THE JOURNAL OF Organic Chemistry

VOLUME 45, NUMBER 11

MAY 23, 1980

Computer-Assisted Mechanistic Evaluation of Organic Reactions. 1. Overview

Timothy D. Salatin and William L. Jorgensen*[1]

Department of Chemistry, Purdue University, West Lafayette, Indiana 47907

Received November 19, 1979

An interactive computer program, CAMEO, is being developed to predict the products of organic reactions given starting materials and conditions. In contrast to earlier programs which use a large empirical data base of reaction tables, a highly mechanistic approach has been taken to reaction evaluation in CAMEO. The mechanistic class that has been implemented first is base-catalyzed and nucleophilic chemistry. Organizing principles have been established to oversee the competitions between proton-transfer, substitution, addition, and elimination reactions. With these rules, a knowledge of electron pushing, and much automatically perceived information on structure, functionality, pK_a levels, and nucleophilic and electrophilic sites, the program is able to make sophisticated, mechanistically sound predictions on the outcomes of organic reactions.

I. Background and Rationale

The novel idea of using computers to help design organic syntheses was first brought to realization by Corey and Wipke 10 years ago.[2] Since that time many diverse approaches to this intriguing problem have been developed[3-10] and a review is available.[11] Virtually all of the programs operate in the antithetic direction by transforming the target molecule to more accessible precursors. Some programs such as LHASA[2,12] and SECS[10] are interactive with the user, helping guide the selection of promising routes, while others[4,5,8] execute without human intervention. The selection of transforms (retroreactions) depends critically on data tables or reaction matrices corresponding to known synthetic operations. Such empirical data bases are often voluminous and restrict their programs to established chemistry.

For the past four years, an interactive computer program has been under development at Purdue whose aim is to predict the products of organic reactions given starting materials and conditions. The program represents a dramatic departure from the earlier work in the computer-synthesis field for several reasons. First, it operates in the synthetic (forward) direction. It has been implemented on a dedicated minicomputer, Texas Instruments 990/10. And, most importantly, it is mechanistically based; i.e., it uses mechanistic principles for broad classes of reactions to make its predictions for products and does not depend on data tables for numerous specific, often mechanistically similar, processes. Thus, the program is not intended so much to design complete synthetic routes as to help assess the feasibility of individual steps. An advantage to the forward mode is that side products of reactions are explicitly displayed, whereas the retrosynthetic programs generally provide only an indirect indication of potential conflicts by decrementing transform ratings.[2,10,12] The mechanistic approach also permits the prediction of new reactions that involve mechanistically sound operations. Furthermore, the project is being used as a vehicle to help seek out general organizing principles for organic reactions. The recognition of such principles is fundamentally important in addition to permitting greater efficiency in the program. The potential pedagogic value of a sophisticated program of this type is also apparent.

The program has been named CAMEO for computer-assisted *m*echanistic *e*valuation of *o*rganic reactions. The

(1) Camille and Henry Dreyfus Foundation Teacher-Scholar, 1978–1983; Alfred P. Sloan Foundation Fellow, 1979–1981.

(2) (a) E. J. Corey, *Pure Appl. Chem.*, **14**, 19 (1967); (b) E. J. Corey and W. T. Wipke, *Science*, **166**, 178 (1969); (c) E. J. Corey, *Q. Rev., Chem. Soc.*, **25**, 455 (1971).

(3) R. Barone, M. Chanon, and J. Metzger, *Rev. Inst. Fr. Pet. Ann. Combust. Liq.*, **28**, 771 (1973).

(4) H. Gelernter, N. S. Sridharan, A. J. Hart, S. C. Yen, F. Fowler, and H. Shue, *Top. Curr. Chem.*, **41**, 113 (1973); H. L. Gelernter, A. F. Sanders, D. L. Larsen, K. K. Agarwal, R. H. Boivie, G. A. Spritzer, and J. E. Searleman, *Science*, **197**, 1041 (1977).

(5) J. Gasteiger and C. Jochum, *Top. Curr. Chem.*, **74**, 95 (1978).

(6) G. Moreau, *Nouv. J. Chim.*, **2**, 187 (1978).

(7) M. Bersohn, *Bull. Chem. Soc. Jpn.*, **45**, 1897 (1972).

(8) I. Ugi, *Rec. Chem. Prog.*, **30**, 289 (1969); *Intra-Sci. Chem. Rep.*, **5**, 229 (1971).

(9) J. B. Hendrickson, *J. Am. Chem. Soc.*, **93**, 6847 (1971); *ibid.*, **93**, 6854 (1971); *ibid.*, **97**, 5763 (1975); *ibid.*, **97**, 5784 (1975); *Top. Curr. Chem.*, **62**, 49 (1976); *J. Chem. Inf. Comput. Sci.*, **19**, 129 (1979).

(10) W. T. Wipke and T. M. Dyott, *J. Am. Chem. Soc.*, **96**, 4825, 4834 (1974).

(11) M. Bersohn and A. Esack, *Chem. Rev.*, **76**, 269 (1976).

(12) (a) E. J. Corey, W. T. Wipke, R. D. Cramer III, and W. J. Howe, *J. Am. Chem. Soc.*, **94**, 421 (1972); (b) E. J. Corey, W. T. Wipke, R. D. Cramer III, and W. J. Howe, *J. Am. Chem. Soc.*, **94**, 431 (1972).

0022-3263/80/1945-2043$01.00/0

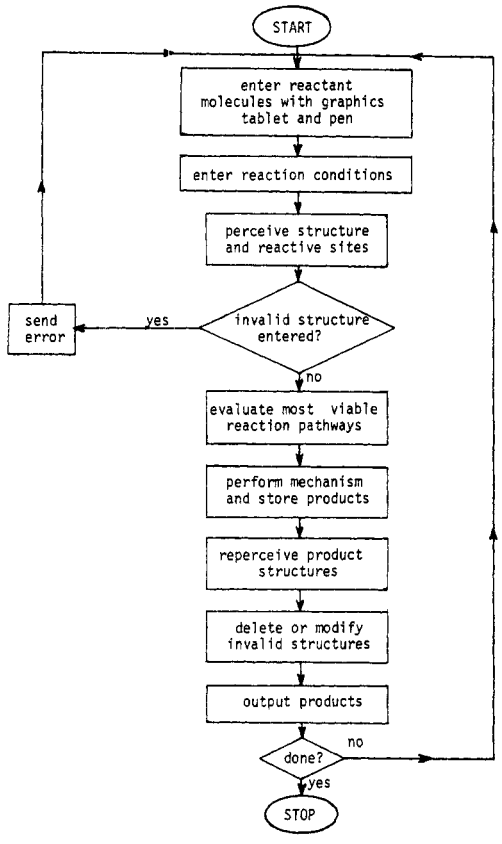


Figure 1. Schematic diagram of the overall program flow in CAMEO.

purpose of this paper is to describe the progress that has been made so far. To begin, several broad classes of organic reactions may be identified partly on the basis of intermediates. These are (1) base-catalyzed and nucleophilic chemistry, (2) acid-catalyzed and electrophilic, (3) radical, (4) carbenoid, (5) thermal pericyclic, and (6) photochemistry. CAMEO as currently planned will have more or less separate mechanistic segments for each category. Other modules may eventually be added for specific oxidations, reductions, and some organometallic chemistry that do not fit the above categories or are mechanistically obscure and require more empirical analyses. The first area to be addressed as described here is base-catalyzed and nucleophilic chemistry, since reactions in this category are mechanistically complex and so important for carbon–carbon bond formation. Development of capabilities for acid-catalyzed and pericyclic processes are actively under way.

