Alpine fault of New Zealand, with generally right-hand strike slip, appears to have a left-hand deflection (*23*). These and other indications of contemporary tectonic change are discussed elsewhere (*26*).

Summary

Earthquake origins are distributed in three dimensions, but ordinary mapping displays only two. Lines of epicenters often trend transversely with respect to the larger surface structures. This is true not only of deep-focus earthquakes but also of ordinary shallow shocks.

Such transverse alignments are often associated with transversely trending boundaries of active areas, defined by all known events during an interval, or by the aftershocks of a given large earthquake.

Both types of transverse trends are presumably due to fractures cutting across the principal structures. In many areas they parallel alternative trends evident elsewhere in the general region, which is then characterized by two principal trends crossing at a high angle. Such conjugate trends are often taken as expressing alternative directions of shearing under approximately uniform regional stresses. As such, they need not conform closely to present strains and tectonic dislocations; rather, they represent established zones of weakness of considerable geologic antiquity. Such zones of weakness may have originated under conditions differing widely from the present ones, but they will still determine the lines along which current dislocations are taking place.

References and Notes

1. E. Tams, *Z. Geophys.* **3**, 361 (1927).
2. B. Gutenberg and C. F. Richter, *Geol. Soc. Amer. Spec. Pap. 34* (1941); J. P. Rothé, *Ann. Geofis.* **4**, 27 (1951).
3. J. P. Rothé, *Ann. Int. Geophys. Yr.* **30**, 9 (1965).
4. A. Sugimura and T. Matsuda, *Geol. Soc. Amer. Bull.* **76**, 509 (1965).
5. A. O. Woodford, J. E. Schoellhamer, J. G. Vedder, R. F. Yerkes, in *Calif. Div. Mines Bull. 170* (1965), vol. 1, chap. 2, p. 65.
6. W. E. Pratt, *Seismol. Soc. Amer. Bull.* **16**, 146 (1926).
7. E. F. Savarensky, I. E. Gubin, D. A. Kharin, Eds., *Zemletryaseniya v SSSR* (Academy of Sciences, Moscow, 1961).
8. T. Matuzawa, *Study of Earthquakes* (Uno Shoten, Tokyo, 1964).
9. C. F. Richter, *Elementary Seismology* (Freeman, San Francisco, 1958).
10. *Int. Seismol. Sum.* (quarterly publication of the Association of Seismology, International Geodetic and Geophysical Union; it gives earthquake data and epicentral locations in chronological order).
11. J. N. Jordan, J. F. Lander, R. A. Black, *Science* **148**, 1323 (1965); see also W. Stauder, *J. Geophys. Res.* **73**, 3847 (1968).
12. C. F. Richter, in *Calif. Div. Mines Bull. 171* (1955), pp. 177–197.
13. E. F. Savarensky, S. L. Solov'ev, D. A. Kharin, Eds., *Atlas zemletryaseniy v SSSR* (Academy of Sciences, Moscow, 1962).
14. J. T. Wilson and D. J. O'Halloran, *Geol. Soc. Amer. Bull.* **69**, 1710 (1958); compare seismic risk map for the conterminous United States, published by the Environmental Science Services Administration and the Coast and Geodetic Survey (1969).
15. H. O. Wood, *Seismol. Soc. Amer. Bull.* **6**, 55 (1916).
16. A. Ryall, D. B. Slemmons, L. D. Gedney, *ibid.* **56**, 1105 (1966).
17. C. F. Richter and J. M. Nordquist, *ibid.* **41**, 347 (1951).
18. J. N. Brune, W. Arabasz, G. R. Engen, *Seismol. Soc. Amer. Bull.*, in press.
19. S. W. Smith and J. M. Nordquist, personal communications.
20. C. F. Richter, *Geol. Soc. Amer. Bull.* **80**, 1363 (1969).
21. E. W. Hart [in *Proceedings, Conference on Geologic Problems of San Andreas Fault System*, W. R. Dickinson and A. Grantz, Eds. (Stanford Univ. Press, Stanford, Calif., 1968), p. 258] has given evidence for geologically younger displacements on the Nacimiento and associated faults; some of his results are quoted by Richter (*20*).
22. H. W. Wellman and R. W. Willett, *Trans. Roy. Soc. N.Z.* **71**, 282 (1942).
23. R. P. Suggate, *Trans. Roy. Soc. N.Z. Geol.* **2**, 105 (1963).
24. M. L. Hill and T. W. Dibblee, Jr., *Geol. Soc. Amer. Bull.* **64**, 443 (1953).
25. S. Kaneko, *Kagaku Asahi (Tokyo)* **1968**, 89 (July 1968).
26. C. F. Richter, *Eos (Trans. Amer. Geophys. Union)* **50**, 318 (1969).
27. This article is contribution No. 1598 of The Division of Geological Sciences, California Institute of Technology, Pasadena.

Computer-Assisted Design of Complex Organic Syntheses

Pathways for molecular synthesis can be devised with a computer and equipment for graphical communication.

E. J. Corey and W. Todd Wipke

Introduction

This article is concerned with the general theory of chemical synthesis and with the application of machine computation to the generation of chemical pathways for the synthesis of complicated organic molecules. The basis for the approach which has been developed comes in large measure from the methods used by chemists in the solution of certain types of synthetic problems. It is appropriate, therefore, to begin with a brief description of the general processes by which chemical syntheses of organic molecules are devised.

The number of discrete organic chemical compounds which are capable of existence as stable entities can be described conservatively as astronomical. The simple formula $C_{40}H_{82}$, which is "saturated" by univalent hydrogen so that only chains of atoms are possible, has been calculated to permit the joining of carbon atoms in 63,491,178,805,831 different ways with regard to topology in two dimensions (*1*). The number of possible and distinctly different molecules of formula $C_{40}H_{82}$ is actually far greater because of the three-dimensional (stereochemical) characteristics of organic structures. For instance, the fact that four single bonds to a carbon are normally tetrahedrally directed in space allows any four unlike groups to be attached to that atom in two different ways. The profusion of organic structures becomes still more impressive when consideration is given to three additional molecular characteristics: first, that organic molecules can contain many thousands of atoms; second, that a large number of the known elements can bond to carbon and to each other in a molecule; and third, that cyclic connections within molecules can lead to a prodigious variety of rings or networks of atoms.

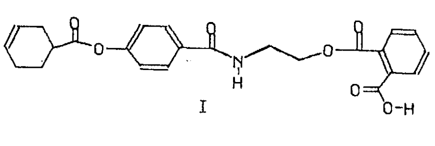
A large majority of the millions of

Dr. Corey is Sheldon Emery Professor of Chemistry at Harvard University, Cambridge, Massachusetts. Dr. Wipke was a postdoctoral research fellow at Harvard University and his present address is Department of Chemistry, Princeton University, Princeton, New Jersey.

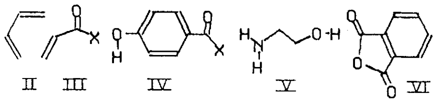
organic structures that are recorded in the scientific literature have been created by chemical synthesis. In many instances a given organic compound may have been synthesized in several different ways. This is even true of "small" molecules such as acetone, CH_3COCH_3, or phenylacetic acid, $C_6H_5CH_2COOH$, although the number of different syntheses leading to a given structure can be expected to increase with molecular size. To effect such syntheses, chemists have at their disposal a variety of basic chemicals derived from fundamental source materials such as petroleum, air, water, salt, and sulfur and, additionally, many thousands of well-known "secondary" compounds obtainable either commercially or by well-described laboratory procedures which can serve as "readily-available" starting materials. They also possess a very large arsenal of chemical reactions, numbering certainly in the thousands, for changing one type of structure into another. For each type of reactive unit in an organic molecule, usually termed a functional group, a wide variety of reactions are known which are useful in chemical synthesis. Such reactions can induce a diverse array of structural modifications including: (i) interconversion, removal, or introduction of functional groups, (ii) extension of the atomic chains or appendages, (iii) generation of atomic rings, (iv) rearrangement of chain or ring members, and (v) cleavages of chains or rings. Finally, synthetic chemists make use of extensive new knowledge regarding the electronic and quantum-mechanical aspects of molecular change.

Unlike macro-construction, the synthesis of a molecule from constituent units does not and cannot depend on the mechanical placement of these units in an appropriate arrangement. The atomic groups to be joined must find each other and bind together in the proper way automatically, without external intervention or delivery. Thus it is not sufficient that the constituent synthetic units required for a synthesis be available; they must also possess chemical properties or affinities which are predictable and which are so specific as to allow selective combination in only one of the many possible modes. It should be noted that the effective application of any synthetic plan may require the use of certain control techniques, for example, involving activating, deactivating, or directing groups, to maximize the specificity of chemical combination. The synthesis of a molecule of any complexity usually requires the utilization of a sequence of carefully chosen chemical reactions in a fairly rigid order. In general, the more complex the molecule, the larger is the number of chemical steps required for synthesis. Clearly, if any one of these stages of synthesis cannot be induced—or modified—to yield the required result in practice, the entire synthetic plan will founder.

How does a chemist choose a pathway for the synthesis of a large organic molecule, given either the great diversity of organic structures and reactions or, in contrast, the critical importance of each step to ultimate success? Surprisingly, this question has not been dealt with in a general and systematic way in chemical textbooks, and only recently has a relevant analysis appeared in the chemical literature (*2*). Not surprisingly, chemists employ a variety of problem-solving techniques of varying sophistication, depending both on the chemist and the problem. It is sufficient for our purposes to distinguish between two extreme approaches and one which is intermediate. The extreme methodologies which differ with respect to analytical and logical sophistication and generality can be termed "direct associative" and "logic-centered" molecular synthesis. The direct associative approach can be used to generate a possible synthetic scheme when the chemist directly recognizes within a structure a number of readily available, undisguised subunits which can be brought together in the proper way using standard reactions with which he is very familiar. For example, the molecule represented by formula I can be assembled by joining the components II to VI in the indicated order. Most synthetic chemists would arrive at this scheme with a minimum of logical analysis or planning simply because of the fact that the subunits II to VI are so obvious and so familiar, like the reactions required to join them, that the simple processes of mnemonic association lead directly to a possible solution. The applicability of the associative approach is limited to relatively simple synthetic problems even if it is broadened to include extensive search and association processes involving the vast chemical literature as well as control techniques.

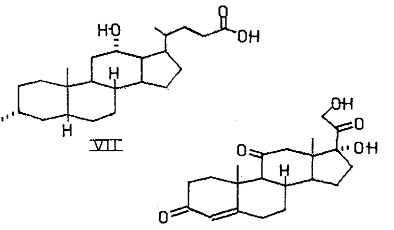


I



II III IV V VI

The intermediate approach to synthetic analysis has been the basis for the great majority of the previously accomplished syntheses of known organic compounds. In essence, the approach involves the recognition of a relation between a critical, and perhaps major, unit in the structure to be synthesized and a structure which corresponds to a known or potentially available chemical substance. This substance then becomes a starting point for the synthesis, and the problem is reduced to finding a set of reactions which will convert each of the starting structures to the desired target structure. The derivation of a sequence for this interconversion may require fairly complex analysis of the logic-centered type (as discussed below); however, it will always be dominated by a set of goals which are determined by the nature of the particular starting and end points of the synthesis. The choice of a particular starting point channels and simplifies the analysis. At the same time, the imposition of a starting point for a synthesis limits the scope and rigor with which a problem can be analyzed, with the result that one or more superior solutions may remain undiscovered. The first synthesis of cortisone (VIII) is illustrative of a synthetic plan which was derived with the use of the "intermediate" approach. The synthesis was accomplished with a readily available natural product as starting material, deoxycholic acid VII, a compound having the same carbon network both with regard to gross structure and stereochemistry (*3*). However, even starting with deoxycholic acid, itself the product of a complicated biosynthesis, over 30 steps were required to effect the synthesis. Despite the general similarities between structures VII and VIII, this particular synthesis is both less efficient and less elegant than other syntheses starting with simple basic chemicals available in unlimited quantity.



