The Performance of a Noninteractive Synthesis Program

M. TAKAHASHI, I. DOGANE,* M. YOSHIDA, H. YAMACHIKA, and T. TAKABATAKE
Sumitomo Chemical Company, Organic Synthesis Research Laboratory, Tsukahara, Takutsuki-shi 569, Japan

M. BERSOHN
Department of Chemistry, Lash Miller Chemical Laboratories, University of Toronto, 80 St. George Street, Toronto, Ontario, Canada M5S 1A1

Received May 19, 1990

The Sumitomo Chemical Company has begun a long-term project to build a program that proposes synthetic routes for organic chemicals. Preliminary results from the program indicate that the interactive method is feasible and preferable because of the superior speed of the computer in decision making. The routes proposed are generally quite reasonable; the efficacy of the program is limited by the fact that there are only about 2000 reactions in this early version.

INTRODUCTION

The performance of a noninteractive computer synthesis program is discussed. The input to the program is the structure of the goal molecule and a set of extra information which is interpreted by the program as instructions for navigating through the synthetic tree. This extra information may include the limit on the number of steps for an allowed route, a limit on the number of nontrivial steps in the route, a limit on the complexity of the intermediates involved at any depth x in the tree, a limit on the number of reactions that do not build the skeleton of the molecule, such as functional group change only reactions, protection–deprotections, etc. Further, we typically provide many kinds of restrictions as to the structural complexity of an acceptable starting material for the route. We can also specify bonds of the goal molecule which must not be made, reaction conditions which are unacceptable, e.g., on the grounds of safety or cost, the number of allowed resolutions along any path, etc.

The general results of the program are that the thoroughness of the exploration exceeds that possible for an interactive program by hundreds of times. We can usually obtain a finite, manageable number of synthetic route proposals from the program. However we cannot be assured, in general, of duplicating a known route. The reason is the vastness of the synthetic tree. Even generating half a million reactants will only give us a small portion of the synthetic possibilities for a goal molecule of any structural complication. So we cannot be certain, a priori, that we will traverse the particular subtree which contains a known synthetic route. The same program with different constraints proposes different routes, and so might a chemist without the use of a computer.

At the current early stage in the development of this program, a typical result of using the program is to find the same route that is being used presently to manufacture the substance or else to find an inferior one. It is not common to find a superior route because the repertory of synthetic reactions inside the program is at present much smaller than that available to a chemist searching among the public sources.

DISCUSSION

A summary of the early work on organic synthesis programs is given in ref 1. A salient aspect even of the earliest work was the contrast between two approaches. One was the interactive approach, using a partnership between the program user and the program; the other was the noninteractive approach, in which the work of the search was done by the program without reference to comments from the user after the run had begun. In the latter case, general instructions can be issued with the input data to the program. Some held that a computer was not capable of being programmed to replicate a chemist's judgement or only capable of this in certain limited cases. In our work we have not come across any case where the judgment of a chemist could not be programmed. To be sure there are terms which lack accurate definition, such as "steric hindrance", the more so as we do not know the exact geometry at the transition state and often do not know the geometry of the molecule in the ground state if it is a molecule not yet prepared. Such terms are vague in their usage by chemists and, correspondingly, are also not concrete in the program. A site which has more quaternary neighbors than another site is seen by the program as being more hindered. If the sites are equal with such a measure, then the degree of substitution of the central carbon as well as the degree of substitution of the neighbor atoms must be considered. In contrast to the vague terms, terms which are definite in the chemist's usage have always also been made unambiguous inside of the program. For example, the program is aware of pK_a values and presumes that when one mole of base is used the most acidic site will alkylate, whereas when more than one mole is used the least acidic site that loses a proton will react preferentially.

It has variously been stated that the judgment of the chemist at the time of running a synthesis program is necessary to rescue such a program from the combinatorial explosion, i.e., that it is impossible for an unaided program to find its way among the host of inferior routes to select some or all of the few really good ones. Carefully examined this argument implies that there exist useful selection criteria among routes which cannot be articulated, i.e., which cannot be logically and clearly stated. Since such criteria could not be described in the literature and could not be explained clearly, their nontransferability would make their existence, if any, quite temporary. We believe that whatever chemical reasons there may be for regarding one incomplete route as more promising than another can be stated clearly and that whatever can be stated clearly can be programmed. This has been our consistent experience. Our program, called SYNSUP-MB, has been under development at Sumitomo Chemical Company since 1984. It is the intellectual descendant of a LISP program.[2] We present some sample problems and solutions with comments to show the rules used by the program for pruning the decision tree.

SAMPLE RESULTS FROM THE PROGRAM

2-Chloro-3-methyl-5-nitrotoluene. The synthesis of this compound presents orientation difficulties as the nitro group is meta to the most powerful ortho-para directing group in the molecule. Further there is the steric difficulty of 1,2,3-trisubstitution.