Several tenets have been adhered to in creating CAMEO. First, the chemist/computer interface must be as facile as possible. Practicing chemists or students lacking any programming skills should be able to interact easily with the program. The use of interactive computer graphics is the most logical way to accomplish this goal. However, interacting graphics requires either a time-sharing or dedicated computer system. The latter was chosen in the form of a moderate minicomputer.

The second rule is that the program must be efficient in terms of both storage and processing requirements. A mechanistic approach to reaction evaluation is particularly attractive in this sense since large numbers of reactions can be segregated into a relatively small set of mechanistic pathways. The program would have no knowledge of specific reactions such as aldol and Dieckmann condensations, Michael reactions, or α-alkylations. Rather, these and innumerable other, even unprecedented, reactions

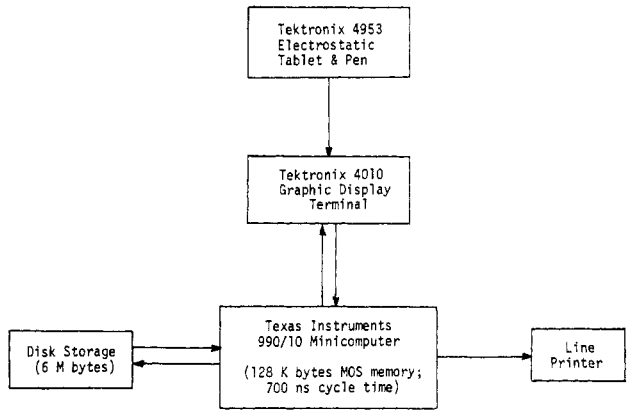


Figure 2. Hardware diagram of the computer system and peripheral devices.

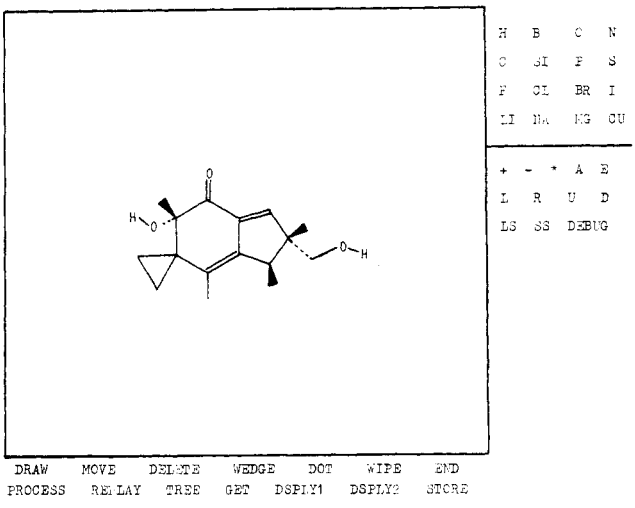


Figure 3. Primary menu and plotting box.

should be accessible through the application of sound mechanistic ideas.

Finally, predicted products must be reasonable. This necessitates some pre- and postmechanism screening for unstable functionality, excessively strained products, tautomers, etc.

There are three main aspects to the program: graphics, perception of structural features and reactive sites, and mechanistic evaluation. An overview of each area is presented below, concluding with examples of reaction sequences predicted by CAMEO. Details on specific procedures are deferred to a series of papers in preparation on the perception[13,14] and mechanistic[15] segments. A general schematic for the program organization and processing order is shown in Figure 1.

II. Graphics

A. Communication with CAMEO. The chemist/computer interface is similar to that of the LHASA program developed by Corey et al.[2b,12a] Reactant molecules are drawn by using an electrostatic graphics tablet and pen and are simultaneously displayed on a storage scope. Input is so closely aligned with drawing chemical structures on a piece of paper that little instruction is required for one to attain proficiency. A hardware block diagram is shown in Figure 2. The tablet is an extension of the CRT and

(13) W. L. Jorgensen and T. D. Salatin, in preparation.
(14) W. L. Jorgensen and B. L. Roos-Kozel, in preparation.
(15) W. L. Jorgensen and T. D. Salatin, in preparation.

Table I. Memory Allocations in Bits for Atom and Bond Table Entries

atom or bond	property	bit allocation
ATOM	XLOC	10
ATOM	YLOC	10
ATOM	TYPE	6
ATOM	CHARGE	2
BOND	ATOM1	5
BOND	ATOM2	5
BOND	ORDER	2
BOND	STEREO	2

has no direct communication link with the computer itself.

All drawing and related operations are performed directly by the chemist via the primary menu and plotting-box display (Figure 3). Operational modes are activated by pressing the pen, which is tracked on the screen by a blinking cursor, within the word representing the desired mode. For example, pressing the pen in the word DRAW sends coordinate information through the CRT buffer to the computer, thus allowing it to recognize that a molecule is to be input. Subsequently, pressing the pen within the plotting box creates an atom, and bonds are drawn to span each two successive atoms. Multiple bonds are created by sequentially pressing the pen on the atoms spanned by an existing bond. All atoms are initially considered to be carbon. Heteroatoms may be created by sequentially activating the atomic symbol desired and the atom whose type is to be modified. Hydrogen atoms are considered to be implicit by the program and are drawn in at the discretion of the chemist. A limited stereochemical package is now operative (vide infra), and stereochemistry may be imparted to a bond through the use of the WEDGE and DOT commands.

Since CAMEO has been written to perform reactions in the forward (synthetic) direction, there must be a way to indicate the reaction conditions desired. This is accomplished by activating the PROCESS mode and associated menus. In its present version, only basic and nucleophilic conditions are in force. If basic reaction conditions are activated, the chemist has the option of accessing a menu of frequently used basic reagents. Pressing the pen within the desired reagent will cause the computer to automatically add this fragment to the atom and bond table (vide infra), just as if it were drawn in from the tablet. This was designed to save user time and, in certain cases, decrease the effective size of the molecular ensemble to be processed. Obviously if an intermolecular reaction is desired, two structures must be recognized by the computer, whether they are input directly or via the reagent menus.