VII VIII

On the other hand, most syntheses

of polypeptides start with the identification of the repeating α-amino acid units and are planned around these as starting materials. This also happens to be the most efficient approach to synthesis because of the availability of the α-amino acid units in the proper stereochemical form, because of the structural redundancy of a polypeptide molecule, and because of the relative operational ease of forming the CO–N bonds which join the α-amino acid units. It might be argued that although polypeptides certainly are complex structures, in terms of synthetic analysis they are sufficiently uncomplicated so that the intermediate form of synthetic analysis is sufficient.

In the subject of logic-centered complex molecular synthesis, at the other end of the spectrum, we encounter a methodology limited only by the frontiers of chemistry and the power of human intelligence and creativity. Central to this methodology is the requirement of a penetrating and rational analysis of the molecular structure which is the synthetic target. This analysis leads to a logically restricted set of structures which may be converted in a single synthetic operation, that is, chemical step, to the synthetic target. Each of these structures is in turn carefully analyzed in detail as was the original target structure, and from each a set of synthetically more accessible structures is derived, again one synthetic operation removed. This process is carried out repeatedly for each synthetic intermediate until a "tree" of such intermediate structures results, which leads down from the synthetic target to structures corresponding to readily available starting materials. Thus a set of possible synthetic pathways corresponding to sequences of intermediate structures is generated. The derivation of these synthetic pathways is strictly carried out in an order which is opposite to the direction in which a synthesis is conducted in the laboratory, and in that sense the analysis is performed backward relative to the execution. Figure 1 shows a small portion of a synthetic tree originating with the synthetic target *T*. In contrast to the "direct associative" approach and the "intermediate" approach, the analysis is in no way encumbered or diverted by assumptions as to starting materials, which serve only to signal the end of a particular analytical sequence. The various synthetic pathways resulting from the generation of a tree of intermediates require further analysis in order to evaluate relative merit. This process demands analysis of each synthetic step of each sequence, which must make maximum use of available chemical information, and, ultimately, experimental substantiation by reduction to practice.

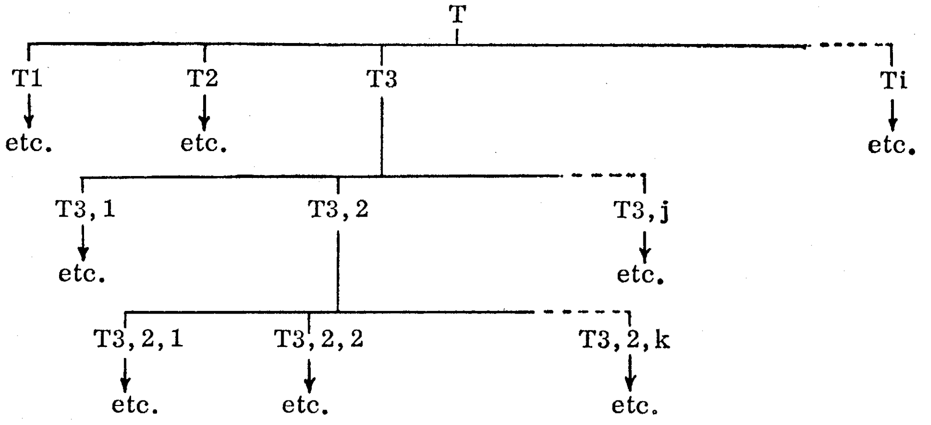
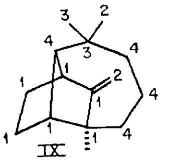
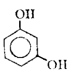


Fig. 1. Synthetic analysis of target *T* generates a "tree" of intermediate precursor structures.

Not uncommonly, a synthetic route so derived and demonstrated by experiment will be sufficiently subtle that the pathway of synthesis will be far from obvious, given the starting materials actually used. For example, the naturally occurring substance longifolene (IX) has been synthesized after a detailed analysis starting from resorcinol (X), methyl iodide (XI), ethyl bromide (XII), and methyl vinyl ketone (XIII), the origin of each carbon atom of synthetic IX being as indicated (*4*).



IX

1, from  (X)

2. from CH_3I (XI)

3, from CH_3CH_2Br (XII)

4, from $CH_3COCH{=}CH_2$ (XIII)

Relatively few of the syntheses of organic molecules recorded in the literature have been based on the logical generation of fairly complete synthetic trees, partly because of the fact that this approach has not previously been clearly and generally defined. Although rigorous analysis of a complex synthetic problem is extremely demanding in terms of time and effort as well as chemical sophistication, it has become increasingly clear that such analysis produces superlative returns. Perhaps the greatest advance in chemical synthesis in coming years will be the repeated and ever more convincing demonstration of this point.

Reference has been made above to the construction of a "synthetic tree" by the stepwise generation of synthetic intermediates, starting from the target structure. It is now appropriate to focus on the details which are central to this process. The analysis starts with the perception of those structural features within the target molecule which are of synthetic significance, including especially the following: (i) individual molecular chains, rings, and appendages, (ii) individual functional groups, (iii) asymmetric centers and groups attached thereto, and (iv) chemical reactivity, sensitivity, or instability at each point within the molecule. The next stage receives direction and selectivity from the all-important goal of reducing molecular complexity, which in a general sense can be considered to involve combinations of the following: (i) simplification of internal connectivity by scission of rings, (ii) reduction of molecular size by disconnection of chains or appendages, (iii) removal of functionality, (iv) modification or removal of sites of unusually high chemical reactivity or instability, (v) simplification of stereochemistry, for example, by removal of asymmetric centers. There are also important subgoals which can direct the analysis toward the above-listed goals without themselves corresponding to molecular simplification. These include (i) interchange of types of functional groups, (ii) introduction of functional groups, (iii) modification of functionality to allow control of levels of chemical reactivity, (iv) introduction of groups which permit stereochemical or positional control, and (v) internal rearrangement, for example, to modify rings, chains, or functional groups.

The most powerful technique for establishing a link between molecular features as perceived and operations which simplify molecular structure makes use of the description of organic reactions in terms of basic electronic reaction mechanisms. Such reaction mechanisms are now known in considerable detail for most of the important synthetic reactions. This knowledge permits a very compact and systematic classification of the hundreds of synthetic reactions and allows, among other things, (i) the correlation of the variation of reactivity with changes in chemical structure, (ii) the consideration of unstable reaction intermediates which intervene between reactants and products, and (iii) the prediction of stereochemical relations between reactants and products. It can be applied to the prediction of what kind of reaction, if any, will occur with a given set of reactants and reaction conditions, or it can be used as a guide to the selection of suitable reagents and reaction conditions to effect a given structural change. Because the electronic properties of the various functional groups are known, these groups can be dealt with systematically and efficiently with regard to their effect on chemical reactions. The fact that the mechanistic pathway for a chemical reaction is the same for forward and reverse directions (principle of microscopic reversibility) ensures that there are no obstacles to the use of mechanisms corresponding to the reverse of synthetic reactions to generate the intermediates of the synthetic tree (*5*).

Thus the perception of the relations between key structural features leads to the selection of an operable mechanism, the logical application of which produces a new structure or set of structures. If the application of this mechanism in the synthetic direction starting from the structure or structures so derived unambiguously re-forms the target structure, then a legitimate component of the synthetic tree has been found. The test of a reaction mechanism in the synthetic direction must be carried out, since the operation of a mechanism on an organic structure can and often does lead to several possible structures.

There are general and powerful principles of stereochemistry which provide goals and processes for synthetic analysis. Therefore, the subject of stereochemistry is linked both to structural features and structural operations based on reaction mechanisms. Further, goals and operations can be derived from topological considerations, especially in molecular structures of high connectivity or those possessing elements of symmetry or redundancy.

The processes of perception, as used by a chemist to derive a synthetic tree or sequence, can be differentiated with regard to subtlety and sophistication. The most complex and subtle levels are, of course, the least well understood and the most interesting for detailed study. They are also of key importance in the simplification of synthetic problems in the sense they can generate higher level strategies of great effectiveness. In some cases, these strategies derive from the recognition of the critical obstacles to the synthesis of a structure.

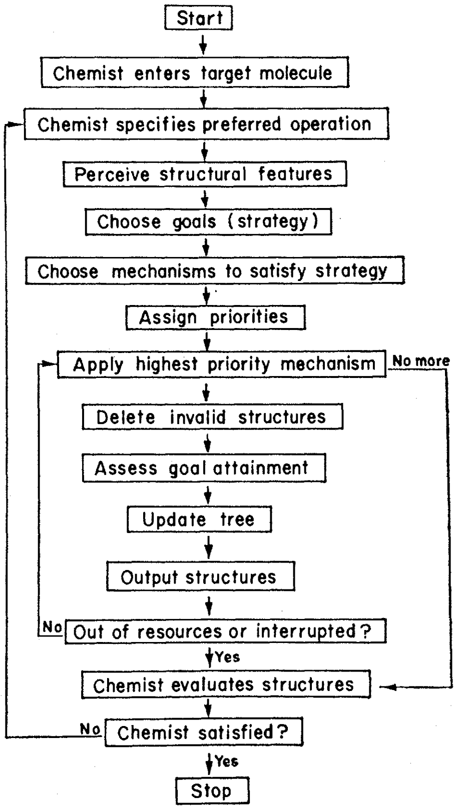

Start
Chemist enters target molecule
Chemist specifies preferred operation
Perceive structural features
Choose goals (strategy)
Choose mechanisms to satisfy strategy
Assign priorities
Apply highest priority mechanism
No more
Delete invalid structures
Assess goal attainment
Update tree
Output structures
No
Out of resources or interrupted?
Yes
Chemist evaluates structures
No
Chemist satisfied?
Yes
Stop

Fig. 2. Processing scheme for machine-assisted logic-centered synthetic analysis.

As a synthetic tree is developed, findings may emerge which suggest new relationships or simplifications that can lead to even more effective analysis of the original target structure. In this sense, it can be said that embedded in each problem is a learning opportunity, which if seized can lead to extraordinarily simple or elegant solutions. Learning opportunities are equally great in the subsequent stages of synthetic work which involve (i) the selection of specific chemical reactions for transforming one synthetic intermediate to the next in the sequence, (ii) the selection of specific reagents, (iii) the design of experiments, and (iv) experimental execution and analysis. During these stages observations, discoveries, inventions, or theories of great importance may result. No one is more keenly aware than the synthetic chemist that much of chemistry remains to be discovered or applied.

