are of lower specificity than the complete structure degree.

Further, as shown in Figure 2, between the atom graph degree and the complete structure degree one can insert degrees where bonds are partially specified, for example, according to their ring/chain character (2b) or their single/unsaturated character (2c). The molecular formula degree (2f) is of lower specificity than the atom graph degree (2e) but incomparable to the skeleton graph degree (2g). The coding of complex 6-rings (2h) by Bedrosian and Milne[3] is of even lower specificity than the skeleton graph degree and incomparable to the molecular formula degree.

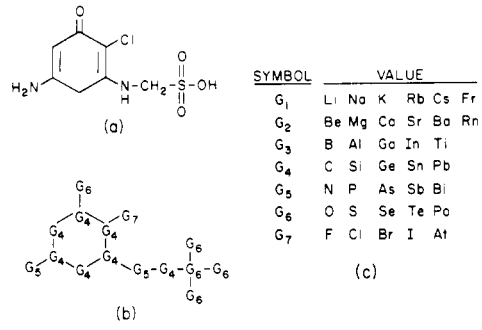


Figure 3. Lower levels of specificity are obtainable by collapsing groupings of identifiers.

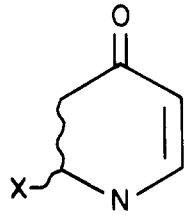


Figure 4. Hybrid structure of more than one degree of specificity.

One can form a variety of intermediate degrees by collapsing elements occurring at the same position in the periodic table as shown in Figure 3c. Used to its full extent this would transform the structure in Figure 3a to that in 3b. Often, this collapsing is not performed uniformly, but only on selected groups such as halogens and metals. Any given degree of atom specificity can be combined with a degree of bond specificity, multiplying the number of degrees. Not all are equally useful.

Furthermore, several degrees of specificity can coexist in a single structure. The example in Figure 4 has one unspecified atom and two unspecified bonds. This kind of structure can be called a hybrid of more than one degree of specificity. A structure of uniform specificity will be called normal.

A set of normal structures at a single degree of specificity will be called homogenous. When two or more homogenous sets are combined, the resulting union will be called heterogenous if it contains structures of differing degrees of specificity. Thus, by definition heterogenous sets contain only normal, not hybrid, structures.

SSS SCREENING SYSTEMS

There are many types of screening systems. Some employ a dictionary; others are open-ended. Some are primarily pragmatic; others algorithmic. There is variation in the types and sizes and specificity of the screens, and in the sophistication of the generation algorithms, where such are used. All, however, have certain characteristics in common.

To a first approximation, all SSS screening systems use a heterogenous set of screens. That is, any departure from normality as just defined turns out to be of a trivial uniform nature. Exhaustive generation of hybrid structures, even at relatively small size, would produce an enormous number of screen structures.

Furthermore, SSS systems, which deal mostly with well-specified queries, need to obtain a high screenout on these. To accomplish this, they usually concentrate screens at the higher degrees of specificity.

For highly specific screens, a small increase in size, under all possible combinations, produces a large increase in the

number of structures. Thus, there are strong limitations on the size or scope of screens obtained by exhaustive generation. The so-called augmented atom screens[4] are an example of a set of screens that is limited in size. Screens that consist solely of rings and ring systems are an example of reduction in scope.

In several SSS systems, low specificity screens have been derived from corresponding more specific screens, usually at little cost. A good example is the augmented atom without bond specificity. These screens can be a useful complement to the corresponding more specific screens, but they are much less numerous. They are usually stored in a heterogenous manner, such as a tree structure.[5]

The NCI system[6] (designed by the University of Pennsylvania) jumps from augmented atom to ring and nuclei screens at several degrees of specificity. Some other systems, e.g., BASIC, WRAIR, have screens of intermediate size, but they compromise by lowering the specificity.[7] Perhaps the largest set of intermediate screens is contained in the TSS screens; these have full specificity but are restricted to chain bonds and they do not branch but indicate branch points.[8]

Most systems have miscellaneous, quasi-structural screens which can include items like atom counts, bond counts, ring counts, etc. These can be considered to have low specificity but, because of their limited structural content and their small number, are not too useful.