Structural modifications are made through use of the MOVE (atom), DELETE (atom or bond), and scaling (LS and SS) modes. The entire plotting box and any associated structures in the present reaction scheme may be erased with a WIPE operation. A synthetic TREE which records by number all reaction products and their genealogical relationships is also displayed on request. Any member of the tree can be retrieved for further processing by pointing to its number with the pen.

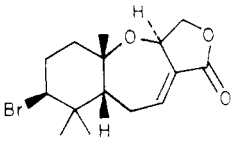
B. Internal Molecular Representation. Dual-connection tables (the atom and bond tables) are used to provide an internal representation of substrate and reagent molecules. This concept was first advanced by Corey[2c,12] and has been modified for our use to provide an efficient method of accessing fundamental atom and bond data. Table I illustrates properties included in the table, together with a list of necessary bit string lengths for each entry. Every atom entry requires two computer words (each 16 bits in length), 20 bits of which consist of graphic coordinate information, (x_i, y_i), while each bond requires one word of memory. Every atom and bond is assigned a unique number for reference. Two words having bits set corresponding to used atom and bond numbers, as well as words containing the total numbers of atoms and bonds, are also stored for each structure. Space for 63 atom types has been provided due to the ever increasing variety of organic reagents. Retrosynthetic programs require fewer types since reagents are not always displayed specifically. Charges are represented by values ranging from 0 to 3 (neutral, cation, anion, and radical, respectively).

The atoms which are spanned by each bond are included in the bond word as they are drawn. Wedged stereochemistry is designated as to whether the wedge is entered aligned with, or in opposition to, the bond-entry packing order. Dotted bonds are handled analogously and are directional in CAMEO.

Molecules containing up to 32 explicit atoms and 32 bonds (the number of bits in a double-precision computer word) are allowed in any one structure ensemble. The effective number is larger through the use of "superatoms". These are groups of atoms which may be manipulated as a unit such as the tosylate (OTS) and lithium cuprate (CULI) superatoms. So, through use of the phenyl superatom (PH), atom table requirements for the aldol condensation between acetophenone and benzaldehyde with sodium hydroxide as the base are reduced from 21 explicit atoms to 11. Naturally, the user must be confident that the superatom remains intact under the reaction conditions. With this feature and implicit hydrogens, sizable natural products such as steroids, alkaloids, and prostaglandins may be processed.

III. Perception

Before any chemistry can be performed, the computer must recognize important structural features and reactive sites in the substrates. This is accomplished in a stepwise manner by first accessing the atom and bond table to establish fundamental atom and bond sets for types, neighbors, and connectivity. Each set is a 32-bit double computer word with bits turned on (binary 1) for the index numbers of the atoms and bonds displaying the property of the set.[2c,12] For example, three bits would be on in the oxygen set for aplysistatin shown below, two bits would



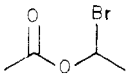
be on in the set for double bonds, etc. The use of sets permits facile application of Boolean operations to build more complex and often more useful sets. For example, the expression (CCBNDS.AND..NOT.BONDS1) yields the set of all carbon–carbon multiple bonds. The neighbor sets which record the atoms and bonds adjacent to each atom and bond in the structure are heavily used throughout the program. With set perception as a basis, more sophisticated recognition processes can occur for rings, functional groups, stereochemistry, and reactive sites as discussed individually below.

A. Functional Groups. The perception of functional groups is critical since they are the centers of synthetic activity. Two approaches to the problem can be identified. The first involves the search of a substructure list. Occurrence of a particular fragment keys further questioning which ultimately leads to the identification of specific functional groups. This avenue is the most popular, and

interesting data-table-driven procedures have been devised.[12,16-18]

The alternative is to use an encoding scheme which yields a unique descriptor for each functional-group type that is easily compared to a reference list. Esack and Bersohn have reported an algorithm which uses a canonical connection table to compile FG machine formulas as the sum of constituent-atom formulas.[19] This problem has been reinvestigated, and a new, efficient, general algorithm based on a unique index is outlined here.

One of the primary questions which must be addressed is where functionality begins and ends. In CAMEO, a functional-group origin is defined as any carbon in a carbon–hetero bond and any carbons in carbon–carbon multiple bonds. The extent of the group is determined by growing paths from the origins until either a carbon–carbon bond is encountered or all atoms in the path have been found. If geminal functionalities are not recognized immediately, then separate paths are grown from the origin and the constituent groups are identified separately. For example, the path growing from either carbon origin of the bromo ester below would establish the two origins, the two



oxygens, and the bromine as part of the functional group. Its index would not be recognized and separate path growing would commence from the non-multiply-bonded origin. The ester and bromide are then separately recognized. CAMEO presently recognizes approximately 100 types of functional groups.

Sixteen-bit (one word) codes are compiled for each group on the basis of seven properties of its constituent atoms and bonds. These properties are defined as follows.

- TSUM The sum of the atom types of the atoms comprising the group minus 4 (since any group will have at least one carbon origin); 5 bits
- BSUM The sum of the orders of the bonds spanning atoms in the group minus 1 (since all groups will contain at least one bond); 3 bits
- CHG Set to one if any atoms in the group are charged; 1 bit
- RING Set to one if a group origin is in a ring; 1 bit
- CBNO The number of carbon origins in the group minus 1; 2 bits
- NTCH The number of carbons attached to a multiply bonded origin or to an origin bonded to two oxygens; 2 bits
- HHBD Set to one if a hetero–hetero bond is present in the group; 1 bit

Hydrogens are not included in FG codes, since their presence is optional. The NTCH parameter distinguishes ketones and aldehydes, hemiketals and hemiacetals, etc. The RING bit permits lactones and lactams to be recognized. If a functional group is not identified with the ring bit on, the ring bit is turned off and comparison is made with the reference list again. This is also done with the charge bit. Charged intermediates occur frequently during processing and it is convenient to recognize, for example, an alkoxy group as an alcohol. The only special cases are, as usual, epoxides, episulfides, and aziridines. After they are recognized as cyclic ethers, sulfides, and amines, a check of ring size is required.