Requirements for Machine-Assisted Synthetic Analysis

Since the process of synthetic analysis can now be defined with much greater clarity, precision, and generality than heretofore possible, the question arises as to whether one can apply modern computers to the solution of problems of molecular syntheses. That there is a need for such an application is made apparent by the fact that a complete, logic-centered synthetic analysis of a complex organic structure often requires so much time, even of the most skilled chemist, as to endanger or remove the feasibility of this approach. It would be a great advantage if at least a part of the necessary analysis could be performed rapidly and accurately by computer. This was one consideration which led to the studies described in this article. Another was the conviction that it is important to test our understanding of synthetic analysis by writing and evaluating machine programs. Such programs, if effective, would constitute a starting point and guide for further elaboration of synthetic analysis and lead, ultimately, to a useful new computer-assisted methodology.

The following general requirements for the computer system were envisaged at the outset: (i) that it be an "interactive" system (*6*) allowing facile graphical communication of both input and output in a form most convenient and natural for the chemist, (ii) that it be capable of rapid processing of molecular structures to generate a legitimate tree, (iii) that the tree be limited in size so as not to include useless or chemically naive structures (that is, "forbidden" regions of the search space), but that it necessarily include as many useful pathways as possible, (iv) that the processing be performed automatically, but with provision of interruption by the chemist at any time to modify or redirect the analysis at any stage, (v) that the depth of search or analysis be decided by the chemist, and (vi) that the evaluation of the various pathways in the synthetic tree be done by the chemist, but that the machine order the output structures in a way tantamount to preliminary evaluation. These requirements limit the task to be performed by computer to the "logic-centered" part of synthesis and leave to the chemist the complex, ill-defined, and "information-centered" part, which is at present beyond the scope of computation. Also left to the chemist is the possibility of entering any learned information or higher strategy resulting from the analysis and, of course, the opportunity to contribute creatively to the solution of the problem. On the other hand, the system should allow considerable opportunity for "internal learning" by the machine.

Process for Synthetic Analysis by Computer

The system used in this study consisted of a Digital Equipment Corporation (DEC) PDP-1 computer and peripheral equipment for graphical communication, including a Rand tablet (*7*) and pen for input, and three DEC cathode-ray tubes (CRT) and a Calcomp plotter for output display. Because of the low cost of PDP-1 time, it is feasible to use this computing facility "interactively" for periods of time (typically 1 hour) which are both adequate and convenient. The program for generating synthetic schemes used with this system, designated as OCSS, involves a sequence of analysis which runs parallel to that used by the chemist in the logic-centered approach. The flow of processing in the program OCSS is shown in Fig. 2.

First the chemist must transmit to the program the target molecule. He then may specify various parts of the molecule or types of disconnections to be considered by the computer or let the program determine the analysis automatically. The program then takes control to begin generating precursors of the target molecule. But first it must perceive all synthetically significant features of the target molecule, for example, rings, ring junctures, functional groups, relations between functional groups, and symmetry. Based on these perceptions, a strategy is developed which includes the setting up of possible simplifications or boundary conditions and the generation of goals (*8*). A priority is assigned to the various goals, and the program then calls for the application of those available "inverse synthetic" operations which are likely to satisfy one or more of the goals. The application of each chosen operation becomes a subgoal which has a priority related to the priorities of the goals for which it was chosen and its relative chance of satisfying the goals.

The highest priority operation is now applied to generate one or more precursors which are checked for valence violations and other structurally unlikely situations. Those structures found to be invalid are deleted from further consideration. The remaining structures are then assessed with respect to how well they satisfy the goal for which they were chosen and the relative simplicity of the structure. The structures are then transmitted to the chemist as part of the updated synthetic tree.

The chemical transformations are applied in the order of their priority until there are no more left, until the program's resources (time allocated by the chemist and available memory) are expended, or until the chemist interrupts the program. Control is then given to the chemist to evaluate the newly generated precursors. He may delete or modify a precursor and may specify which structure in the tree should be examined next. If the chemist is satisfied that one or more intermediates in the tree are "readily available," then he terminates work on this problem and takes away a permanent record of the structures and synthesis tree. If not, then the whole sequence begins again with a selected precursor as the new target molecule, which of course need not be entered again. The program thus stops when the chemist is satisfied and not necessarily when the first solution is found.

Program Modules

A. Graphics. The program is organized into five functional modules which are shown in Fig. 3 in their relation to each other. The first of these modules, graphical communication, provides the interface between chemist and program. A major concern in designing the system was that it be convenient for use by a chemist without his having to learn a code. Couper, in 1858 (*9*), introduced the structural diagram, a convenient two-dimensional graphical language for representing molecules and reactions. (For examples of this language in its present form, see formulas I to VIII.) As this language is used internationally and is familiar to all chemists, it is the ideal language for communication between chemist and computer. Consequently, OCSS was developed to communicate in this language with the use of currently available hardware and software techniques. All input to the program comes from a Rand tablet (*7*) and pen, a device by which the computer can sense pen position on the tablet and sense whether the pen is being pressed down or not. The tablet is used for drawing in target molecules, for making modifications to structures, and for selecting options by pointing to a particular option from a "menu."

A CRT display is used in conjunction with the tablet, for, unlike pencil on paper, the Rand pen leaves no trace on the tablet. Instead, the program creates the trace on the CRT display. Thus the chemist draws on the tablet but observes his drawing on the scope. A small cross on the scope tells him where the pen is on the tablet. He can then point to objects on the scope by moving the pen until the cross is on the object and then press the pen down. These objects may correspond to parts of a drawing, or control words which initiate action by the program. From the standpoint of both the chemist and the computer programmer, the Rand tablet is easier and more natural to use than the conventional light pen.

For output, the program uses three DEC type-340 display units and a Calcomp plotter. The three scopes are used for display of a structure during input and modification, the synthesis tree, and any structure selected from the tree. Figure 4 (left) shows the three scopes during the processing phase. Each of the scopes displays control words and other buttons to allow the chemist to control graphically the operations of OCSS. On scope 1 buttons pertaining to structural input appear. Scope 2 displays the synthesis tree. Each node of the tree is a button which causes the structure for that node to appear on scope 3. Also on scope 2 are the buttons for requesting hard-copy output of either a structure or the synthesis tree (the structure index). Scope 3 has the buttons for specifying parameters and chemical transformations to the program. The Calcomp plotter is used to generate a permanent copy of both the structures and the synthesis tree (Fig. 4, right). To make efficient use of the plotter, which is a relatively slow output device, the program interleaves computation with plotting. In this way the plotter may run continuously, while the computer is generating more precursors and the chemist is evaluating them. Normally, the plotter is able to keep up with precursor generation, and by the time the session ends, all structures are in hard-copy form.

At this point it is appropriate to describe the structural diagram notation system as it is used by OCSS. The vocabulary consists of atoms, C, H, N, O, S, P, and X (halogen); charges, + and −; radical, ↑; and bonds, single, double, triple, and dotted (stereochemical configuration). In this notation system, the molecule is represented as a graph, with the nodes being atoms and the branches being bonds. Unlabeled nodes are assumed to be carbon atoms, and atoms other than carbon must always be labeled with the symbol of that atom type. Hydrogen atoms are not represented except when necessary for denoting stereochemistry or when attached to a hetero atom such as oxygen (O), nitrogen (N), or sulfur (S). Multiple bonds are represented by the appropriate number of parallel lines connecting the two atoms. Atoms having a formal charge are shown with a + or − sign to the right of that atom. Similarly, the presence of an unpaired electron, a radical, is shown by an up arrow (↑) to the right of that atom. Stereochemistry is shown by designating one bond as "down" and having that bond appear dotted. The other bonds attached to the asymmetric atom are drawn as they would actually appear if the asymmetric atom were oriented with the "down" bond perpendicular to and below the plane of the writing surface. The geometrical arrangement of substituents about a double bond (*cis-trans* isomerism) is indicated by drawing the bond with the substituent properly oriented.

On scope 1 are displayed control words applicable to the drawing process including DRAW, MOVE, DELETE, and ERASE. To begin drawing a structural diagram, one points to the control word DRAW (Fig. 5). The word DRAW becomes brighter, indicating that the program is in the drawing state. The pen is then placed inside the box at the point at which a bond is to begin, pressed down, and, while in the down position, moved to the point at which the bond is to end. A straight line appears connecting the initial point to the current pen position, giving the effect of a "rubber band bond." When the pen is lifted, the line becomes a permanent bond between two atoms. If the start or end points do not correspond to existing atoms, then new carbon atoms are created. Multiple bonds are made by drawing more bonds on top of a single bond. Addition of another bond to a triple bond results in a single bond. Bonds may cross without introducing a spurious atom at the intersection. OCSS analyzes the diagram for well-formedness as it is drawn; if the valence of an atom is exceeded, the message VALENCE EXCEEDED appears, and the offending bond is erased.

To specify charges or atom types other than carbon, one points to the correct atom type and then to the atom in the drawing. Bonds may be designated as "down" by pointing to the word DOWN and then to the bond. The bond then appears dotted. An atom or bond may be erased by pointing to the word DELETE and then pointing to the bond or atom. If an atom is designated, the atom and all bonds connected to it are deleted. If a bond is designated, only the bond is deleted, unless it leaves an isolated atom, in which case that atom is deleted. A drawing may be straightened up or the orientation of the two-dimensional drawing changed by moving the atoms one at a time by use of the MOVE control. The order ERASE erases the entire molecule. When

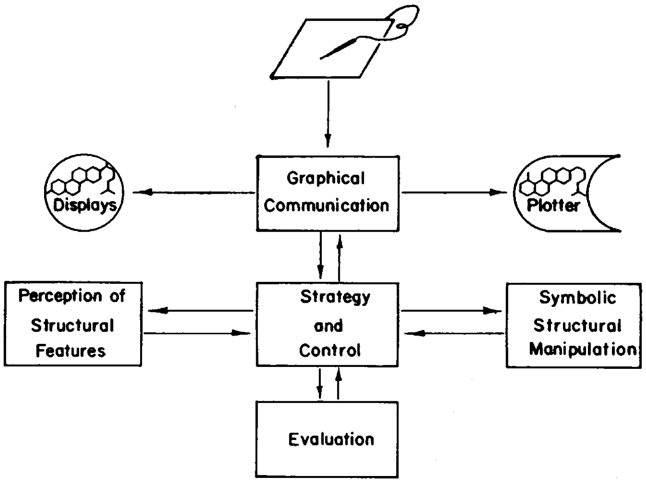


Fig. 3 (left). Functional subdivisions of the OCSS program.

