sighted administrators who initiated this project and sustained it in its early stages. SECS program development at Santa Cruz has been supported by the National Institutes of Health, Research Resources Grant No. RR-01059, and through computer support from the Stanford University SUMEX-AIM Project Grant No. RR-00785.

REFERENCES AND NOTES

(1) E. J. Corey and W. T. Wipke, *Science*, **166**, 178 (1969).
(2) Review: M. Bersohn and A. Esacks, *Chem. Rev.*, **76**, 269 (1976).
(3) Review: P. Gund, *Annu. Rep. Med. Chem.*, **12**, 288 (1977).
(4) "Computer-Assisted Organic Synthesis", W. T. Wipke and W. J. Howe, Eds., ACS Symposium Series, No. 61, American Chemical Society, Washington, D.C., 1977.
(5) L. H. Sarett, unpublished speech before Synthetic Organic Chemical Manufacturers Association, June 1964; quoted in ref 6.
(6) W. T. Wipke, "Computer Representation and Manipulation of Chemical Information", W. T. Wipke, S. R. Heller, R. J. Feldmann, and E. Hyde, Eds., Wiley, New York, 1974, p 147.
(7) P. Gund, J. D. Andose, and J. B. Rhodes, in Contributed Papers to the 3rd International Conference on Computers in Chemical Research, Education and Technology, E. V. Ludeña and F. Brito, Eds., Centro de Estudios Avanzados del Instituto Venezolano de Investigaciones Cientificas (IVIC), Caracas, 1977, p 226.
(8) Review: W. T. Wipke, H. Braun, G. Smith, F. Choplin, and W. Sieber, in ref 4, p 97.
(9) "Utilization of Stereochemistry and Other Aspects of Computer-Assisted Synthetic Design", T. M. Dyott, Ph.D. Thesis, Princeton University, 1973 (University Microfilms, No. 74-9677).
(10) Our current computer environment consists of an IBM 370/168 computer running under the OS/VS2 MVS-3.7A operating system. We have also run SECS on TSO under the OS-MFT and OS-MVT operating systems. We are using DEC GT-42 and GT-43 graphics display terminals with 16 K words of core memory. Communication with our host computer is at 1200 baud using VADIC 3400 series modems.
(11) Different releases of the SECS program are identified by different version numbers, starting with 1.0. The current release of the program corresponds to version 2.7.
(12) Initially available through First Data Corp., Waltham, Mass., the program is now available on ADP Network Services, Inc., Ann Arbor, Mich.
(13) B. Dominy, "SECS and the Information Scientist", presented at the Science Information Subsection of the Pharmaceutical Manufacturer's Association Meeting, Washington, D.C., March 6, 1977.
(14) E. J. Corey, W. J. Howe, H. W. Orf, D. A. Pensak, and G. Petersson, *J. Am. Chem. Soc.*, **97**, 6116 (1975).
(15) R. F. Shuman, S. H. Pines, W. E. Shearin, R. F. Czaja, N. L. Abramson, and R. Tull, *J. Org. Chem.*, **42**, 1914 (1977).
(16) S. H. Pines, R. F. Czaja, and N. L. Abramson, *J. Org. Chem.*, **40**, 1920 (1975).
(17) J. Ten Broeke, A. W. Douglas, and E. J. J. Grabowski, *J. Org. Chem.*, **41**, 3159 (1976).
(18) H. Bruns, private communication.
(19) P. Gund, J. D. Andose, and J. B. Rhodes, in ref 4, p 179.
(20) H. Bruns, *Naturwissenschaften*, **66**, 197 (1979).

The Evaluation of an Automatically Indexed, Machine-Readable Chemical Reactions File

PETER WILLETT

Postgraduate School of Librarianship and Information Science, University of Sheffield, Western Bank, Sheffield, S10 2TN, United Kingdom

Received April 24, 1979

An automatic system for the analysis and retrieval of chemical reaction information is described which allows searches to be carried out for both reacting and nonreacting substructures in the reactant and product molecules of a chemical reaction. The techniques could be implemented easily in a conventional substructure search system. Applications of chemical reaction files to the use of computer-aided synthesis programs are discussed.