Functional group numbers (ca. 1–100) are arranged according to their electronic properties as withdrawing (W), neutral (N), or donating (D). In this way groups of a particular class may be accessed by using only their respective FG numbers. This is of great utility during pK_a and leaving-group perception (vide infra). Extended conjugation is recognized for W and D groups. This substructure is reflected in the FG number without destroying the integrity of the base value by setting bits 9 or 10 for a conjugated W or D group, respectively. This means that the α,β-unsaturated lactone in aplysistatin has a FG number of 13 (ester) + 256 (vinylw). Bits denoting conjugation may be masked to recover the original group number. Origins of these groups are the union of the original-group origins and those of the unsaturated bonds making up the chain of conjugation.

The atoms, bonds, origins, and FG numbers of each functional group are stored in arrays for later reference. Up to 12 explicit groups in addition to carbon–carbon multiple bonds may be present in any reactant ensemble.

B. Rings and Aromaticity. Algorithms for the detection of rings and aromaticity have been reported.[20] New procedures have been implemented in CAMEO as will be presented elsewhere.[14] The information obtained in these routines yields the smallest set of smallest rings, aromatic rings, and sets of fusion, bridgehead, and spiro atoms and bonds. Aromatic tautomers are also recognized and tautomerized if appropriate.

C. Stereochemistry. The stereochemical capabilities of CAMEO are still being developed. Stereo definitions are made from the wedged and dotted bond information.[21] Axial and equatorial groups and cis–trans relationships about double bonds and rings are recognized. Stereochemical conversions in substitution reactions are treated. Future enhancements will focus on stereochemical effects in bridged and fused ring systems and possibly three-dimensional structural representations. In many areas, the forward processing creates programming challenges not found in the retrosynthetic projects. Fragment moving is necessary to display aesthetically acceptable product structures. The stereochemistry of elimination reactions, particularly in acyclic substrates, can also require substantial manipulation.

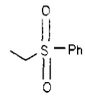
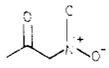
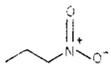
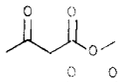
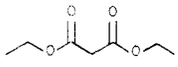
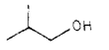
D. pK_a Perception. No program presuming to predict reaction products in a mechanistic fashion under basic conditions would be useful without a reliable algorithm for the identification of relative pK_a values. This makes the internal simulation of a proton-transfer step feasible. The fact that this step is usually the first in a base-catalyzed reaction further elevates its importance. To date no algorithm of this type has appeared in the literature.

CAMEO uses a data driven algorithm in which FG numbers stored as bits set in an array of acidity levels are used to identify the most acidic hydrogens (explicit or implicit) present. The current version of CAMEO recognizes 15 acidity levels ranging from hydrohalic acids (HCl, pK_a = −7) to simple alkyl groups (pK_a > 40). Thus, the average pK_a range for a level is about 3 units. However, more levels have been concentrated in the pK_a region most useful to the organic chemist (ca. 5–35).

Two parallel arrays are used in the routine to identify acidic protons. The first, KEYL, is searched level by level until a functional group is found which is present in the substrate. The second array, WD2ND, is then searched

(16) D. A. Pensak, Ph.D. Thesis, Harvard University, 1973.
(17) J. Figueras, *J. Chem. Doc.*, **12**, 237 (1972).
(18) M. Bersohn and A. Esack, *Chem. Scr.*, **6**, 122 (1974).
(19) A. Esack and M. Bersohn, *J. Chem. Soc., Perkin Trans. 1*, 2464 (1974).
(20) (a) E. J. Corey and G. A. Petersson, *J. Am. Chem. Soc.*, **94**, 460 (1972); (b) M. Bersohn, *J. Chem. Soc., Perkin Trans. 1*, 1239 (1973).
(21) E. J. Corey, W. J. Howe, and D. A. Pensak, *J. Am. Chem. Soc.*, **96**, 7724 (1974).

Table II. Examples of Compounds in the 15 Acidity Levels

level	example	level	example
1	HCl	9	
2	$PhCO_3H$	10	
3		11	
4	PhSH	12	LDA, Me_2SO
5		13	
6		14	
7		15	
8			

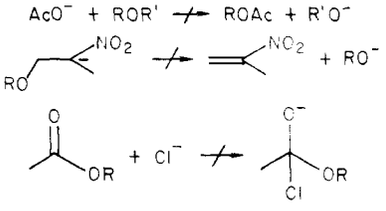
within that level to see whether another activating group is necessary to attain this pK_a range. Only two activating groups are allowed at this point, although extensions to this would be simple provided that pK_a values for the combinations of groups produced could be found. Table II lists examples of substrates that fall in each of the 15 acidity levels.

If conditions conducive to proton transfer prevail, it is important to determine whether or not the base used is strong enough to abstract any protons to a significant extent. Whatever line is drawn here will be somewhat arbitrary; at the present time, CAMEO allows proton transfer to occur unless it results in a thermodynamically uphill step of greater than one acidity level. This enables identification of the vast majority of reactions without burdening the program with unlikely intermediates.

E. Nucleophiles and Electrophiles. The fundamental process in ionic chemistry is joining nucleophilic and electrophilic sites. The recognition of potential nucleophiles is allied with pK_a perception. Any anionic sites obtained from proton transfer are deemed nucleophiles. For example, treating acetophenone with sodium ethoxide yields three nucleophilic sites: the alkoxide oxygen, enolate carbon, and enolate oxygen via resonance.[22] The alkoxide oxygen is retained since ethanol is within one acidity level of ketones. Treating the same ketone with LDA yields only the two sites on the enolate as nucleophiles. If no anionic sites are avilable, the central atoms in the following neutral groups are considered to be potential nucleophilic sites: amines, phosphines, sulfides, sulfoxides, selenides, and selenoxides. Neutral oxygen is not currently in this list, although oxonium ions can be formed with potent electrophiles. Conjugation as in enamines is also taken into account.

An electrophile can be an atom in an X—Y, X=Y, or X≡Y bond, where X and Y are carbon or heteroatoms and X can equal Y. This broad definition encompasses carbon–oxygen double bonds, carbon–halogen bonds, epoxides, sulfenyl halides, bromine, peroxides, etc. Associated with each electrophilic site X is the leaving group bond X—Y, X=Y, or X≡Y. A few restrictions are needed to make these definitions manageable. A measure of leaving-group ability is ascertained in CAMEO by evaluating an effective pK_a level for the leaving group Y^-, $X—Y^-$, or $X=Y^-$. For a few cases, like epoxides and acid halides, the effective pK_a is reduced from the pK_a of an alcohol to reflect the loss of ring strain and potential elimination of halide ion in these instances. Then, for any candidate bond only the electrophilic site with the better leaving group (lower effective pK_a) is retained; e.g., sulfur is the electrophile in phenylsulfenyl chloride. If there is a tie as in unsymmetrical peroxides, then two electrophilic sites are recognized. A further restriction is that C–C bonds are not considered as leaving groups except in doubly activated cyclopropanes[23] or in E1 eliminations breaking cyclopropyl or cyclobutyl bonds. Finally, conjugation is considered so carbons or heteroatoms β, δ, etc. to and conjugated with an electrophilic site are also electrophiles as in enones or allylic halides.