From the foregoing, it can be concluded that a SSS screening system usually has several fairly homogenous classes of relatively high specificity screens. Where low specificity screens occur they fall into the following three categories:

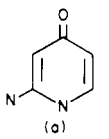
1. They are a simple collapsing of existing high specificity screens, or another representation of the same structures.
2. They are forced, because higher specificity screens would yield too many screens. In this case one finds intermediate specificity rather than low specificity for its own screening power.
3. They are fairly trivial screens such as atom counts and bond counts. These are undiscriminating under relatively small SSS queries.

SCREENING FOR QUERIES OF VARYING SPECIFICITY

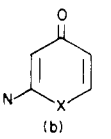
The submission of a low specificity query to a large file would normally produce an overwhelming number of responses. As a consequence, low specificity queries are not often experienced, unless they are submitted for statistical purposes or the structures are so large that size or complexity limits the response. However, there are queries that contain more than one degree of specificity, queries that, in our terminology, are hybrid structures. These queries usually have a reasonable number of expected responses.

Difficulties may easily arise in handling these hybrid queries if a system is not designed for them. As an example, just one unspecified atom in the query, illustrated in Figure 5b, nullifies 10 out of 16 screens, including the most effective ones. The failure occurs because screens can be used only if they are at the same, or at a lower, degree of specificity than that in which they query is expressed—never at a higher degree. In current SSS screens this failure is especially noticeable with small hybrid queries and with larger low specificity queries, e.g., those involving the complex 6-rings ("chickenwire") mentioned earlier.

For an SSS to obtain power on these reasonable types of queries, it is necessary to provide a sufficient number of low and intermediate specificity screens. Because lower specificity screens generally apply to larger portions of a file than the corresponding higher specificity screens, the former, to be useful in comparison to the latter, must be extended in size. There should be, in other words, an inverse relationship between specificity and screen size. This means, for example, that skeleton graph screens should extend to greater size and variety than the corresponding higher specificity screens.



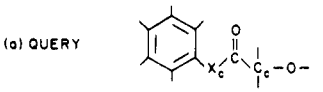
(a)

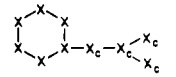
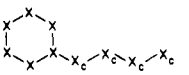


(b)

SCREEN NUMBER	INCIDENCE %	DESCRIPTION	USED FOR (a)	USED FOR (b)
1-13	59.8	N	x	x
1-10	62.8	O=	x	x
3-7	40.5	C(C,C,O)	x	x
3-5	56.3	N(C,C)	x	
3-16	18.5	C(N,N)	x	
3-24	7.8	C(C,N,N)	x	
4-3	61.0	C-C-C-O	x	x
4-2	62.1	C-C-C-N	x	x
5-12	15.0	N-C-C-C-O	x	x
6-16	12.2	C-C-C-C-C-N	x	
6-14	13.0	C-C-C-N-C-N	x	
7-23	8.7	C-C-N-C-C-C-O	x	
7-63	3.1	N-C-N-C-C-C-O	x	
7-178	0.8	C-C-N-C-C-C-O	x	
7-129	1.1	C-C-C-C-N-C-N	x	
7-30	7.0	N-C-C-C-C-C-N	x	

Figure 5. Screens obtained for a specific structure (a), and screens obtained for the same structure with one atom made nonspecific (b).



(b) SCREEN FRAGMENTS	(c) SPECIFICITY LEVEL *)	(d) % INCIDENCE
	1	22.70
	1	36.52
O~C~C~O	2	11.59
C—O	3	41.04
C=O	3	49.59

*) LEVEL 1 = SKELETON GRAPH, WITH DISTINCTION OF ATOMS AS TO RING OR CHAIN MEMBERSHIP.
LEVEL 2 = SPECIFIC ATOMS, UNSPECIFIED BONDS.
LEVEL 3 = SPECIFIC ATOMS AND BONDS.

Figure 6. Query with one unspecified atom, screened by screens of various specificities.

The practice of developing screens in a hierarchy from low to high specificity, as exhibited in the NIH/EPA system[5] or by the CAS graph/node/bond ring descriptor,[9] does not provide this inverse relationship.