I. INTRODUCTION

Two previous papers in this series have described automatic methods for the characterization of chemical reactions using Wiswesser Line Notations (WLN)[1] and connection tables[2] as the machine-readable structure representations; a comparison of the techniques[3] and an evaluation of a printed index of reactions produced by the WLN approach[4] have also been presented. In this paper, we describe an experimental system for the retrieval of reaction information based, primarily, upon the reaction sites identified by our structure-matching algorithm.[2] The retrieval system described here uses the analyses produced by the structure-matching procedure to allow searches to be made for both reacting and nonreacting substructures in the reactants and products; the WLN analyses allow easy access only to the former type of feature.[3]

II. CREATION OF THE SEARCH FILE

The source file for this work was the 7415 reactions successfully analyzed by the WLN reaction indexing program described earlier.[1] For each such reaction the WLNs of the reactant and product molecules were converted to CROSSBOW connection tables using software provided by ICI Ltd. (Pharmaceuticals Division), and then the tables were written out to tape together with the fragment strings resulting from the WLN analysis, the original WLNs, and the bibliographical reference. In this way, a file of 5226 one-reactant, one-product reactions was obtained for analysis by the structure-matching procedure; it should be noted that the procedure is extensible to more complex transformations by merging the sets of reactant or product structure representations so as to represent a single, discontinuous graph.

Many types of structure search system have been described in the literature.[5,6] We have used a simple sequential organization in which each of the items in the file is characterized by a fragment bitscreen; corresponding query bitstrings are held in core and matched against each of the reactions in the search file in turn. Boolean AND, OR, and NOT logic is available, together with a string search facility for the fragments resulting from the WLN analysis.

The multifarious nature of chemical reaction information and the need to differentiate between reacting and nonreacting substructural features require a variety of modes of access to the data, and we now outline the screening system which has been used to characterize each of the reactions in the file.

An analysis by Clews[7] showed that a file of reaction site residues contained a higher percentage of heteroatoms than the corresponding file of reacting compounds, and thus different sets of screens are required for assignment to the two types of structural feature. We have used atom, bond, ring, and molecular formula screens and as these are to be assigned to both reactant and product reaction sites and parent molecules, a total of 16 different types of screen are used for each reaction; in view of the small number of reactions in the search file, this variety of screen types is probably excessive but the retrieval results described below indicate that the screening

0095-2338/80/1620-0093$01.00/0

Table I. Creation of the Reactions Search File

reactions processed	4729
successful analyses	4388
overflow	8
no atoms matched	296
detected failures	37

system is applicable to very much larger files of data.

The reacting molecules were sorted into a WLN-ordered list and connection tables generated for each distinct compound; these tables were then used to produce the atom and bond screen sets which were to be assigned to the parent compounds. The corresponding reaction site screen sets were produced from a 1-in-3 interval sample of the data base; for each reaction in this subfile, redundant adjacency matrices were produced and then compared so as to isolate the reaction sites, and the atoms and bonds contained therein were used for set generation. Each of the four sets contained 240 members, including a conflated screen, and was produced by the automatic screen set generation procedure described by Willet.[8] An adaptation of this technique was used to produce 48-member screen sets describing the rings present in the reacting molecules and in the reaction sites.

The molecular formula screens for the reacting molecules were obtained from the numbers and types of atoms in these compounds while the reaction site screens described the difference in molecular formula between them; thus, in a reaction involving the formation of an oxime from a ketone, a bit would be set in the product reaction site screen to indicate the presence of one nitrogen atom. The use of molecular formula changes in characterizing reactions has been described by Nunn.[9]

A program has been written to process the file of 5226 reactions mentioned above. The first segment takes a CROSSBOW connection table and converts it to a redundant adjacency matrix; the process is repeated for both of the reacting molecules. This routine does not handle tables derived from bridged or peri-fused ring systems and was thus able to process only 90.5% of the reactions in the source file. In the second segment, the two matrices are compared using the structure-matching algorithm while the final section validates the analysis, assigns the various types of screen, and then writes the bitscreens, WLNs, WLN analysis fragments, and bibliographical details out to tape. The program contains about 1000 lines of Algol 68 code and has been run in 125K of core on the University of Sheffield ICL 1906S computer.