Aside from this pK_a analysis, the strengths of nucleophiles and electrophiles have received limited ranking (vide infra) since the implementation of solvent and counterion effects is still under development.[24] In the meantime, a ΔpK_a criterion has been found useful in evaluating the feasibility of substitutions, additions, and eliminations. Specifically, *the pK_a of the conjugate acid of the nucleophile or base should be no more than one level below the effective pK_a for the conjugate acid of the leaving group.* This procedure eliminates much naive chemistry such as the following examples.



However, the ΔpK_a rule has exceptions particularly for addition reactions in protic media. Kinetic effects such as anion trapping by proton transfer can occur in hydroxylic solvents. For example, thiophenoxide ion undergoes Michael addition with vinyl sulfones and cyanide ion adds to enones (hydrocyanation) in ethanol. These considerations are handled currently via the "protic" and "aprotic" solvent buttons on the reagent menu. Competition between substitutions and eliminations receives further screening as discussed below.

In closing this section, it is noted that full perception is performed on all reactants and products (Figure 1). This includes checks for valence violations, unstable functionality, Bredt's rule violations, strained trans fusion bonds, and tautomers. Offending structures are either rejected or converted to more stable forms.

IV. Mechanistic Evaluation

In the retrosynthetic approaches, reactions have been treated as transforms,[2c] matrices,[5] and classes of "half-reactions".[6] None of these directly involve individual reaction mechanisms. However, the internal simulation of reaction pathways appears as a viable and attractive alternative. A literature search reveals that a small number of general pathways are operative under basic conditions: proton transfer, metalation, addition, substitution, and elimination. It must be stated that the intent here is not

(22) A general routine for resonating charges has been developed. One notable feature is that a resonance structure is not retained (as a potential nucleophile) if it involves forming a 2p–3p π bond. This prohibits, for example, subsequent O-alkylation of an α-sulfonyl carbanion. Also, in symmetrical cases such as a nitro group or sulfinate anion, only one of the equivalent resonance structures is retained. Resonance, of course, does not occur if it renders a Bredt's rule violation.

(23) S. Danishefsky, *Acc. Chem. Res.*, **12**, 66 (1979).
(24) A. J. Parker, *Chem. Rev.*, **69**, 1 (1969).

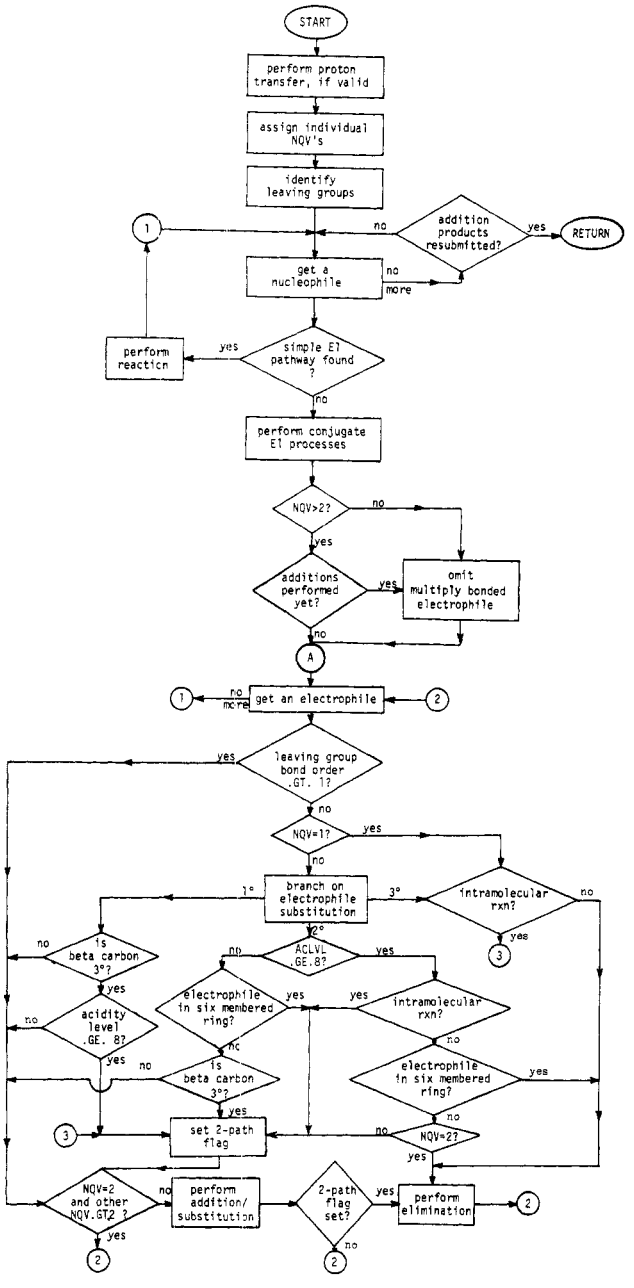
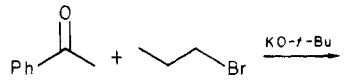


Figure 4. Flow chart for the mechanistic executive in CAMEO.

to delve into many of the current discussions surrounding the actual intermediates and transition states through which these reactions run their course. Rather, it is desired to use a fundamental, textbook-type outlook, so that reactions can be used in a systematic manner capable of being handled efficiently by a computer. It seemed reasonable that the large amount of data in the literature would reveal patterns for changes in rates and mechanisms with changes in molecular structure and conditions. Procedures could then be devised to predict viable products for a majority of reactions in the class. The treatment of base-catalyzed and nucleophilic reactions in CAMEO is outlined below and corresponds to the box on evaluation of pathways in Figure 1.

A. Base-Catalyzed Chemistry. Figure 4 shows the mechanistic logic through which CAMEO predicts reaction products. The flow chart may be clarified by following the processing of a specific reaction such as the alkylation of acetophenone shown below.