The number and the nature of the routes proposed varied with the user's instructions at input. Not surprisingly, the more simple the structure demanded of any acceptable starting

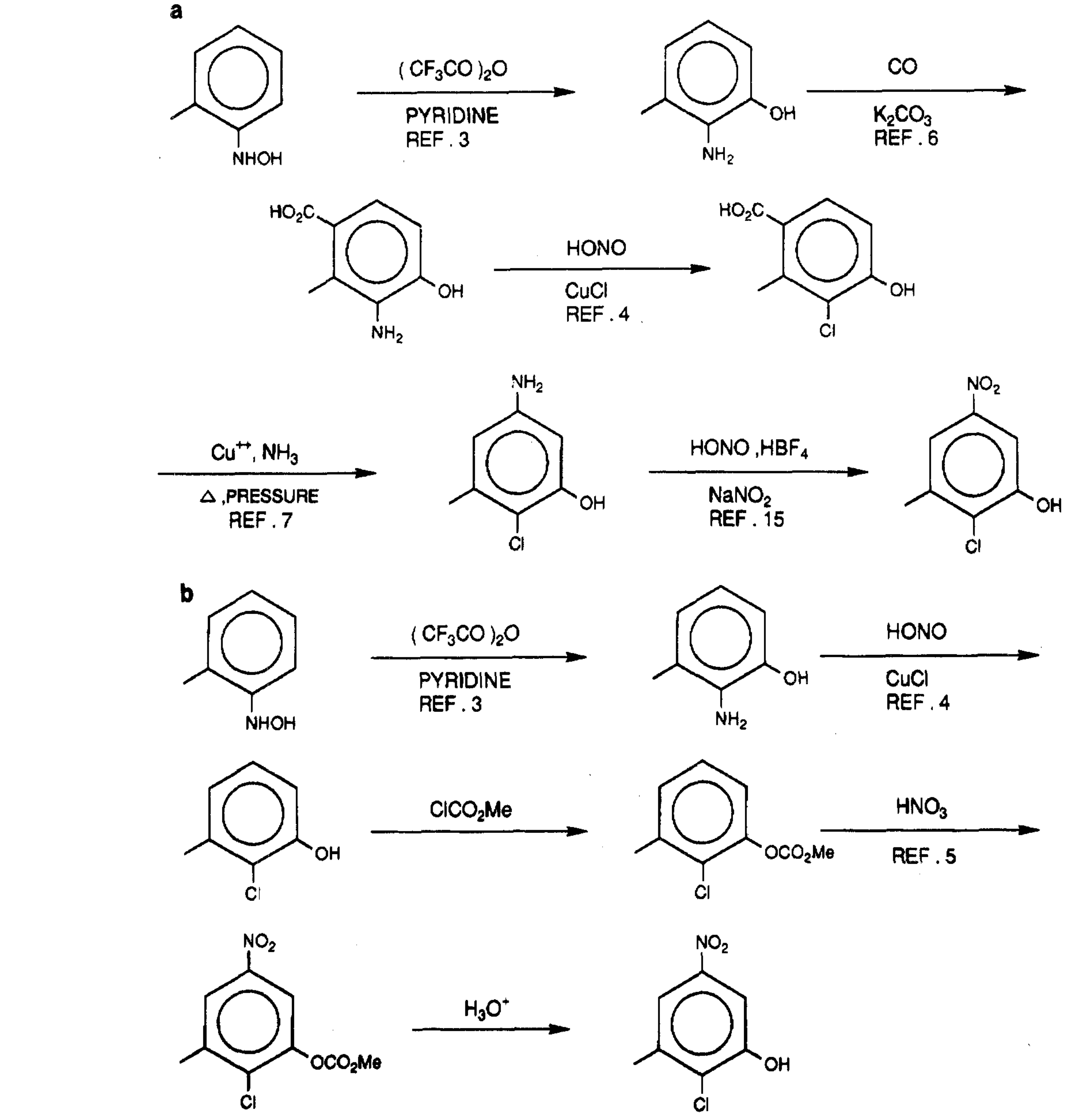

Figure 1. Five-step routes to 2-chloro-3-hydroxy-5-nitrotoluene.

material, the longer the routes had to become. On the first trial our input instructions specified that:

1. An acceptable starting material had to have no more than one functional group and no more than two substituents in an aromatic ring.
2. All reactions must be regiospecific, to avoid having to separate ortho and para isomers.
3. The synthesis could not be longer than four steps.

With these input instructions no routes were obtained. When the limit on the number of steps was changed to five and no other changes were made, the two routes shown in Figure 1 were proposed. The tree explored contained 4456 nodes, i.e., reactions were simulated and reactant descriptions generated a total of 4456 times.

We comment that the reaction of carboxylation of a phenol discovered by Yasuhara is paraspecific.

The program's basic routine is to examine the structural description of the current molecule and from this description to generate the description of the structure of the reactant(s) in a promising reaction to produce the current molecule. To shorten the verbiage we will frequently write, e.g., that "acetone was generated", meaning that a description of the molecular structure of the reactant acetone was generated.