The processing of the source file is detailed in Table I. Of the 4729 reactions for which adjacency matrices could be produced, analyses were obtained for 4388 of the reactions, a net success rate of 92.8%. In eight of the reactions which were not processed, the integers describing circular substructures in the reacting molecules became too large for the computer words reserved for them, while in 296 cases no reactant–product mappings were obtained;[2] these reactions could have been included in the search file simply by marking the whole molecules as the reaction sites. The remaining 37 cases corresponded either to reactions in which all of the reactant or product atoms were eliminated or to reactions in which the difference in the numbers of reactant and product atoms in the reaction sites was not equal to the difference in the numbers of atoms in the reacting molecules; these simple checks encompassed many of the failures noted in the previous report.[2]

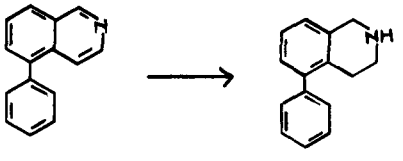
The software that has been developed to search the file of reactions is, in large part, identical with that used to characterize the analyses. In particular, bitscreens are assigned using a connection table as the primary input query representation. These tables are processed in the same way as the connection tables of the reacting molecules to produce integer strings which may then be matched against the appropriate screen set.[8] Molecular formula and ring requirements may be specified, and provision has also been made for string searches[17] of the fragments arising from the WLN analysis; although not currently implemented, it would be useful to extend this facility to the notations of the parent compounds. No atom-by-atom search facilities have been included though these might be required for very large files of reactions; in such a case, a note would need to be kept of the atoms included in the reaction sites.

A full description of the screening system, together with examples of query encoding techniques and a listing of the program used to create the search file, is given in the author's thesis.[10]

III. EVALUATION OF THE SEARCH SYSTEM

The analyses were tested with a set of 102 queries which was matched against the search file of 4388 reactions described above. Three main sources were used for the compilation of the set of queries. The first of these was 34 real enquiries supplied by the Research Information Department of Pfizer (UK) Ltd; the second group was the 18 questions used in the Derwent–WLN comparison described by Bawden et al.,[4] while the remaining 50 queries were culled from a variety of literature sources, all of which contain illustrative reaction types that should be easily searchable in a reaction's retrieval system.[11-15]

In toto, of 102 queries searched, 79 produced no output at all; the remainder gave rise to between one and 65 reactions, the median screenout for these queries being 99.8%. The determination of precision is somewhat difficult in that while an exact match may be obtained between a reaction and the query requirements, concurrent changes may have taken place elsewhere in the reacting molecules. Thus the reaction shown below might, or might not, be considered relevant to a request



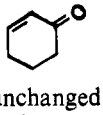
requiring the hydrogenation of carbon–carbon double bonds due to the simultaneous hydrogenation of an analogous carbon–nitrogen bond. We have taken such retrievals to be false drops; with the proviso that the figures err on the side of caution, 328 of the 399 reactions retrieved were regarded as relevant to their query with the median precision being 65.7%.

Only the two queries

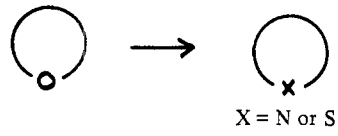
$$-NH_2 \rightarrow -H$$

$$=CH-CH_2-CH= \rightarrow -CH=CH-CH=$$

resulted in noticeably poor retrieval with 10 out of 11 and 7 out of 7 of the reactions, respectively, being false drops. More generally, the availability of a variety of screen types permitted a stringent definition of the query. Thus, consider a reaction in which the ring below, whether fused or not, was opened



while a lactone remained unchanged; specification of the ring carboxyl grouping in both the reactant and the product together with a WLN string search for the analysis fragment corresponding to the ruptured carbonyl ring was sufficient to eliminate the entire file. A request for the ring change produced four hits out of six by specifying a divalent ring oxygen atom in the reactant reaction site, ring nitrogen, or sulfur atoms in the product reaction site together with the appropriate



molecular formula change. Whereas substructure searching corresponds to the search for an inclusive bitstring match, searches for molecular formula changes in reactions may be either inclusive or exact. In cases where the query is only partially defined, only a search for an inclusive match is possible, but in other cases exact-match searching may lead to a marked increase in the precision of the search. Thus, a request for the hydrolysis of acetates produced 92 reactions if an inclusive molecular formula change was specified, as against 56, all of which were hits, for an exact search.

The distribution of retrieval set sizes is similar to that given by Adamson et al.[16] with the majority of the queries producing little or no material. Such queries often involved the formation of specific rings or the reaction of complex functionalities; most of the queries in this class, some of which are shown in Chart I, were from the group of real industrial questions. Conversely, some queries produced a large output; as noted earlier,[1] such changes are generally very simple in character (see Chart II).