As discussed above, pK_a considerations yield three nucleophilic sites, two from the enolate and the alkoxide oxygen. The potential electrophilic sites are recognized as the carbonyl and bromide carbons. A structure is stored corresponding to each nucleophile. Qualification values (NQV) from 1 to 4 are assigned to each nucleophilic site in order to categorize its reactivity pattern. A value of 1 or 2 identifies elimination as the dominant pathway with an NQV of 1 representing virtually total elimination. Strong bases that are sterically hindered are in these categories, e.g., DBU and LDA have an NQV of 1 and *tert*-butoxide ion has an NQV of 2. Nucleophiles in class 3 are weaker, unhindered bases that prefer addition or substitution, e.g., thiophenoxide and chloride ion. Nucleophiles with an NQV of 4 may participate in substitutions, additions, or eliminations. These include strong unhindered bases such as enolate ions and *n*-alkoxide ions. It should be recalled that proton transfer has already been considered before this point. Whether any or all paths allowed by the NQV occur is dependent on the available electrophiles.

Although not a factor in the present example, there exist instances where proton transfer is not a valid first step. When lithium is used as the counterion, equilibration among proton-transfer structures may be retarded to the point that addition to polarized multiple bonds dominates the reaction. Thus, addition of alkyllithium reagents to ketones and esters is of great synthetic utility. If proton transfer were allowed to occur initially in these cases, the nucleophile performing the addition would be destroyed. However, a base such as LDA will not participate in addition due to its steric bulk. Counterion effects are hence of importance in any reaction analysis. These phenomena are currently taken into consideration by CAMEO for lithium and magnesium and further work in this area is continuing.

Since large rate enhancements are shown for simple E1 elimination processes,[25] these are reviewed prior to other competing pathways. The ΔpK_a rule is used in selecting leaving groups. If such a process is found, its resulting products are formed, and no further mechanistic processing takes place. Conjugate E1 processes are considered competitive with other reactions.

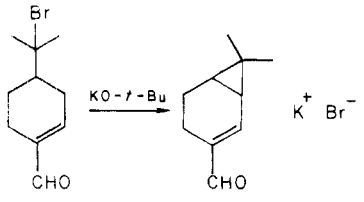
Rate data for conjugate E1 processes (e.g., Grob fragmentations) are rarer and in cyclic systems are often dependent on antiperiplanarity of the intervening bonds. At this time, CAMEO does not have the capability to test for this condition. In view of these shortcomings, E1 eliminations of path greater than two are allowed somewhat indiscriminantly, relying on a postmechanism filter to remove undesirable structures. Paths of up to six atoms between nucleophile and electrophile are permitted. In the alkylation of acetophenone, no E1 processes are possible. Competition among additions, substitutions, and E2 eliminations is then considered.

The formal similarity of additions and substitutions is somewhat fortuitous from a programming point of view. They are subdivisions of a general mechanism in which the incoming nucleophile becomes bonded to atom Y, the order of the Y–X bond is decreased by one, and two electrons shift from the nucleophile to X. Thus, both reaction paths may be handled by a single routine which performs this operation. Heuristics reflecting reactivity differences between the two are applied before the actual mechanism is performed.

(25) G. Biale, D. Cook, D. J. Lloyd, A. J. Parker, I. D. R. Stevens, J. Takahashi, and S. Winstein, *J. Am. Chem. Soc.*, 93, 4735 (1971).

Competition between S_N2 and E2 pathways constitutes the majority of heuristics used in CAMEO at this time. Either process may completely dominate the other depending on the nucleophile and electrophile. The ΔpK_a and protic/aprotic solvent considerations are taken into account in selecting potential leaving groups for a nucleophile. Furthermore, yields of substitution products decrease with increasing hindrance of the electrophile in the order primary > secondary > tertiary. This is exemplified by the fact that ethyl *tert*-butyl ether must be prepared from an ethyl halide and *tert*-butoxide anion; reaction of *tert*-butyl halide with ethoxide results in total elimination. Nucleophile NQV's and base strengths as well as electrophile substitution are used in pathway evaluation. Molecules may undergo (1) addition or substitution, (2) elimination, or (3) both processes resulting in two products as indicated in Figure 4.

Some comments on details in the flow chart are in order. The propensity of bases of NQV 1 (e.g., LDA) and of tertiary electrophiles to participate in elimination reactions may only be abated if the reaction is intramolecular. Thus, the cyclopropane annelation shown below is feasible.[26]



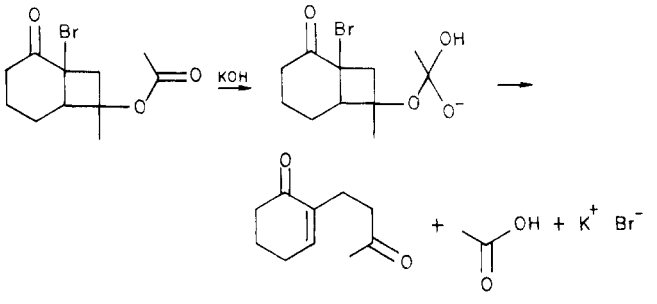
Conversely, primary electrophiles give high yields of substitution products. Elimination is important only when the electrophile has a tertiary β-carbon and a strong base is used. The cutoff value is now set at acidity level 8 and includes bases of pK_a ca. ≥ 16. This situation occurs in the reaction of isobutyl bromide with $NaOC_2H_5$, which gives 60% elimination at room temperature.[25]

The fate of secondary electrophiles is strongly dependent upon the strength of the base. Figure 4 shows that, in most cases, substitution and elimination are both important processes, with some exceptions. A nucleophile having an NQV of 2 will give elimination with a secondary substrate. A six-membered ring tends to increase elimination, as does a tertiary carbon β to the electrophile. The former situation leads to total elimination with a strong base and competing reactions with a weakly basic nucleophile (e.g., Ph_3P). Hence, cyclohexyl bromide gives about 50% elimination with sodium thiophenoxide and total elimination with $NaOC_2H_5$.[27]

Each nucleophile is tested with all candidate electrophiles before the next nucleophilic site is processed. In the present example, the enolate is recognized as a strong base and a good nucleophile. The environment of the electrophilic site is found to be primary with no tertiary carbon β to the bromine, so substitution will predominate. The two resonance structures then lead to the C- and O-alkylated products. Control is passed to subprograms which aesthetically position the two fragments and perform the bond making and breaking. The modified atom and bond tables are stored in a disk file.

This leaves the *tert*-butoxide anion as the only remaining nucleophile to be tested. Although reaction with the carbonyl group is formally possible, the steric bulk of the nucleophile precludes this pathway. Since the NQV of *tert*-butoxide is 2, the carbonyl group is omitted from the list of electrophiles to be searched. The flow diagram shows that for unhindered *n*-propyl bromide even a bulky base may participate in a substitution route. However, the low steric accessibility of this nucleophile will cause substitution to be slow with respect to a highly nucleophilic species. For this reason, the *tert*-butoxide anion is not allowed to react in an S_N2 process in the presence of the more reactive α-keto carbanion (NQV = 4). Thus, *tert*-butyl propyl ether is not returned as a reaction product in this case.