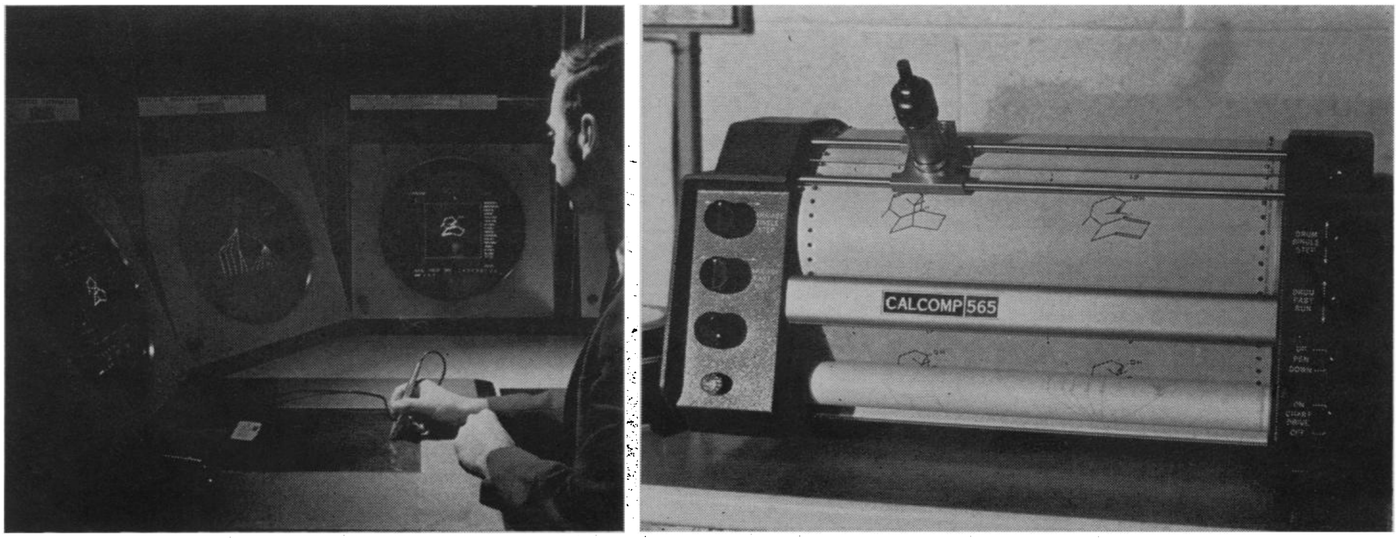
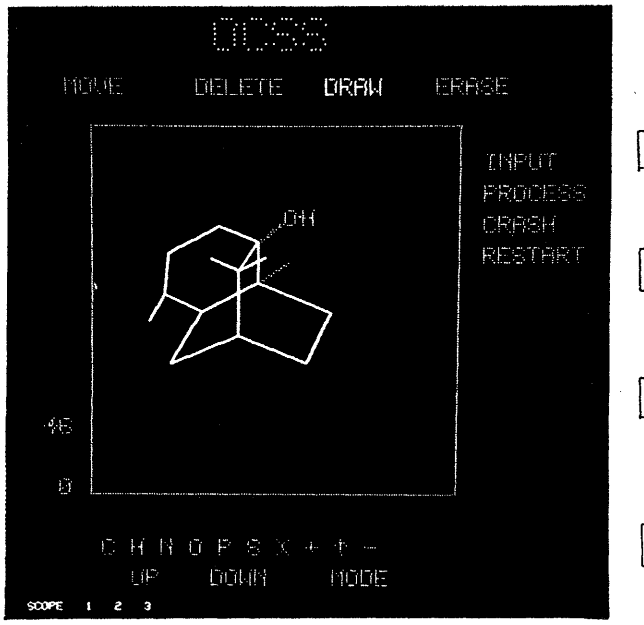


Fig. 4 (below). (Left) The OCSS system during the processing phase. The left scope is used for input with the Rand tablet shown beneath it. The middle scope shows the synthesis tree, and the scope on the right displays mechanisms and structures. (Right) Hard-copy output of structures and the synthesis tree is obtained from a Calcomp plotter.



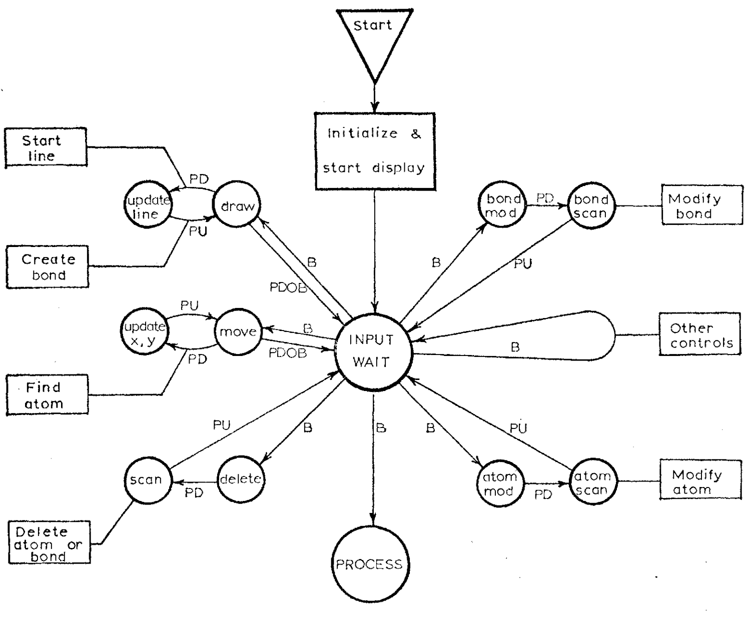


Fig. 5 (left). Scope 1 displays control words for input and the structural diagram as it is drawn. Fig. 6 (right). State diagram for OCSS graphic input. The conditions for changing states are PD, pen down; PU, pen up; B, pen down on appropriate control word; PDOB, pen down outside drawing box.

the operator points to PROCESS, the molecule represented by the drawing becomes the target molecule, and the processing state is entered.

The operation of the graphical input processor may be summarized with the state diagram shown in Fig. 6. In this diagram, circles represent states of the program, codes on the arrows indicate conditions for changing states, and attached rectangles represent actions performed while in a state or in the act of changing states.

During the structural input phase described above, the graphics module translates the molecular drawing (the external representation) into a form of connection table, the representation used internally by the computer. A connection table was first suggested solely for computer use by Mooers (*10*), and a variant of that table format is now used by Chemical Abstracts Service (*11, 12*) and others (*13*). The basic form of a connection table has, for each atom in the molecule except hydrogen, an entry containing the atom type, the attached atoms, and the connecting bond types. The connection table used in OCSS differs in that bonds are represented explicitly as entities having names instead of being represented implicitly as the thing between two atoms. Explicit bond representation facilitates all operations important to OCSS—structural analysis, structural manipulation, and structural display. Alternate representations (*14*) have various shortcomings. The connection matrix (*15*), for example, cannot represent absolute stereochemistry or stereoisomerism about a double bond and in some implementations (*16*) leads to ambiguities regarding double bond placement. The linear chemical notations, for example, Wiswesser (*17*), Dyson (IUPAC) (*18*), Hayward (*19*), and MCC (*20*), although economical with respect to storage space, and useful with respect to certain types of information storage-retrieval systems, prove very unsuitable for synthetic analysis because of the complexity of the codes and the implicit representation of structural features in the codes.

The connection table provides a rich base of fundamental information free from orientational prejudice, artificially imposed structural hierarchies, or enumerating systems, and is in a form convenient for computer manipulation. At present OCSS allows for as many as 36 explicit atoms. Each atom requires an atom table entry of four words allocated as shown in Fig. 7. The superscript refers to the number of bits assigned to each data item. The symbols are defined as:

U, this table entry in use
D, this atom to be displayed
S, this atom carries a charge or is not carbon
INFO, miscellaneous information
CHG, charge (neutral, cation, radical, anion)
NBDS, the number of valences explicitly used
NATCH, the number of attached explicit atoms
ATYPE, atom type (C, H, N, O, P, S, or X)
BOND 1–5, names of bonds to this atom

Each bond requires a bond table entry consisting of two words allocated as shown in Fig. 8, where the symbols are defined as:

U, this table entry used
D, this bond to be displayed
BINFO, miscellaneous
BSTEREO, bond "up" or "down"
BTYPE, bond type (single, double, triple)
ATOM 1} names of the atoms between
ATOM 2} which this is a bond

These two lists together with the separately stored atom position coordinates have the same informational content as the original structural diagram, and hence are sufficient to recreate the diagram.

The use made of graphics in this program is very important. It allows the chemist to use the structural diagram, a representation convenient for him, while the computer uses the connection table, a representation convenient for it. Absent are the burdens inherent in making the computer use a representation preferred by chemists or in forcing the chemist to use a representation designed for the computer. In actual practice chemists with only a 3-minute introduction to the system have entered complicated organic molecules successfully. Furthermore, since in the OCSS system drawing is natural

^{1}U	^{1}D	^{1}S	^{5}INFO	^{4}CHG	^{3}NBDS	^{3}NATCH
^{9}ATYPE				^{9}BOND1		
^{9}BOND2				^{9}BOND3		
^{9}BOND4				^{9}BOND5		

^{1}U	^{1}D	^{9}BINFO	^{3}BSTEREO	^{4}BTYPE
^{9}ATOM1			^{9}ATOM2	

Fig. 7 (left). The structure of an atom table entry.

Fig. 8 (above). The structure of a bond table entry.

and requires only one hand, one can enter complex molecules, such as cortisone (VIII), as rapidly as one can draw, in less than 30 seconds. Finally, the output from the program is in a form readily intelligible to a chemist. In fact, the structural formulas in this paper were drawn by OCSS. This graphical approach has effectively removed the long-standing communication barrier between organic chemist and computer.

B. Perception. The perception module, the eye of the program, starts with the basic connection table after input and derives higher level concepts about the structure, which are required for the functioning of the strategy and control module and the manipulation module. The perception module recognizes functional groups, chains, rings, appendages on rings or chains, atoms common to two or more rings, and redundancy or symmetry in the molecular skeleton or network. The perception of asymmetric centers and salient stereochemical interactions is planned as an addition in the next stage of program development.

In addition to the perception of functional groups, there is also a categorization of groups according to electronic properties, as a result of which one or more general group electronic descriptors are associated with each group. These are, for example, W for n- or π-electron withdrawing, D for n-electron supplying, and X for σ-heterolytic to form X^-. The relation between functional groups is perceived for all possible pairs of functional groups in terms of the number of bonds and atoms separating each pair. Similarly, pairwise relationships need to be perceived for (i) each appendage and each functional group, (ii) each ring and each functional group, (iii) each pair of common atoms, (iv) each pair of rings, and others. All possible rings and the size of each are perceived, as well as the relationship of one ring to another, for example, isolated (no common atoms), spiro (one common atom), fused (two common atoms directly bonded), and bridged (three or more common atoms). All this structural information is transmitted to the control and strategy module.

Important features of the key perception algorithms used in OCSS will now be described. The ring perception algorithm finds all cycles in the chemical graph. The algorithm begins by arbitrarily choosing an atom as origin. A path is then grown out from this origin along the molecular network until the path doubles back on itself, that is, the last atom A_n appears earlier in the path, say as A_i. The ring then consists of the

$$\text{path } A_1, A_2, \ldots, A_i, \ldots, A_{n-1}, A_n$$

sequence A_i, A_{i+1}, . . ., A_{n-1}. The ring is stored in the ring list if it does not duplicate a ring already in the list. The path is then shortened to the last atom having a still untraveled branch, and the growth is continued. If, when all paths from the origin have been traversed, all atoms in the structure have not been covered, then the structure consists of more than one fragment. In that case a new origin in the next fragment is chosen, and the process is continued.

Each ring is represented canonically as an ordered set by a binary string in which a 1 in the ith position indicates that the ith atom is a member of this ring. This representation permits the use of standard set operations and concepts, for example, union, intersection, and set inclusion, for handling rings and finding ring intersections.

The number of rings in the chemical sense (*21*) is given by $nrealrings = nb - na + nf$, where nb is the number of bonds without reference to bond multiplicity, na is the number of atoms, and nf is the number of fragments in the structure. These *real* rings are intuitively recognized by a chemist. The remaining rings, which a chemist does not normally recognize, are termed *pseudo* rings. A pseudo ring or envelope is a ring which is formed by the combination of one or more real rings. In Fig 9, ring 8 is pseudo because it includes real rings 4 and 10.