The high screenout and precision figures show clearly the ability of the screening system to provide rapid and accurate access to the data for a large fraction of the substructural reaction queries that might be expected in an operational environment. The results may also be taken to show the effectiveness of the structure-matching algorithm in identifying the reaction sites within the pairs of molecules involved in a reaction. Moreover, the entire process of reaction site detection, screen set generation, screen assignment, and search is fully automatic with manual intervention required only at the query-encoding stage. Even here, significant savings of user effort have been achieved by the use of a connection table as the primary input query medium. The table is processed to produce atom, bond, and molecular formula screens without the need for the user to have any idea as to the contents of the various screen sets; the search system is hence ideally suited to on-line usage via an interactive graphics terminal. However, the reaction analyses are in no way dependent on the particular method used to search them, and any substructure search system based on connection tables could be used as long as screens were made available for searches of both reacting and nonreacting features. Since the structure-matching procedure is both simple in concept and efficient in operation, large reaction files could easily be implemented and searched in a conventional structure-handling system.

IV. USE OF REACTION FILES IN COMPUTER-AIDED SYNTHESIS

Computer-aided synthesis design programs are becoming widely available in both commercial and academic environments.[18,19] Such programs inspect an input target molecule for substructures which may be made by application of transforms available in the program's internal reaction library. Application of the transform gives rise to a precursor molecule which may then act as a target compound in its own right. The transform dictionaries contain reactions of proven synthetic value which may be applied in a wide variety of structural environments, but limitations on core storage and speed of execution mean that the size of the dictionary is quite small, typically about 300 entries.[20,22] Many types of reactions reported in the literature will not be available to a synthesis program because of their lack of generality even though they may well proceed in excellent yield under certain circumstances; thus while a program may suggest possible substructural changes which will bring about the synthesis of the target molecule, it is not unlikely that transforms other than those included in the reaction dictionary will carry out these changes more effectively. Thus the use of a transform might be assigned a low merit rating because of the presence of conflicting functionality elsewhere in the molecule; however, use of a more specific reagent not available to the program might well result in the reaction proceeding in good yield. The scope and power of a synthesis program could be dramatically increased if it could gain access to such reactions, and we now describe one way in which this could be achieved.

Chart I. Examples of Queries Which Retrieved No Reactions

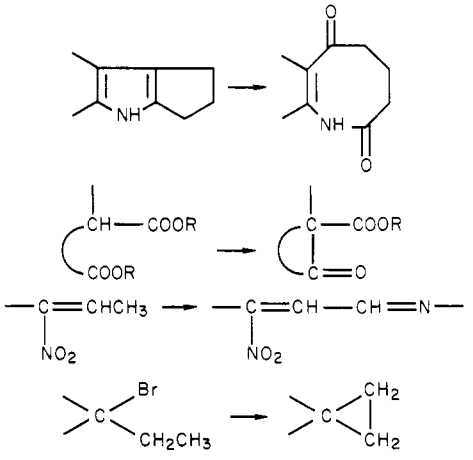
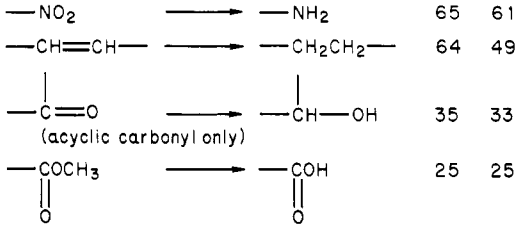


Chart II. Typical Queries Retrieving a Large Number of Reactions[a]



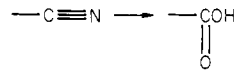
[a] The figures are the number of reactions retrieved and the number of hits.

A data base of chemical reactions is potentially very large and reaction indexing programs must accordingly be simple in concept and efficient in operation if economical processing rates are to be achieved; conversely, synthesis programs must perform highly sophisticated manipulations upon limited numbers of reactions. The proposed technique, which is analogous to suggestions made by Gund et al.,[23] uses both of these characteristics. The required target molecule is input to the synthesis program and potential syntheses are obtained in the normal manner using the internal reaction dictionary. These syntheses will contain one or more substructural transformations, and each such change is then used as a query for a search of a separate file containing as many reactions as are available. Since a given transformation may generally be brought about in several ways, it will be clear that very many more possible synthesis routes will become available, all of them involving the same series of substructural changes.