Figure 6 contains a simple illustration of a hybrid query using screens at three degrees of specificity. The incidence of each of the screens in a sample file is shown.[15] The joint incidence of all five screens yielded a 98.5% screenout. Even with that screenout, the ratio between screen hits and atom-by-atom matches was 22:1. This demonstrates the principle of screening for hybrid queries by means of homogenous screens of different degrees of specificity.

In order to make searches for hybrid queries efficient, several sets of screens, at as many degrees of specificity, are likely to be required. Each set should be a maximally specific fallback position for the next higher set of screens. Further, to be effective, each homogenous set must be developed independently of the screens available at other degrees.

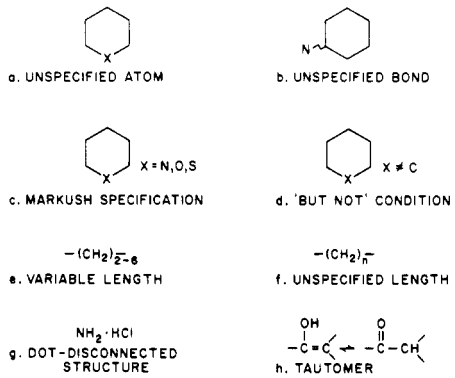


Figure 7. Some common problem queries.

It is apparent that the hierarchical method used at WRAIR to generate screens[10] can be used to generate a homogenous set at any given degree of specificity. The method enables generation of a "horizontal" hierarchy within a degree, which becomes preferable to the "vertical" hierarchies across degrees mentioned earlier. Moreover, the screens so generated are economical, so that several sets can be supported without taxing a system.

The WRAIR method allows screens to "grow" one step at a time, producing a number of offspring. The marginal discrimination[11] of a new screen is computed according to its incidence relative to the incidence of its parent of least incidence. The process continues until a screen occurs with incidence below a cutoff where parenthood is proscribed. This process would be carried out independently at selected degrees of specificity, resulting in independent sets of screens. Most likely, all the screens for a single compound can be generated during one pass through its connection table.

In this method, the storage requirement would be at most on the order of 100 bits per compound for each homogenous set of screens. This economy has been achieved for the entire set of WRAIR screens by coding each screen with a number of bits in accordance with its marginal discrimination, and then superimposing the codes for all screens applicable to a given compound. These storage requirements compare favorably with other methods and would allow the storage of several homogenous sets of screens.

OTHER PROBLEM AREAS

Of course, there are several other types of problem queries in SSS systems. Some of the most common ones are shown in Figure 7. So far the discussion has concentrated on query types 7a and 7b. Any of the exhibited types can occur in combination with others.

Queries of type 7c and 7e can be separated into individual cases and treated as disjoint queries. In some SSS systems the common parts can introduce efficiency. Generally, combinations of this kind of indeterminacy multiply the number of disjoint queries so that a limitation is imposed at the screening phase of SSS.

With some care, this procedure can also be used on queries of type 7f which contain indeterminate length chains. Even though all possible lengths are specified by the query, only a limited number of screens will be applicable. This is a simple consequence of the finiteness of the screening system.

Tautomers as in (7h) can always be treated as though the tautomeric bonds are unspecified. The shifting hydrogen presents less of a problem, since hydrogen atoms are often excluded from SSS schemes. See Feldman[12] for a discussion of the automatic detection and search for tautomeric structures.

Dot disconnected structures such as the one shown in Figure 7g, include many salts, hydrated compounds, and mixtures.

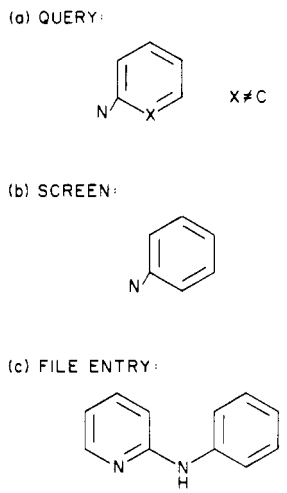


Figure 8. (a) Query specifying X as representing any atom with required valence, but not carbon. (b) Negated screen. (c) File entry matching, in part, the query request, and in part, the negated screen.