Let S_r be the set of all the rings in a structure, let a ring R be the set of all bonds in that ring, let $\mathcal{B}$ be the set of all bonds which are in rings, and let

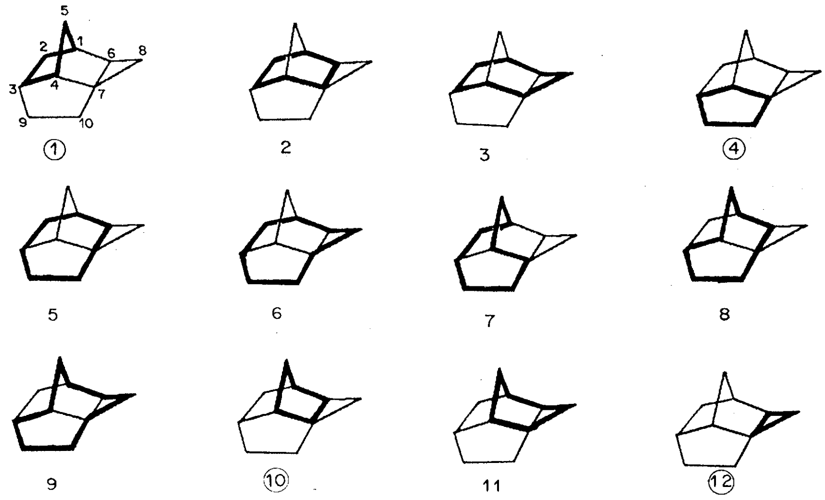


Fig. 9. Rings are perceived and classed as *real* (indicated by circles) and *pseudo.*

$*X$ be the number of members in set X (*22*). The intuitive notion of a *real* ring may now be formalized: a real ring is a ring which is a member of a *maximum proper covering set of rings* S_{mpc}, where an S_{mpc} is a set of rings S such that

1) S is a subset of S_r; S covers all bonds in $\mathcal{B}$ (*covering set*); and removal of any ring from S leaves a set which does not cover all bonds in $\mathcal{B}$.

2) The intersection of any two rings in S involves not more than one-half the bonds of either of the two rings (*proper set*).

3) The number of rings in S is greater than or equal to the number of rings in any other set that also satisfies conditions 1 and 2 above (*maximal set*).

In certain structures there is no unique S_{mpc} but instead of a collection $\mathcal{C}_{\mathrm{mpc}}$ of S_{mpc}'s, for example, in cubane there are six S_{mpc}'s, each containing *five* four-membered rings. More precisely, then, if $\mathcal{C}_c$ is the collection of covering sets,

$$\mathcal{C}_c \overset{\mathrm{def}}{} \Big\{S \subset S_r : \bigcup_{R' \epsilon S} R' = \mathcal{B} \text{ and } (\forall\ R \epsilon S)\ (\bigcup_{\substack{R' \epsilon S \\ R' \neq R}} R' \neq \mathcal{B})\Big\}$$

$$\mathcal{C}_c \overset{\mathrm{def}}{} \Big\{S \epsilon \mathcal{C}_c : (\forall\ R, R' \epsilon S, R' \neq R) \times [*(R \bigcap R') \leqslant *R'/2] \text{ and } (\forall\ S' \epsilon \mathcal{C}_c)\ (*S' \leqslant *S)\Big\}$$

Pick an $S_{\mathrm{mpc}}\ \varepsilon\ \mathcal{C}_{\mathrm{mpc}}$, then R is *real* $\overset{\mathrm{def}}{\rightleftharpoons} R\ \varepsilon\ S_{\mathrm{mpc}}$.

The OCSS perception module finds *one* S_{mpc} from the collection $\mathcal{C}_{\mathrm{mpc}}$ as follows:

1) Let sets S_{mpc} and S equal the *null* set, where S is the set of bonds covered.

2) Begin scanning the rings.

3) Proceeding from smallest to largest rings, if $R_i \bigcup S \neq S$ and R_i satisfies the intersection requirement, then replace set S by $R_i \bigcup S$ and replace set S_{mpc} by $\{R_i\} \bigcup S_{\mathrm{mpc}}$.

4) If $S = \mathcal{B}$, then go to 7.

5) If there are unscanned rings, go to 3.

6) Reinstate ring dropped in S_{mpc} and S_r and go to 8.

7) If S_{mpc} is maximal, that is, equals *nrealrings*, then *done*.

8) Temporarily drop a different ring from S_{mpc} and S_r. Let $S = \bigcup_{R\ \epsilon\ S_{\mathrm{mpc}}} R$ and go to 2.

Figure 9 shows sample results from this algorithm. The previously reported concept of *fundamental ring* (*23*) requires a more complex algorithm and seems less useful in synthetic analysis than the concept of *real* ring reported here. For all six four-membered rings in cubane to be classed as *real*, then the definition of a real ring must be

$$R \text{ is real} \rightleftharpoons R\ \epsilon \bigcup_{S\ \epsilon\ \mathcal{C}_{\mathrm{mpc}}} S$$

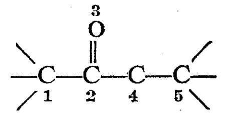
which defines a unique set of real rings. For this definition it is necessary to find *two* S_{mpc}'s, for the union of any two S_{mpc}'s in $\mathcal{C}_{\mathrm{mpc}}$ equals the union of all S_{mpc}'s in $\mathcal{C}_{\mathrm{mpc}}$.

Perception of functional groups is performed efficiently by a flexible table-driven recognizer (*24*). This recognizer makes use of a previously unreported node-by-node matching method which involves no backtracking and is more efficient for this application than normal node-by-node graph matching (*25*) or set-reduction methods (*26*). By normal graph matching techniques a graph of each functional group would be matched node-by-node against the substrate structure. By the method used in OCSS, a node in the substrate structure is matched against a hierarchal table of node-bond types (see Fig. 10).

The table entries consist of a pattern part and an instruction part. The instruction part of a matched entry tells the recognizer which atoms in the substrate to examine next and against which subtable to match. When a complete functional group has been found, the instruction part indicates that fact and also indicates the name of the functional group. The entries in the table are ordered with singly bonded carbon coming last in the table. When we examine the attachments to a given atom, the attachments are examined in the same order. Thus, if the first examined attachment is singly bonded carbon, then it is known that the other attachments are also singly bonded carbon. This hierarchal system speeds analysis and simplifies the table required.

Initially all atoms are eligible for starting the recognizer, but as a functional group is recognized, the atoms in that group are removed from eligibility. The recognizer is first initiated at eligible primary multiply bonded hetero atoms, then primary singly bonded hetero atoms, and finally, nonprimary hetero atoms.

In recognizing an ester group, for example, the initiating atom is the doubly bonded oxygen atom O(3). The recognizer finds a match for I° O. The code *nu* in the matched entry directs the recognizer to examine next the attachments to the atom just matched, O(3),



and the remainder of the instruction directs the recognizer to the next subtable. The recognizer finds a match for the attachment, =C(2), and is directed to examine the attachments to C(2), excluding O(3) which has already been traversed. The match on —O(4) directs examination to O(4) attachments. Finally, the –C(5) match indicates that an ester has been recognized. The names of the atoms and bonds in the ester group are recorded, and the group is assigned a unique name for later reference.

Part of the perception process is the reduction of the OCSS connection table to unique representation so that OCSS may recognize the structure if encountered again. By the method of Morgan (*11*), a unique connection table is generated by renumbering the atoms of the structure in such a way as to obtain the lowest valued connection table. The unique table is then compacted into a simple binary string that becomes the canonical name of the structure.

Graphical symmetry is recognized during the search for the canonical connection table. The program recognizes ties between alternative numbering schemes for the structure. A tie indicates the presence of an element of graphical symmetry. The program stores this symmetry element as a list of pairs of equivalent atoms and a list of atoms unchanged by the symmetry element. Graphical symmetry is a necessary but not sufficient condition for real molecular symmetry because graphical symmetry is based only on connectivity, node types, and bond types and not on three-dimensional properties of the structure. Nevertheless, graphical symmetry can be useful for synthetic planning.

C. Strategy and control. On the basis of all perceptions, the strategy and control module attempts to use a set of fundamental heuristics (*27*) of organic synthesis which are supplied to the program. These heuristics have been derived by careful analysis of the field of organic synthesis to determine the most powerful and general principles which lead from a chemical structure to a

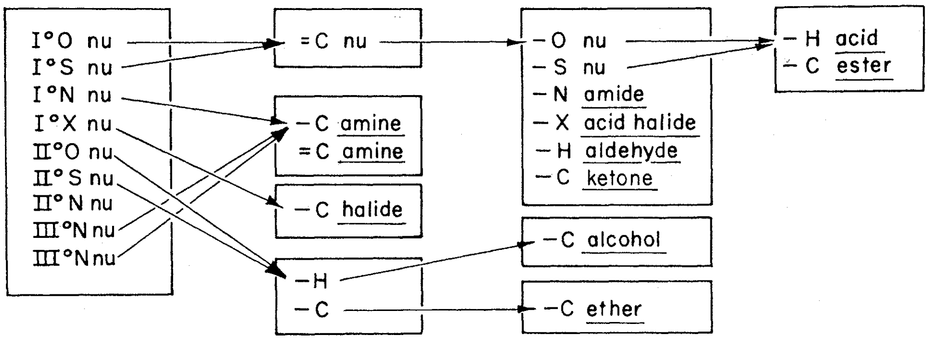


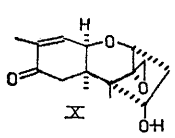
Fig. 10. Sample table for functional group recognition.

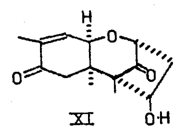
valid set of precursors in the synthesis tree. In many respects these heuristics are similar to the principles which have led chemists to effective and often unusually simple solutions of highly complex and subtle synthetic problems. The heuristics are specified in a form which is generally and systematically applicable to the great variety of organic structures. They lead to the development of goals by the strategy module and, directly or indirectly, the connection of these goals to specific commands in the manipulation module. The collection of heuristics in OCSS at present is far from complete, both because a number of powerful strategies which have been recognized have not yet been introduced into the program and because many new strategies remain to be devised or evaluated. The heuristics in the program can be categorized, depending on whether they relate primarily to functional groups, molecular skeleton or network, appendage groups, molecular geometry (stereochemistry), or combinations of these.

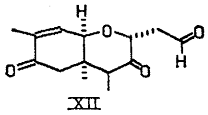
With regard to functional groups, for example, it is important to remove or modify highly reactive functional groups first, since it is true of a large majority of synthetic problems that unstable groups should be formed at the latest possible point in a synthesis. Another heuristic leads to the examination of all pairs of functional groups and the intervening atoms and bonds to ascertain whether mechanistic disconnections are possible. Certain pairs of functional groups in certain relations to each other allow mechanistic disconnections corresponding to highly effective synthetic operations. This heuristic functions in terms both of the general electronic group descriptors and the specific structure of each functional group. Another heuristic leads to the generation of subgoals involving the replacement of one functional group or general electronic descriptor by another when these subgoals lead to a skeletal or network disconnection. In all these strategies the fundamental approach depends upon the recognition of pairwise relations between functional groups. To our knowledge this approach has not previously been recognized by synthetic chemists, although it constitutes a most powerful and systematic procedure for human as well as computerized analysis.