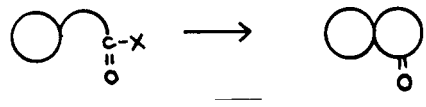
The method has been tested with the assistance of the Medicinal Research Centre, Beecham Pharmaceuticals, who both funded and provided facilities for carrying out a series of searches using the commercially available version of Wipke's on-line SECS[24] program; printouts were also provided which had been obtained for the synthesis of three molecules using Gelernter's SYNCHEM system.[22,25] A total of 52 distinct substructural changes were identified in the output from these two sources; these changes were encoded and searched as described above. It is a measure of the general level of sim-

plicity of the transformations currently available to synthesis programs that almost half of the queries produced some output, a much larger percentage than in the previous query set; in all, 25 queries gave rise to between 1 and 58 retrievals.

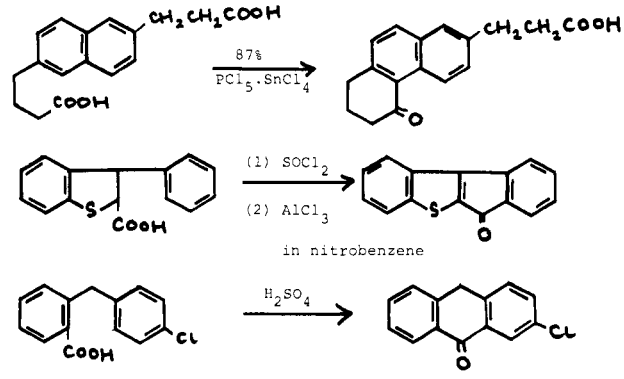
The file of reactions used was derived from issues 307 to 330 (October 1969 to March 1970) of *Current Abstracts of Chemistry* (CAC) and the output from the search program for a reaction satisfying the query requirements contained the WLNs of the reacting molecules together with a reference to the source abstract in CAC. Accordingly, for each query, the appropriate issues were consulted to identify the change in the reaction flow diagrams included with each abstract. Ideally, a search of the original journal articles should have been made, but many of these were not available so that we were limited to the reaction details given in the diagrams; this information was not always present. Thus a search for the ten reactions retrieved in answer to a request for the change



did not produce a single method for carrying out the reaction; another seven of the queries which produced relevant output yielded a similar lack of detail. Other cases, however, suggest that the procedure could be of some help in the design of practical synthetic pathways: thus a search for ring closures of the form (X = OH or halogen)



produced the three reactions



Ten different methods for the oxidation of secondary alcohols to the corresponding ketones were identified, these including the use of chromium trioxide in pyridine or acetic acid, manganese dioxide, and lead tetraacetate, as well as less familiar reagents, such as dicyclohexylcarbodiimide in dimethyl sulfoxide and 2,3-dichloro-5,6-dicyanobenzoquinone; additional methods were doubtless included in the many abstracts which did not give details of the reaction conditions.

It seems likely that this two-step approach could yield several alternatives for any substructural change that may be required in a synthesis if a sufficiently large data base were to become generally available. However, direct data capture from some printed source, such as CAC, would inevitably lead to the same change being recorded many times over so that some form of registration procedure would need to be adopted.

V. CONCLUSIONS

The evaluation has been carried out in two ways. In the first part, a set of 102 substructural reaction queries were searched against a file of 4388 analyses produced by our reaction site detection procedure; the median screenout and precision for the 23 queries that retrieved some material were 99.8 and 65.7%, respectively. Although the number of reactions in the file is quite small, the results suggest that the analyses provide sufficiently accurate characterizations to permit a wide range of reaction queries to be searched with a reasonable degree of retrieval effectiveness. Secondly, searches for reaction transformations obtained from two computer-aided synthesis design programs have been used to illustrate the ability of files of reactions to substantially increase the number of reaction types available to such programs.

ACKNOWLEDGMENT

Thanks are due to Michael Allen and Michael Lynch for helpful discussions, Pfizer (UK) Ltd. for the set of user queries, ICI Ltd. (Pharmaceuticals Division) for software, and the Department of Education and Science for the award of a British Library Postdoctoral Research Fellowship. Funding and facilities for the SECS searches were provided by Beecham Pharmaceuticals and this support is gratefully acknowledged.