They are not too common as SSS queries but are included here because of some experience with them in the WRAIR system. Originally, as is often done, such structures were assigned all screens applicable to each of the components. But for filing purposes, the "major" component had to be identified. It was thought that it could be obtained from a substructure search. In actual operation, however, it was found that SSS during file update[13] resulted in an excessive number of file accesses and atom-by-atom searches. The remedy was to keep on file two sets of screens, one for the mixture and one for the major component. For file update the latter could then be searched on the basis of an identity match.

The most troublesome questions remain those involving the "not" operator (7d). The use of screens to eliminate undesired structures entails the risk of eliminating acceptable structures as well. This occurs when the screen used for elimination is matched elsewhere on a file structure, as illustrated in Figure 8. It is possible to design a screening system using "not" screens to deal with negation as described elsewhere.[14] Lacking this, the proper way to handle this query is to consider the negated parts as unspecified (7a in this case) for the purposes of screening and to satisfy the negation upon atom-by-atom searching. This again demonstrates the need for powerful low specificity screening.

ACKNOWLEDGMENT

We wish to thank Mr. Ray Theisen for running searches yielding the statistics used here.

REFERENCES AND NOTES

(1) L. C. Ray and R. A. Kirsch, "Finding Chemical Records by Digital Computer", *Science*, **126**, 841 (1957).
(2) See Goppelt structures in ref 12.
(3) S. D. Bedrosian and M. B. Milne, "Graphical Representation for Automated Retrieval of a Class of Fused Six-Rings", *J. Chem. Inf. Comput. Sci.*, **17**, 47 (1977).
(4) G. W. Adamson, M. F. Lynch, and W. G. Town, "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. II. Atom-Centered Fragments", *J. Chem. Soc. C*, 3702–6 (1971).
(5) R. J. Feldmann et al., "An Interactive Substructure Search System", *J. Chem. Inf. Comput, Sci.*, **17**, 157–64 (1977).
(6) S. Richman, G. F. Hazard, and A. Kalikow, "The Drug Research and Development Chemical Information System of NCI's Developmental Therapeutics Program", in "Retrieval of Medicinal Chemical Information", ACS Symposium Series, American Chemical Society, Washington, D.C., 1978.
(7) The BASIC screens include linear sequences of length four to six atoms with bond specificity at the ring/chain level and some collapsing of atom types. The WRAIR screens have chains of length four to seven atoms without bond specificity except for ring closures. They then continue from eight to eleven atoms in size with hybrid ring chain structures.

Relevant publications are H. R. Schenk and F. Wegmuller, "Substructure Search by Means of the Chemical Abstracts Service Registry II System", *J. Chem. Info. Comput. Sci.*, **16**, 153–61 (1976), for the BASIC system; and J. Page, R. Theisen, and F. Kuhl, "The Walter Reed Army Institute of Research Chemical Information System", in "Retrieval of Medicinal Chemical Information", ACS Symposium Series, American Chemical Society, Washington, D.C., 1978.
(8) M. Milne, D. Lefkovitz, H. Hill, and R. Powers, "Search of CA Registry (1.25 Million Compounds) with the Topological Screens System", *J. Chem. Doc.*, **12**, 183–189 (1972).
(9) P. G. Dittmar, R. E. Stobaugh, and C. E. Watson, "The Chemical Abstracts Service Chemical Registry System. I. General Design", *J. Chem. Inf. Comput. Sci.*, **16**, 111–121 (1976).
(10) A. Feldman and L. Hodes, "An Efficient Design for Chemical Structure Searching. I. The Screens", *J. Chem. Inf. Comput. Sci.*, **15**, 147–52 (1975).
(11) L. Hodes, "Selection of Descriptors According to Discrimination and Redundancy. Application to Chemical Structure Searching", *J. Chem. Inf. Comput. Sci.*, **16**, 88–93 (1976).
(12) A. Feldman, "An Efficient Design for Chemical Structure Searching. III. The Coding of Resonating and Tautomeric Forms", *J. Chem. Inf. Comput. Sci.*, **17**, 220–223 (1977).
(13) The WRAIR system uses the SSS screens for full structure search. See L. Hodes and A. Feldman, "An Efficient Design for Chemical Structure Searching. II. The File Organization", *J. Chem. Inf. Comput. Sci.*, **18**, 96–101 (1978).
(14) L. Hodes, "A Square Root Algorithm for Inclusive Matching. Application to Chemical Structure Searching", Proceedings of the Conference on Computer Graphics, Pattern Recognition, and Data Structures, IEEE Computer Society, 1975.
(15) This file contained about 17 000 entries, representing a random sample of the quarter-million entry WRAIR file.