Other heuristics in the functional group category deal with the strategies based on the properties of π-conjugated groupings, relationships which effectively reduce the number of functional groups, and strategies for connecting functionalized atoms to form common ring systems. In the category of network-based strategies are heuristics such as (i) the conversion of seven- to ten-membered rings to the more common five- or six-membered rings by rearrangement or internal elimination processes, (ii) the maximum reduction of rings by mechanistically allowed cycloelimination, (iii) the conversion of bridged rings to fused rings by rearrangement, and (iv) the cleavage of bonds to nucleophilic hetero atoms (such as N, O, and S) in the network. The underlying justification for these heuristics, although not immediately obvious, rests firmly on established synthetic chemistry [see, for example, (*2*)]. The useful heuristics involving appendages and molecular geometry follow closely common synthetic practice. For example, in the case of appendages, those attached to atoms bearing certain functional groups (such as OH or C=O) should be disconnected by the appropriate operation. Thus, sets of independent goals are derived by the strategy module which are based (separately) on functional groups, molecular skeleton or network, appendage groups, and molecular geometry. A transformation which serves more than one goal simultaneously will have a correspondingly higher priority which is effectively the summed inherited priorities of the goals served.

The operation of this scheme for setting worthwhile strategies can be illustrated by a specific example, for instance, the synthesis of the interesting antibiotic trichothecolone (X), a substance that has not yet been synthesized. The appendage functional group heuristics lead to generation of intermediate XI which is disconnected in a high-priority operation based on pairwise consideration of functional groups to intermediate XII. The priority of this last disconnection is further enhanced because it also satisfies the network-oriented heuristics. Most chemists would agree that this route of synthesis of X possesses considerable merit.


X

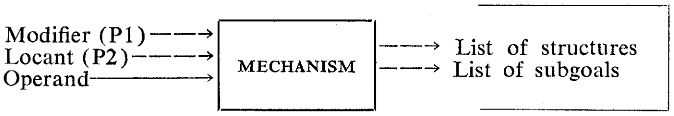

XI


XII

The effectiveness of the heuristics referred to above is one of the critical factors in the performance and quality of a computer program for synthetic analysis. It is not surprising, therefore, that present studies on the refinement of OCSS are of great usefulness both to the evaluation of the applicability of a given heuristic to a wide range of problems and to the development of new heuristics.

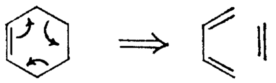
D. Manipulation. The manipulation module performs the symbolic chemical transformations to create precursor structures. In the module are subroutines which operate on the connection table to perform the symbolic inverse of synthetic interconversions. The functioning of these subroutines results in the making and breaking of bonds, the addition or loss of atoms, and the addition or loss of charge. These routines also check the structural factors which are required for a given manipulation to be valid. Two kinds of symbolic chemical transformations are used, *symbolic mechanism* and *symbolic functional group modification.*

A symbolic mechanism may or may not correspond to one complete chemical reaction, and the detailed electronic mechanism of a chemical reaction need not be known to be symbolized. Schematically, a symbolic mechanism takes as input P1, a modifying parameter, P2, a location parameter, and an operand structure. It produces as output a list of structural precursors of the operand and a list of subgoals (either may be a null list).



Taking these parameters in turn, P1 may specify a variation of the mechanism, for example, whether, for a 1,3-diene addition reaction, the entering groups are placed 1,2 or 1,4. The P2 parameter specifies the location on the operand molecule on which the mechanism is to work. It may specify an atom, a bond, a functional group, or a ring. If P1 and P2 are not specified, then the mechanism omits P1 and assumes all reasonable P2's. Obviously, an operand must always be specified. Since one mechanism may create several intermediates, its output is in the form of a list. The mechanism also creates a list of subgoals to be tried in case none of the intermediates created are acceptable. These subgoals are lists of individual mechanisms or sequences of mechanisms and are often proposed to circumvent a block or difficulty in application of the original mechanism. They may, however, be simply other reasonable things to try next.

The cyclic elimination reaction is an important process for molecular disconnection which can serve to illustrate the nature of a symbolic mechanism. The cyclic elimination process is the reverse of synthetic reactions such as the Diels-Alder reaction (reaction 1), photoaddition (reaction 2), valence tautomerism (reaction 3), oligomerization (reaction 4), and photocyclization (reaction 5). The manipulation corresponding to the cycloelimination mechanism entails the finding of a double bond in an even-membered ring and the alternate breaking and making of bonds around the ring, that is, the equivalent of shifting bonds as shown below.



A four-membered ring needs no double bond, but other even-membered rings do. If there is no double bond, then before the cyclic elimination can operate, a double bond must be introduced; this is accomplished by creation of the subgoal for the replacement of a single bond by a double bond.

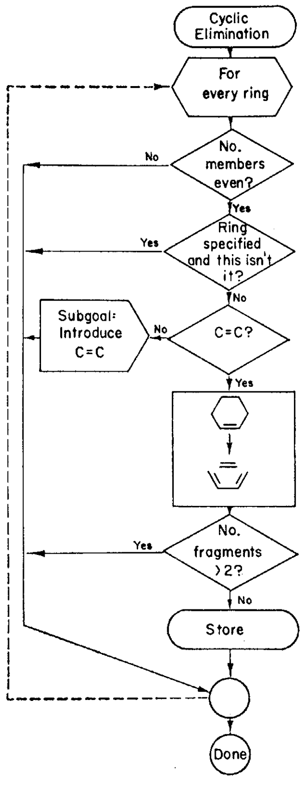
Figure 11 illustrates in flow chart form the operation of the symbolic cycloelimination mechanism.

Another very important symbolic mechanism is that of "OH-type" cleavage which encompasses the wide range of carbonyl addition and analogous reactions, some of which are illustrated by reactions 6 to 11. The mechanism operates as shown in reaction 12. A flow chart is given in Fig. 12.

An equally important symbolic mechanism is "W cleavage" which encompasses nucleophilic displacements and nucleophilic conjugate additions in which the nucleophile is an anion stabilized by an adjacent electron-withdrawing group (W). Examples are provided by reactions 13 to 15, and the flow chart is given in Fig. 13.

Certain of the symbolic mechanisms follow closely actual chemical mechanisms and involve the generation of structures corresponding to reactive intermediates. For example, carbon-carbon cleavage or rearrangement processes via carbonium ions involve manipulation and output of the intermediate cationic structures as well as the neutral species derived by elimination or association.

The second type of manipulation, symbolic functional group modification, results in the exchange, introduction, or removal of functional groups but does not in itself affect the skeletal connections in the molecule. Such manipulation is usually called for by functional group–oriented heuristics when it leads to the realization of a goal, for example,



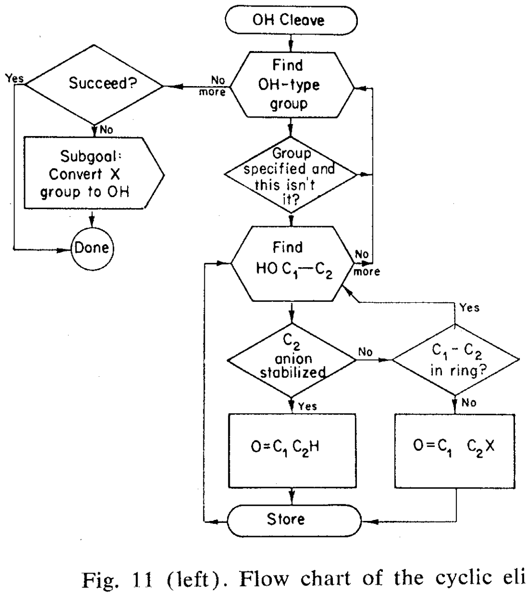
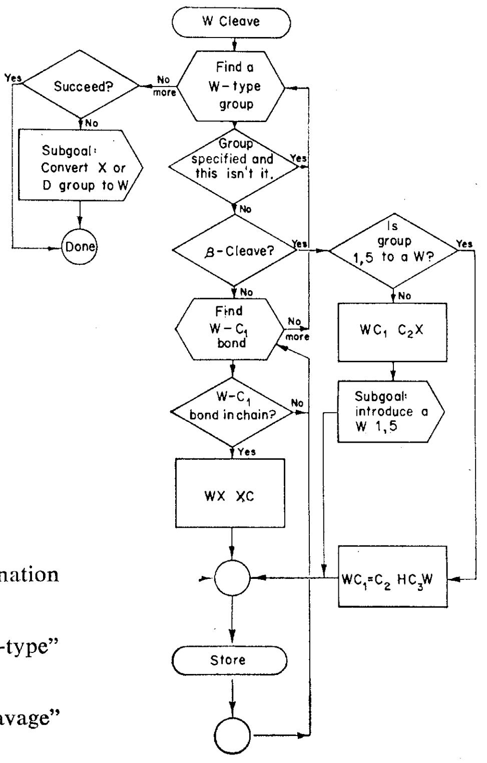


Fig. 11 (left). Flow chart of the cyclic elimination mechanism.

Fig. 12 (above). Flow chart of the "OH-type" cleavage mechanism.

Fig. 13 (right). Flow chart of the "W cleavage" mechanism.