REFERENCES AND NOTES

(1) M. F. Lynch and P. Willett, "The Production of Machine Readable Descriptions of Chemical Reactions Using Wiswesser Line Notations", *J. Chem. Inf. Comput. Sci.*, **18**, 149–154 (1978).
(2) M. F. Lynch and P. Willett, "The Automatic Detection of Chemical Reaction Sites", *J. Chem. Inf. Comput. Sci.*, **18**, 154–159 (1978).
(3) P. Willett, "Computer Techniques for the Indexing of Chemical Reaction Information", *J. Chem. Inf. Comput. Sci.*, **19**, 156–158 (1979).
(4) D. Bawden, T. K. Devon, F. T. Jackson, S. I. Wood, M. F. Lynch, and P. Willett, "A Qualitative Comparison of Wiswesser Line Notation Descriptors of Reactions and the Derwent Chemical Reactions Documentation Service", *J. Chem. Inf. Comput. Sci.*, **19**, 90–93 (1979).
(5) J. E. Ash and E. Hyde, Eds., "Chemical Information System", Ellis-Horwood, Chichester, 1975.
(6) W. T. Wipke, S. R. Heller, R. J. Feldmann, and E. Hyde, Eds., "Computer Representation and Manipulation of Chemical Information", Wiley, New York, 1974.
(7) L. A. Clews, "Characterization of a Reaction Data Base as an Aid to the Evaluation of a Chemical Reaction Retrieval System", M.Sc. Thesis, University of Sheffield, 1973.
(8) P. Willett, "A Screen Set Generation Algorithm", *J. Chem. Inf. Comput. Sci.*, **19**, 159–162 (1979).
(9) P. R. Nunn, "The Automatic Analysis of Chemical Reactions", M.Sc. Thesis, University of Sheffield, 1974.
(10) P. Willett, "Computer Analysis of Chemical Reaction Information for Storage and Retrieval, Ph.D. Thesis, University of Sheffield, 1978.
(11) J. Valls and O. Schier, "Chemical Reaction Indexing", ref 5, pp 243–258.
(12) J. Valls, "Reaction Documentation", ref 6, pp 83–104.
(13) G. E. Vleduts, "Concerning One System of Classification and Codification of Organic Reactions", *Inf. Storage Retr.*, **1** (2/3), 117–146 (1963).
(14) D. R. Eakin and W. A. Warr, "Computerized Aids to Organic Synthesis in a Pharmaceutical Research Company", ref 18, pp 217–226.
(15) E. J. Corey, R. D. Cramer, and W. J. Howe, "Computer-Assisted Synthetic Analysis for Complex Molecules. Methods and Procedures for Machine Generation of Synthetic Intermediates", *J. Am. Chem. Soc.*, **94**, 440–459 (1972).
(16) G. W. Adamson, J. A. Bush, A. H. W. McLure, and M. F. Lynch, "An Evaluation of a Substructure Search Screen System Based on Bond-Centered Fragments", *J. Chem. Doc.*, **14**, 44–48 (1974).
(17) J. E. Crowe, P. Leggate, B. N. Rossiter, and J. F. B. Rowland, "The Searching of Wiswesser Line Notations by Means of a Character-Matching Serial Search", *J. Chem. Doc.*, **13**, 85–92 (1973).
(18) W. T. Wipke and W. J. Howe, Eds., "Computer-Assisted Organic Synthesis", Symp. Ser., No. 61, American Chemical Society, Washington, D.C., 1977.
(19) M. Bersohn and A. Esack, "Computers and Organic Synthesis", *Chem. Rev.*, **76**, 269–282 (1976).
(20) H. W. Orf, "Computer-Assisted Synthetic Analysis", Ph.D. Thesis, Harvard University, 1976.
(21) M. Bersohn and A. Esack, "A Computer Representation of Synthetic Organic Reactions", *Comput. Chem.*, **1**, 103–107 (1976).
(22) R. H. Boivie, "Heuristic Search Guidance and Subgoal Analysis in the SYNCHEM2 Organic Synthesis Discovery Program", Ph. D. Thesis, State University of New York at Stoney Brook, 1977.
(23) P. Gund, J. D. Andose, and J. B. Rhodes, "Computer-Assisted Synthetic Analysis in Drug Research", ref 18, pp 179–187.
(24) W. T. Wipke, H. Braun, G. Smith, F. Choplin, and W. Sieber, "SECS-Simulation and Evaluation of Chemical Synthesis: Strategy and Planning, ref 18, pp 97–127.
(25) H. L. Gelernter, A. F. Sanders, D. L. Larsen, K. K. Agarwal, R. H. Boivie, G. A. Spritzer, and J. E. Searleman, "Empirical Explorations of SYNCHEM", *Science*, **197**, 1041–1049 (1978).