A Systematic Organization of Synthetic Reactions†,[1]

JAMES B. HENDRICKSON

Edison Chemistry Laboratories, Brandeis University, Waltham, Massachusetts 02154

Received September 12, 1978

A base of description of organic structure first defines the kind and number of four important kinds of attachments to a single carbon. From this may be derived two interconnected descriptions of a reaction. Linear digital descriptions of the strand of carbons involved in the reaction may be listed for substrate and for product. Also the change in the reaction may be denoted by the kind of attachment made and that broken for each carbon. A single reaction step is defined as one with only a unit exchange in attachment at any involved carbon. This system allows a systematic organization of all possible synthetic reactions in a simple but rigorous format. The basic parent reactions are developed first, then a set of ways to modify these to more complex variants. The reactions can be presented in a compact graphical form, and the system can be utilized to describe and find all possible pathways for multistep reaction sequences between specified generalized substrates and products.

The reactions of organic chemistry constitute an enormous quantity of information, their number reflecting the fineness of description. There are also many possible ways to organize this collection for presentation, depending on the focus of interest, and no single mode dominates present practice. The system developed here is aimed at the needs of the synthetic chemist and consists of a set of nested categories, the logic of which assures that all possible reactions are included. Hence any possible reaction has a place in the system fixed by its logic. This approach assures that one can find any wanted reaction and also can see what reaction possibilities are yet unknown in practice and waiting to be invented.

Chemists commonly express generalized reactions as partial structures of substrate and product bearing only those atoms and bonds which change in the reaction. With several adjacent carbons involved, there can be a great number of possible variants. The present system offers a concise logical notation of these part structures in order to generate and organize all possible variants in a structured way. The basis for the notation is a numerical, or digital, one easily adapted to a rigorous system, the completeness of which is then simply defined by mathematical combinations, unbiased as to current chemical practicality.[1b]

The system describes each involved carbon separately in terms of four kinds of attachment basic to synthetic operations. Any reaction or reaction sequence is then the net change in these attachments at each involved carbon, between the substrate and product partial structures. The kinds of attachments defined for any carbon are: H for hydrogens, R for single (σ) bonds to carbon, Π for double (π) bonds to carbon, and Z for bonds (σ or π) to electronegative heteroatoms.

The numbers of each kind of attachment are, respectively, h, σ, π, and z, limited by valency to $h + \sigma + \pi + z = 4$. The "skeletal level" of any carbon is then its value of σ ($\sigma = 1$, primary; $\sigma = 2$, secondary, etc.), and the "functional level" (f) is the sum of attachments to heteroatoms (z) and Π bonds to other carbons (π, limited to 2), symbolized as $f = z + \pi$. The distinction in kind of functionality is denoted by placing one or two overbars over the value of f to indicate $\pi = 1$ or 2, respectively (examples in Figure 1). Metals or other electropositive attachments are included in H, i.e., as the conjugate acid. Hence any carbon is characterized by two digits (each of 0–4), the skeletal value, σ, and the functionality value, f, such that $\sigma + f = 4 - h$. The oxidation state of any carbon is then simply given by $x = z - h$, with values of $-4 \leq x \leq 4$. The functional oxidation state, $x' = 2f - \pi$, is useful in describing reactions and is discussed below. This digital mode of description is based on fundamentals and is very easy to assimilate. For cataloging purposes and for most general use, the system does not require the use of a computer. Used by hand, this notation system can drastically simplify the chemist's consideration of synthetic problems.

CHARACTERIZATION OF REACTIONS

In any synthetic reaction only a few of the involved carbons change their attachments, and these carbons are always

† Presented in the symposium, "Retrieval, Analysis, and Indexing of Chemical Reactions", 176th National Meeting of the American Chemical Society, Miami Beach, Fla., Sept 12, 1978.

0095-2338/79/1619-0129$01.00/0