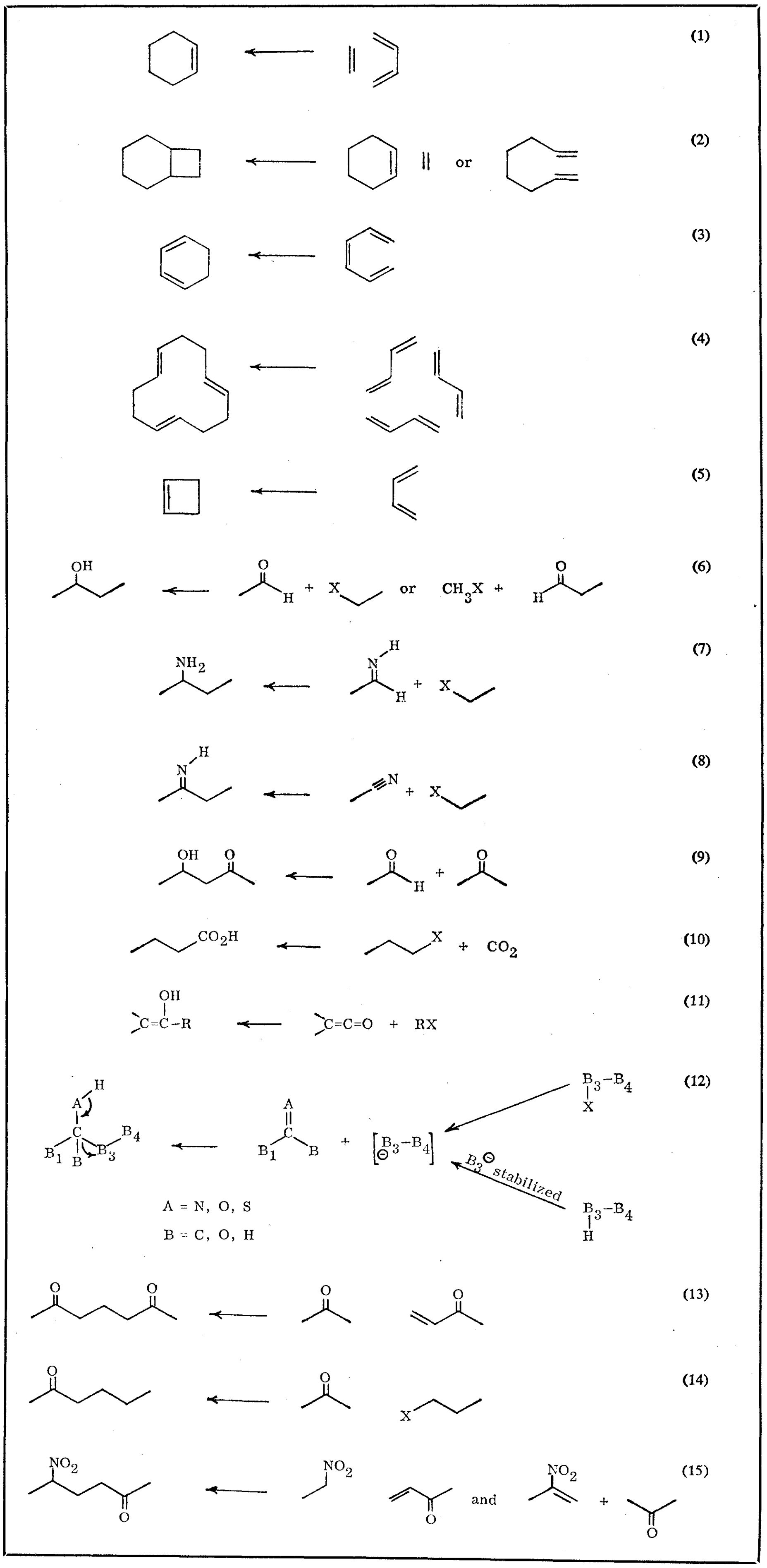


when it makes possible network disconnection by a symbolic mechanism. There is a close parallel in human synthetic analysis. Symbolic functional group modification accomplishes transformations which may correspond to several chemical reactions or steps. It is not concerned with intermediates or mechanistic details of such reactions, the justification simply being that there are a sufficient number of synthetic techniques to allow with high probability the interconversion of any two types of functional group.

E. Evaluation. The evaluation module is the least well-formulated part of the OCSS system. From the outset it was intended that the bulk of evaluation be done by the chemist for reasons which have been outlined above. However, there are several aspects of the evaluation process which can be adequately delineated and have in part been implemented in OCSS.

There are two levels of evaluation performed by the evaluation module. The first pertains only to properties of the structure itself, while the second pertains to (i) changes in these properties in going from one structure to its precursor, (ii) goal satisfaction, and (iii) properties of the synthesis tree itself. In evaluating the structure itself, the evaluation module checks for elementary conditions such as valence violations (these could occur under circumstances not anticipated by the generating operation), unlikely charge distribution (dication or dianion), and implausible topology (such as implied planarity of the bonds to an atom when such planarity is energetically unfavorable). A structure which fails this simple evaluation is removed from further consideration and, consequently, never reaches the chemist.

The module next evaluates the structure for uniqueness. The canonical name of the structure is simply matched against the names of all other structures in the synthesis tree. A match indicates that the structure has been generated previously and is represented by another node in the tree. The action taken depends upon the relations between the two duplicate structures. If both structures have the same parent (the node above this node as in a family tree), then the structure created last is deleted. This situation arises from mechanistic equivalence or structural symmetry. If the two structures have different parents, then the overall characteristics of the two paths in the



synthesis tree must be evaluated at a later stage to determine if one path is better than the other.

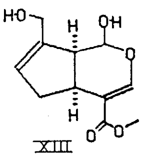
Structures surviving the previous tests are then evaluated by the program with respect to their relative simplicity. The structures are awarded points on an arbitrary scale in various categories, for example, network simplicity, appendage simplicity, and functional group simplicity. Ring system simplicity VTOPO2 is assessed by

$$\text{VTOPO2} = 55 - (\mathit{nfused} + 2 \times \mathit{nbridged} + 3 \times \mathit{nspiro}) - \sum_{i=1}^{\mathit{nrealring}} (|\mathit{rsize}_i - 6| + 2)$$

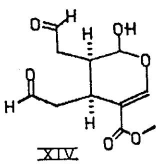
where *nfused, nbridged,* and *nspiro* correspond, respectively, to the number of fused, bridged, and spiro ring junctures, *rsize* corresponds to the ring size, and the summation occurs over all *real* rings. Appendage simplicity VTOPO3 is assessed by

$$\text{VTOPO3} = (\mathit{ncarbon\ fragments} \times 8) + 8 - \mathit{nprimary\ atoms}$$

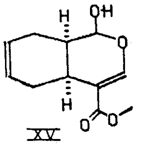
The assessed value for each category then is placed into the evaluation vector. Comparison of evaluation vectors starts with the leftmost (most significant) elements of the two vectors and proceeds to the right only if the elements are equal. The relative importance of the categories of evaluation changes at the direction of a goal in order to reward attainment of that goal. An example of the need for this is in the synthesis of genipin XIII (*28*). In generating XV, the precursor of XIV, the ring simplicity is decreased, but, in the overall transformation from XIII, the ring simplicity is increased as well as that of the appendages. Thus, in evaluating the step from XIV to XV, since the goal was reconnection, the increased complexity of ring structure is less important than the increased simplicity of appendages. Crude as this method of evaluating structural simplicity is, it does provide a basis for directing further search in the synthesis tree. This evaluation is used only to rate structures and not to eliminate structures.



XIII



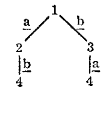
XIV



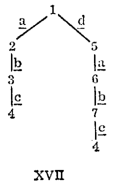
XV

The structures are also ranked as to how well the goal or goals responsible for their creation were satisfied. Some goals can be attained in one step as the reconnection goal from XIV. Other goals such as ring expansion in XIII ⇒ XV can only be attained after several steps. The present version of OCSS only considers one-step attainment of goals, and the goal structure is still fairly simple.

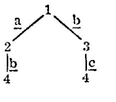
The last stage of evaluation is that of considering the overall situation in the synthesis tree. As mentioned previously, this is necessitated when duplicate structures are encountered which have different parents. The four possible cases, illustrated by XVI to XIX, differ in the relation between the paths to the lowest common ancestor (PLCA). In case XVI the paths differ only in the ordering of transformations. This presents the problem of choosing that path with the better ordering. In many cases the choice of ordering is based on subtle operational factors related only to the laboratory execution of the synthesis. In case XVII one path is a subset of the other, indicating step *d* was unnecessary. Clearly, the longer sequence should be deleted. In both remaining cases, XVIII and XIX, there exists a nontrivial alternate route for transforming 4 to 1. Such a plan is good, as it allows flexibility in experimental execution—failure of one route still leaves one route open which utilizes the *same starting materials.* Thus, the value of both paths is raised, and the lowest common precursor is regarded as more valuable. In case XIX the shorter of the paths may be preferable, but that is not necessarily true.



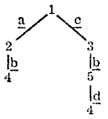
XVI



XVII



XVIII



XIX

Thus the evaluation module performs three tasks: (i) elimination of obviously poor structures, (ii) measurement of goal satisfaction, and (iii) provision of guidance to the control module for further exploration of the synthesis tree. The strategy of tree exploration is to concentrate on the paths that provide rapid simplification. It is uncertain at this time what the optimum strategy should be.

Illustration

In order to complete the outline of our approach to computer-assisted synthetic analysis, an illustrative example is included in this section. Formula 1 in Fig. 14 represents the structure of patchouli alcohol, a naturally occurring substance which finds wide use in perfumes. This example was chosen as typical of many problems involving molecules of moderate complexity, the synthesis of which is fairly subtle. In this case functionality, carbon network, and appendage groups all play a role in the derivation of possible synthetic routes. Also the patchouli alcohol structure is small enough so that a relatively small number of chemical transformations produce simple precursors. Even so, it was necessary to delete a considerable number of the precursors generated by the computer in order to limit the space required for the illustration. It should be noted that patchouli alcohol has already been synthesized in the laboratory by Büchi (*29*) and by Danishefsky and Dumas (*30*) by quite different routes.

One of the synthetic pathways shown in Fig. 14 consists of sequence 35 → 16 → 11 → 1. This corresponds to an extremely direct and plausible route which, while related to the Danishefsky synthesis, is considerably simpler and in certain respects more elegant. The sequence 34 → 16 → 11 → 1 is interesting but more unconventional and more speculative. The synthesis of 11 via 17 (read NO_2 for OH) is simple, novel, and plausible. Another interesting route of considerable merit is represented by the sequence 62 → 43 → 33 → 1. Further analysis of precursors 37, 40, and 42, which are related to 33 by a 1, 2 shift of carbon, not shown because of space limitations, leads to other, less direct synthetic routes. In this connection it should be mentioned that the Büchi synthesis proceeded via such an indirect path through intermediates 55, 39, and 51 (α-patchoulene). Finally, the cyclization of 49 to 33 represents another plausible route to 1 which is also related to the Danishefsky synthesis.

The evaluation of computer-assisted synthesis from the point of view of the chemist requires the scrutiny of many different synthetic problems involving the wide range of structures. However, the illustration shown in Fig. 14 gives a fairly good idea of the performance which has already resulted from the first phase of the present study. One of the most interesting and important characteristics of the computerized analyses is the high frequency with which intermediates of a very novel and provocative nature are generated.

Conclusion

The application of digital computers to the generation of paths for the chemical synthesis of complex molecules has in the past seemed improbable or at best inconsequential to many chemists. Perhaps the clearest result of the investigation here reported is that such a development can now be seen as a distinct and promising probability. A number of the obstacles barring the way have been removed, including the problems of facile graphical communication between chemist and computer, the development of techniques and software for perception and manipulation of chemical structure by machine, and the formulation of heuristics which allow the setting of strategies and goals needed to direct the analysis. The task is complex, however, and much remains to be done. Useful improvements in hardware can be expected. The use of large computers on a time-sharing basis together with a real time graphics terminal would allow the operation of still larger and more powerful programs and permit more extensive synthetic analysis. Further gains would accrue from impending developments in peripheral equipment, for example, to speed hard-copy graphical output.

Given the requisite computing and graphic communications hardware, it is possible to envisage the development at some future time of a program or set of programs of such power as to raise the technique of computerized synthetic analysis to the status of an indispensable aid to the chemist. The start which has been made is especially meaningful because it points up sharply the areas where further software and heuristics are needed. Three-dimensional structural information will have to be included in the analysis. This information, which is required for the use of stereochemical heuristics and manipulations, will have to be generated by computations of three-dimensional molecular geometry from the atom and bond connection tables. Although methods for calculations of a similar nature have already been devised (*31*), the introduction of three-dimensional information and stereochemical heuristics constitutes a major undertaking (*32*). The development of heuristics involving electronically delocalized systems (and including quantum mechanical considerations and complex, many-center functional groups) represents another area where future effort is needed. Further progress also demands the expansion and improvement of the types of heuristics already in use. Eventually, the hierarchal ordering of all heuristics and the allowance of variation of this ordering with regard to key structural parameters can be made in a way which enhances effectiveness still more. It is also clear that the expansion of the available data base in the direction of a fairly complete set of structural manipulations corresponding to synthetic reactions would be in order.

These requirements add up to a substantial long-range effort. Indeed, the nature of the endeavor is such that any problem-solving methods developed are subject to further improvement either as a result of knowledge gained through their application or from new discovery in chemistry. Thus a lively and prolonged evolutionary development of this new discipline can be anticipated, one result of which is certain to be a much deeper understanding of the foundations and principles of chemical synthesis.