After all nucleophiles have been processed, products of addition reactions are allowed to participate in a subsequent E1 or S_N2 process. Thus, the cyclobutane ring cleavage reaction[28] shown below is predicted in only one pass through the program.



A simple E1 step also often completes the formation of various other carbonyl functions such as esters, amides, lactones, etc.

The alkylation products of the example do not involve addition reactions, so the mechanistic module is exited. Products are then reperceived to check for unstable functionality, tautomers, valence violations, and highly strained structures before display on the CRT.

This example shows the adaptation of heuristics created from structure–reactivity correlations obtained from a survey of existing literature data. The heuristics originate from four fundamental areas of concern: (1) nucleophilicity, (2) base strength, (3) leaving-group ability, and (4) steric accessibility of both the substrate and base. Refinement of the mechanistic schemes and the search for organizing principles continues. The incorporation of valuable prescriptions such as the Woodward–Hoffmann rules and Baldwin's rules is planned.[29] The accurate screening of pathways is key to the sophistication of the program. The current philosophy is to err on the side of leniency rather than miss a viable product.

Since CAMEO basically pushes electrons in one direction at a time, some reactions possessing multistep mechanisms or requiring addition of greater than 1 equiv of a reagent yield intermediates which must be resubmitted before the actual products are created. This potential drawback also has an advantage, however, since the course of a reaction whose mechanism is obscure may be clarified by the output of intermediate structures.

B. Sample Sequences. CAMEO has been used to predict the products for hundreds of organic reactions. Sophisticated predictions are often made and the program is particularly adept at pointing out potential side products. Output for several sequences including some related to existing natural-product syntheses is shown in Figure

(26) G. Büchi, W. Hofheinz, and J. V. Paukstelis, *J. Am. Chem. Soc.*, 88, 4114 (1966).
(27) D. J. McLellan, *J. Chem. Soc. B*, 705 (1966); D. Cook, A. J. Parker, and M. Ruane, *Tetrahedron Lett.*, 5715 (1968).
(28) N. R. Hunter, G. A. Macalpine, H. J. Liu, and Z. Valenta, *Can. J. Chem.*, 48, 1436 (1970).
(29) (a) R. B. Woodward and R. Hoffmann, "The Conservation of Orbital Symmetry", Academic Press, New York, 1970; (b) J. E. Baldwin, *J. Chem. Soc., Chem. Commun.*, 734 (1976).

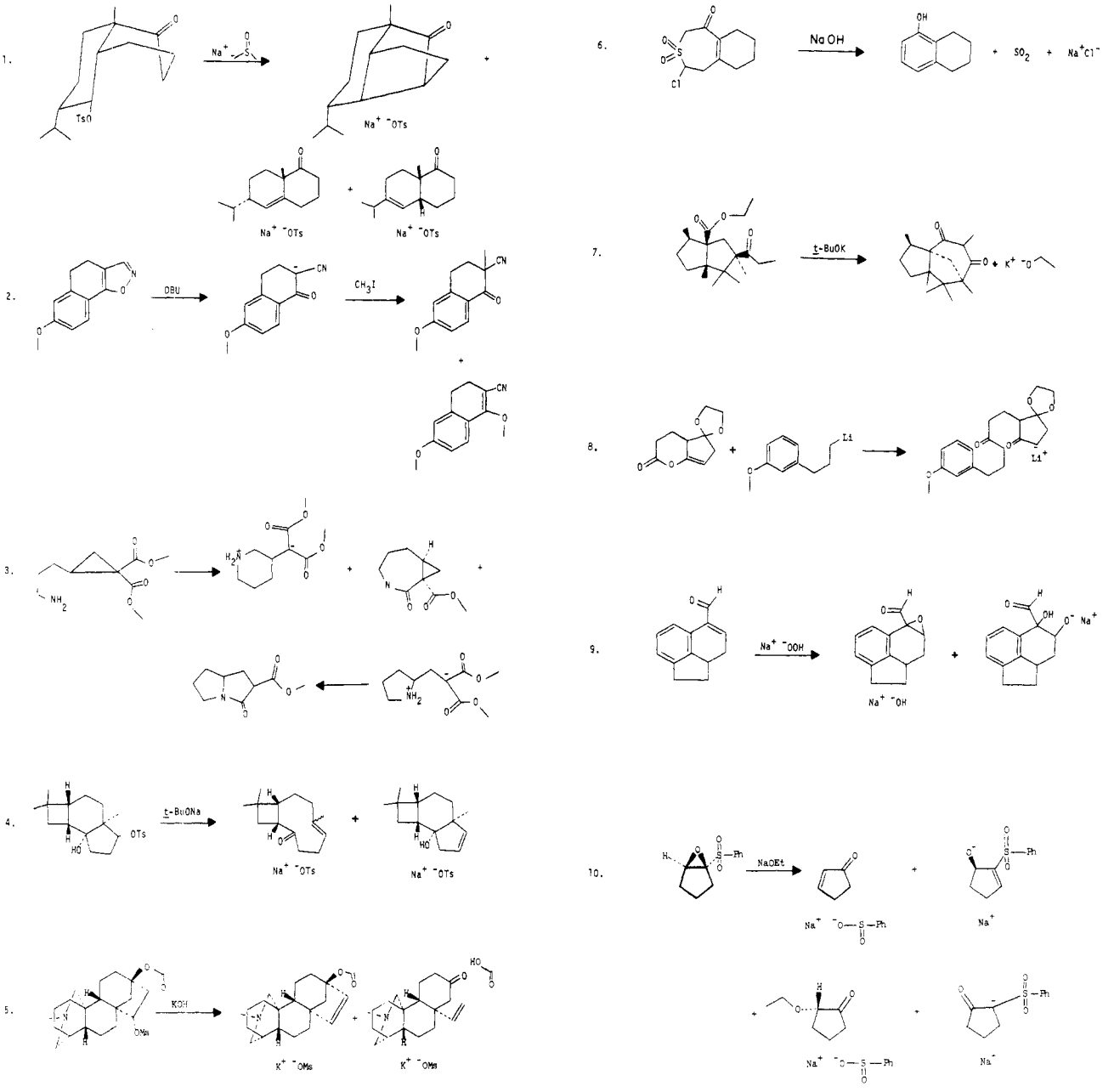


Figure 5. Representative synthetic sequences predicted by CAMEO.