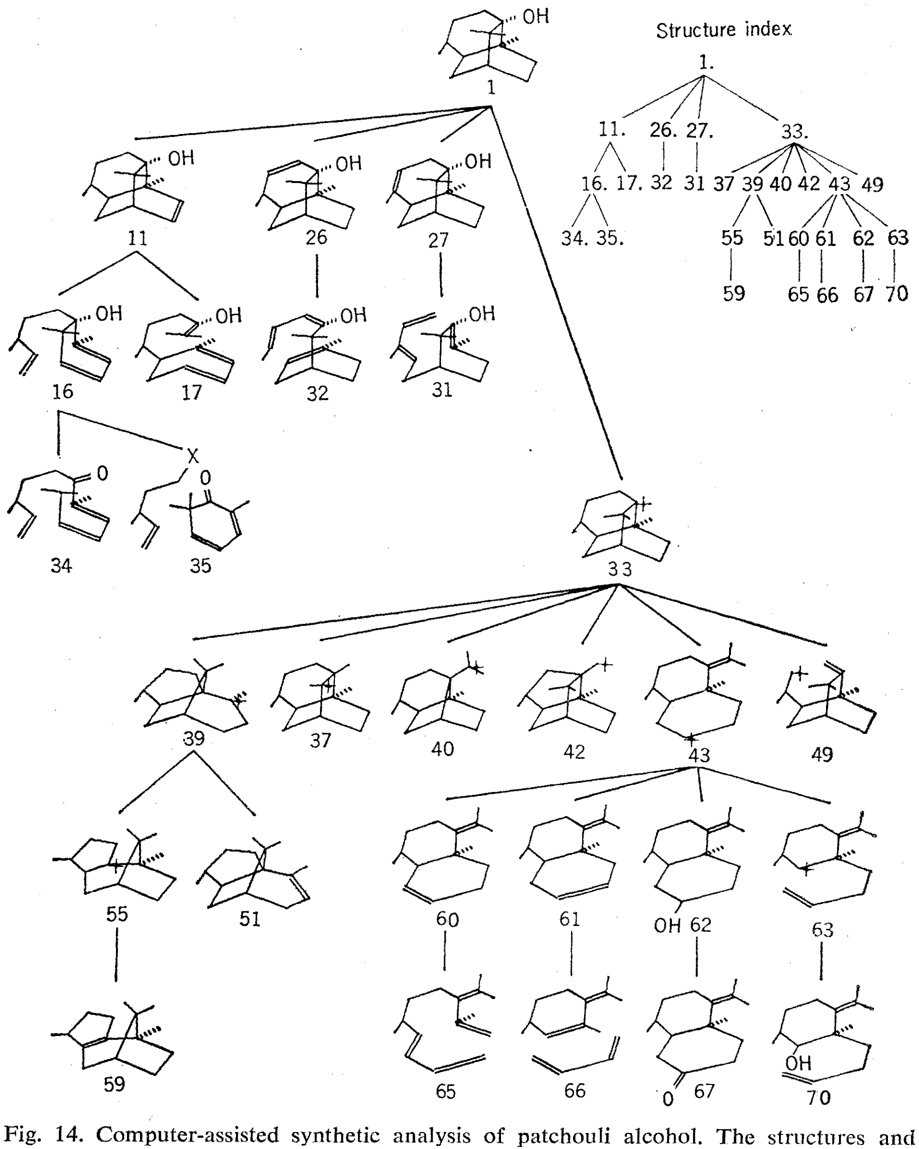


Fig. 14. Computer-assisted synthetic analysis of patchouli alcohol. The structures and structure index were generated and drawn by OCSS.

References and Notes

1. H. R. Henze and C. M. Blair, *J. Amer. Chem. Soc.* **53**, 3077 (1931).
2. E. J. Corey, *Pure Appl. Chem.* **14,** 19 (1967).
3. See L. F. Fieser and M. Fieser, *Steroids* (Reinhold, New York, 1959), pp. 644–650.
4. E. J. Corey, M. Ohno, R. B. Mitra, P. A. Vatakencherry, *J. Amer. Chem. Soc.* **86**, 478 (1964).
5. For a general reference on electronic reaction

mechanisms, see E. S. Gould, *Mechanisms and Structure in Organic Chemistry* (Holt-Dryden, New York, 1959).
6. I. E. Sutherland, "Sketchpad: A man-machine graphical communication system," Technical Report No. 296, M.I.T. Lincoln Laboratory, Lexington, Mass., 30 January 1963; see also L. D. Harmon and K. C. Knowlton, *Science* **164**, 19 (1969).
7. I. E. Sutherland, *Sci. Amer.* **215**(3), 86 (1966), and other articles in that issue; M. R. Davis and T. O. Ellis, in *Proc. Amer. Fed. Inform. Processing Soc. Joint Computer Conf.* **25**, 325 (1964); J. F. Teixeira and R. P. Sallen, in *ibid.* **32**, 315 (1968).
8. For a discussion of other goal-directed programs, see *Computer and Thought*, E. A. Feigenbaum and J. Feldman, Eds. (McGraw-Hill, New York, 1963).
9. M. A. Couper, *Compt. Rend. Acad. Sci. Paris* **46**, 1157 (1858).
10. C. Mooers, *Zator Tech. Bull. No. 59* (Zator Co., Boston, Mass., 1951).
11. H. L. Morgan, *J. Chem. Doc.* **5**, 107 (1965).
12. A. E. Petrarca and J. E. Rush, *ibid.* **9**, 32 (1969).
13. D. J. Gluck, *ibid.* **5**, 43 (1965).
14. For a comprehensive review through 1966, see F. A. Tate, *Annu. Rev. Inform. Sci. Technol.* **2**, 285 (1967).
15. G. W. Wheland, *Advanced Organic Chemistry* (Wiley, New York, ed. 2, 1949), p. 87; L. Spialter, *J. Amer. Chem. Soc.* **85**, 2102 (1963); *J. Chem. Doc.* **4**, 261 (1964).
16. N. Jochelson, C. M. Mohr, R. C. Reid, *J. Chem. Doc.* **8**, 113 (1968).
17. W. J. Wiswesser, *A Line-Formula Chemical Notation* (Crowell, New York, 1954); E. G. Smith, *Wiswesser Line-Formula Chemical Notation* (McGraw-Hill, New York, 1968).
18. G. M. Dyson, W. E. Cossum, M. F. Lynch, H. L. Morgan, *Inform. Storage Retrieval* **1**, 66 (1963).
19. H. W. Hayward, H. M. S. Sneed, J. H. Turnipseed, S. J. Tauber, *J. Chem. Doc.* **5**, 183 (1965).
20. D. Lefkovitz, *ibid.* **7**, 186, 192 (1967).
21. M. Fre'rejacque, *Bull. Soc. Chim. France* **6**, 1008 (1939).
22. For a description of the set notations used here, see P. R. Halmos, *Naive Set Theory* (Van Nostrand, Princeton, N.J., 1960).
23. R. Fugmann, U. Dölling, H. Nickelsen, *Angew. Chem. Int. Engl. Ed.* **6**, 723 (1967).
24. T. E. Cheatham and K. Sattley, in *Proc. Amer. Fed. Inform. Processing Soc. Joint Computer Conf.* **24**, 31 (1964).
25. L. C. Ray and R. A. Kirsch, *Science* **126**, 814 (1957); W. E. Cossum, M. L. Krakiwsky, M. F. Lynch, *J. Chem. Doc.* **5**, 33 (1965).
26. E. H. Sussenguth, Jr., *J. Chem. Doc.* **5**, 36 (1965).
27. Heuristic is used here as a noun to mean heuristic principle, a "rule-of-thumb" which may lead by a shortcut to the solution of a problem or may lead to a blind alley.
28. G. Büchi, B. Gubler, R. S. Schneider, J. Wild, *J. Amer. Chem. Soc.* **89**, 2776 (1967).
29. G. Büchi and W. D. MacLeod, *ibid.* **84**, 3205 (1962).
30. S. Danishefsky and D. Dumas, *Chem. Commun.* **1968**, 1287 (1968).
31. K. B. Wiberg, *J. Amer. Chem. Soc.* **87**, 1070 (1965); N. L. Allinger, M. A. Miller, F. A. Van Catledge, J. A. Hirsch, *ibid.* **89**, 4345 (1967); C. Levinthal, *Sci. Amer.* **214**(6), 42 (1966).
32. L. Velluz, J. Valls, J. Mathieu, *Angew. Chem. Int. Engl. Ed.* **6**, 778 (1967).
33. Supported by the NIH (contract PH-43-68-1018) and by the Advanced Research Projects Agency (PDP-1 maintenance grant to Harvard University). We thank Prof. Ivan Sutherland and Thomas Cheatham for discussions and suggestions and Scott Steketee and Mrs. Edward Dennis for programming assistance. We also thank Dr. R. E. Stobaugh of Chemical Abstracts Service for providing documentation of the Morgan algorithm (*11*).

The Siege of the House of Reason

Max Tishler

It seems rather simplistic to talk, as some have, about recent campus disruptions solely in terms of irresponsible youth or college administrators who have too much bone in their heads or too little in their backs. The university is under siege today as much from without as from within. The demands of our dynamic society have transformed it too quickly and under too much pressure from an important, but not central, institution into a full-fledged member of the American establishment. I propose to examine some of these demands and pressures, focusing my comments mainly on those aspects of the universities that are related to science and with which I am familiar.

First, however, let me explain what I mean when I use the word "establishment." I am not talking about a power center. The university does not make any decisions for American society, as do politicians, labor unions, or industry. The university as an institution has power only within its own confines. And groups of students have demonstrated recently that even that limited power is open to challenge.

But the university is a key institution. Nearly half our young men and women spend important years of their lives on its campus. It is the indispensable educator of skilled people without whom we could not run our complex industrialized nation. Its faculty shares its knowledge, judgment, and ideas not only with the students but with practically every facet of the culture—from experimental kindergartens to the White House. The university is the cradle of most of the basic research and much of the new technology that are powering our economic growth, shielding our republic, and transforming the quality of our lives. It is the forerunner of change, the critic of the status quo, and the guardian of objective rationality, without which both our civilization and mankind may be doomed.

Is the university doing all of these things well? Of course not. No single institution could take on so many assignments at once and stay on the Dean's list. I would pass out a few A's and B's but they would be well balanced, I'm afraid, by C's and even a few D's. Some of the tasks are mutually conflicting. No professor, for instance, can simultaneously counsel his government, conduct important research, and satisfy his students. Time is not a rubber band.

But the basic trouble that afflicts the university, it seems to me, derives from the pressures of the outside society. Take, for instance, the enormous expansion in enrollments. This puts a painful strain on the whole institution—students, faculty, and administration. It is the result of irresistible demands for increasing quantities of trained graduates by government, industry, education, the sciences, and other professions. College becomes the only gateway to rewarding jobs in the adult world. Some adults criticize student rebels on the grounds that these young people do not understand that a college education is a privilege. They are mistaken. They are thinking of the world in which they were brought up, not today's world. A college education is no longer a privilege. It is a necessity.

This is why the student population in our colleges and universities has more than doubled during the period of 1955–65, from 2.7 million to 5.7 million (*1*). Today almost 50 percent of our college-age population is enrolled in these institutions of learning, in contrast to Great Britain, France, Italy, and Germany where the percentages are below 20.

The author is senior vice president for research and development of Merck & Co., Inc., Rahway, New Jersey 07065. This article is adapted from an address given on 23 May 1969 at Kent State University, Kent, Ohio.