5. Some comments are in order explaining the screening of potential products. Example 1, a key step in the synthesis of sativene,[30] shows both a strength and a weakness of CAMEO at the present time. While the O-alkylation product is deleted because it would violate Bredt's rule, the steric inaccessibility of the enolate to the protons which would be abstracted in the E2 reactions is not gauged. A reaction in which both S_N2 and E2 processes are feasible is depicted in example 10. Note that an α-alkoxy sulfone is deemed unstable and elimination occurs during postmechanism perception. Example 2 shows CAMEO's ability to handle functionality which it does not explicitly identify. Although an isoxazoline is not specifically recognized, the N–O bond is identified as a potential leaving-group bond spanning an enol and an imine.

As mentioned earlier, doubly activated cyclopropanes[23] (example 3) may act as Michael acceptors as may other conjugated functionality (example 9). Example 3 also illustrates lactam formation under neutral conditions. Fragmentation reactions may take place directly (example 4)[31] or after an initial addition (example 5). In the latter case, no substitution is allowed on the neopentyl-like mesylate, and no reaction may involve the bridgehead carbon as an electrophile. If a product is an aromatic tautomer (example 6), the more stable form is output. Episulfones are recognized as unstable in the postmechanism screening. Creation of aromaticity in a product will cause others to be deleted. Example 7[32] shows a standard condensation. Finally, after the Michael addition in example 9, in addition to the 3-exo-tet[29b] path leading to epoxide formation, there is also a 4-endo-tet path leading to the diol. The program will eventually have to recognize the steric problems in the substitution leading to the latter product. These transformations and unlimited variations involving alternate nucleophiles, electron-withdrawing groups, and leaving groups are all handled in a consistent

(30) J. E. McMurry, *J. Am. Chem. Soc.*, **90**, 6821 (1968).

(31) E. J. Corey, R. B. Mitra, and H. Uda, *J. Am. Chem. Soc.*, **85**, 362 (1963); *ibid.*, **86**, 485 (1964).

(32) G. Stork and F. H. Clark, Jr., *J. Am. Chem. Soc.*, **77**, 1072 (1955); *ibid.*, **83**, 3114 (1961).

and efficient manner by the mechanistically based program.

V. Conclusion

A program, CAMEO, is being developed to predict the products of organic reactions. Reactant molecules are input via a graphics tablet and CRT, through which all chemist-computer communication takes place. Routines for the perception of sets, rings, functional groups, stereochemistry, aromaticity, and acidities have been written. This information provides the foundation for the internal simulation of basic reaction mechanisms, utilizing general structure-reactivity correlations to control program flow among competing pathways. Products are output on the CRT, where the chemist can select, modify, and resubmit structures. Repetition of the procedures causes multistep reaction sequences to be created, which are recorded in a synthetic tree.

One of the advantages of the program's design is that new, mechanistically sound reactions may be discovered without specific programming. Work is currently taking place to expand both the number of reaction classes treated and the heuristics used in directing flow among the pathways involved.

Acknowledgment. The original equipment purchases for this project were made possible by grants from the Du Pont Young Faculty fund and Research Corporation. Acknowledgment is made to the donors of the Petroleum Research Fund, administered by the American Chemical Society, for partial support of this work. Gratitude is also expressed to the Dreyfus and Sloan Foundations for assistance. We are indebted to Barbara L. Roos-Kozel for developing the ring and stereochemical perception routines in CAMEO. Synthetic consultation with Professor P. L. Fuchs has been most valuable.

Computer-Assisted Synthetic Analysis. Techniques for Efficient Long-Range Retrosynthetic Searches Applied to the Robinson Annulation Process

E. J. Corey,* A. Peter Johnson,[1] and Alan K. Long

Department of Chemistry, Harvard University, Cambridge, Massachusetts 02138

Received November 23, 1979

One of the major fundamental problems in computer-assisted synthetic analysis is the development of methods for conducting goal-driven, deep (or long-range) search. Among the many problems associated with this approach is the avoidance of unfruitful lines of analysis which not only slow the problem-solving process but inundate the user with an unacceptable number of possibilities. Since chemists face similar difficulties during their attempts to plan syntheses, progress with the computer-assisted approach can provide perspective. The newest developments in multistep analysis in the Harvard program LHASA are illustrated for the Robinson annulation search. LHASA treats a target structure systematically, examining each of the 12 possible positions for an α,β-unsaturated ketone in each six-membered ring. Sequences of retrosynthetic steps called "procedures" are applied to remove obstacles blocking performance of the key simplifying disconnection, the Robinson annulation transform. A new technique for preevaluation of these procedures is used to ensure that only the sequences most likely to succeed in the laboratory are actually displayed to the chemist. Several chemical examples are shown.

One of the most powerful strategies available for simplifying the analysis of a complex synthetic problem is the key-reaction-based strategy. This strategy depends on the selection of an important reaction to construct a crucial section of the synthetic target. After application of the key reaction, a series of reactions are chosen which convert the functionality and structural features of the key-reaction product into the corresponding features in the target molecule.

In antithetic (or retrosynthetic) analysis[2] an analogous strategy is correspondingly effective. After selection of a key transform (or retroreaction), structural features in the target molecule are correlated with those required by the transform (that is, the features characteristically produced by the corresponding key reaction). Finally, "subgoal" transforms are selected which remove (antithetically) features deleterious to the application of the key transform and introduce features required by that transform.

This strategy has been employed at two levels in the LHASA program for computer-assisted synthetic analysis.[3] At the first level a number of structurally simplifying transforms have the ability to request other transforms which either exchange one functional group for another (functional group interchange, or FGI)[4] or introduce a desirable functional group (functional group addition, or FGA) and so pave the way for the operation of the main transform. At a higher level, certain powerfully simplifying (or key) transforms, for example, the Diels-Alder,[5] Robinson annulation, and halolactonization[6] transforms, have much more extensive subgoal capabilities. The search procedures driven by these transforms can generate retrosynthetic sequences of up to 20 steps in order to set up the structure required for valid operation of the key transform.

Machine application of these latter transforms uses a data-driven, binary search technique[5] to identify and re-

(1) Department of Organic Chemistry, The University, Leeds LS2 9JT, England.

(2) Corey, E. J. *Q. Rev., Chem. Soc.* **1971**, *25*, 455.

(3) For recent LHASA publications, see: Corey, E. J.; Long, A. K. *J. Org. Chem.* **1978**, *43*, 2208.

(4) Sequential FGI's of up to four steps are currently available in LHASA. See: Corey, E. J.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1976**, *98*, 203.

(5) Corey, E. J.; Howe, W. J.; Pensak, D. A. *J. Am. Chem. Soc.* **1974**, *96*, 7724.

(6) Corey, E. J.; Long, A. K.; Mulzer, J.; Orf, H. W.; Johnson, A. P.; Hewett, A. P. W., in preparation.

0022-3263/80/1945-2051$01.00/0