Routes involving nitration of *o*-toluidine or acetylated *o*-toluidine were not proposed because the input instructions demanded regiospecificity and this nitration was not considered to be regiospecific. In the program, regiospecific electrophilic aromatic substitutions are considered to be unique reactions, are separately retrievable, and are labeled as regiospecific. We are not aware of a method for paraspecific nitration of anilines; consequently, the program also has no special note of this.

Phenols can be converted to aryl chlorides by the reagent Ph_2PCl_3 at temperatures of 140 for sites activated by withdrawing groups and at 220 with somewhat lower yields for a hydroxyl group at unactivated sites.[3] When we added this reaction to the program and reran the problem, the program proposed the three-step synthesis shown in Figure 2.

Since the paraspecific nitration of phenols is achievable[9] there were only two positions for the nitro group to enter the ring and one of these was totally unhindered and judged to be most favorable. In the following step, the activated hydroxyl, para to the nitro group, was deemed by the program to be able to be replaced preferentially. Experiment might show otherwise.

Thiophene Saccharine. In this case there is a known synthesis, and the aim was to find a cheaper way to make the

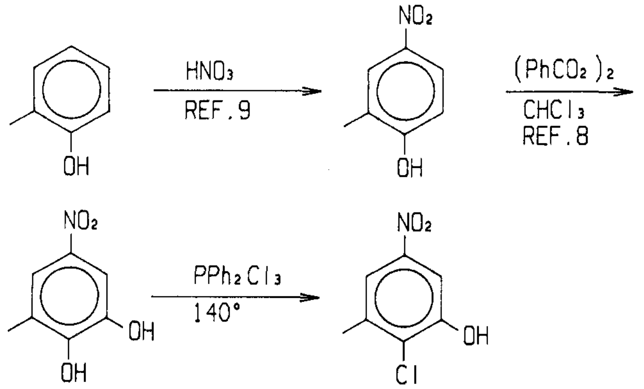

Figure 2. Three-step route to 2-chloro-3-hydroxy-5-nitrotoluene.

compound. This aim was not achieved. The best route found by the program was the existing known synthesis. The input specified that the starting materials should contain no more than four carbon atoms and the number of steps should not exceed eight. About 162000 reactions were simulated in the course of the search. When the limit on steps was seven, no routes were found. The reader will note that the dimeric intermediates are substantially more complicated in structure than is the goal molecule. This is a crucial matter. If we permit an arbitrary increase in complexity as we descend the tree, then it will become impossible to find synthetic routes. So there has to be a limit on the increase in complexity as we descend the synthetic tree. On the other hand, as this route shows, there must be exceptions. We make one exception for dimers, in which case the equivalent functional groups are only counted once. We must make another exception for reactions that remove groups previously needed for protection or stereocontrol. The atoms of these groups must not be considered in any assessment of the number of chiral centers or number of functional groups, etc. But there are numerous reactions in the program which remove complexity, such as decarboxylations, reduction of carbonyl or dichloromethylene to methylene, reverse Diels–Alder, etc. The reactant in each such reaction is in some way more complicated than the product so if we are too strict in our prohibitions, routes involving several of these reactions will automatically be ruled out. There is a trade-off here which is managed by the user at input time. (See Figure 3.)

A Chiral Cyclic Carbonate. Here the search was restricted by requiring a particular starting material. This implied that the program could not propose any route which makes carbon–carbon bonds. The step limit was set at four, but a three-step synthesis was found as shown in Figure 4. The ketone was noted as aryl–alkyl and the enantioselective reagent

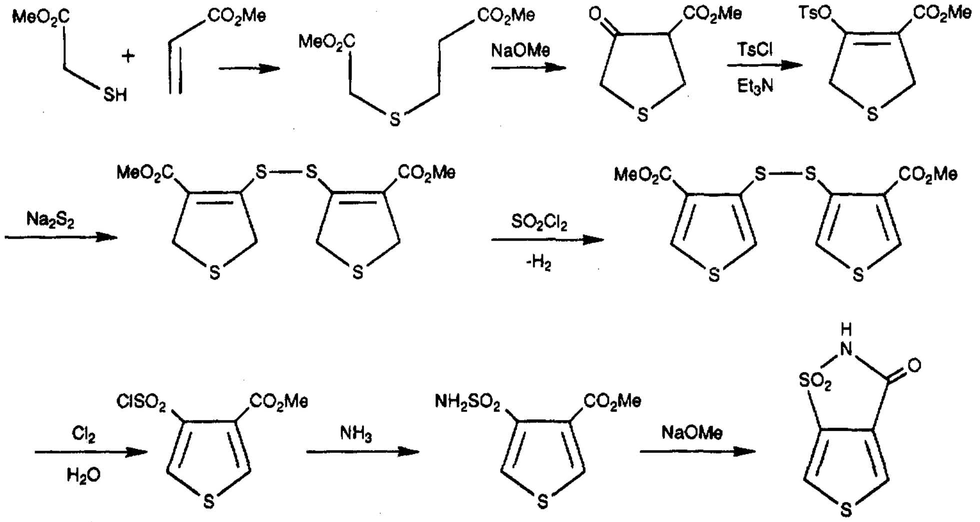

Figure 3. Eight-step route to thiophene saccharine.

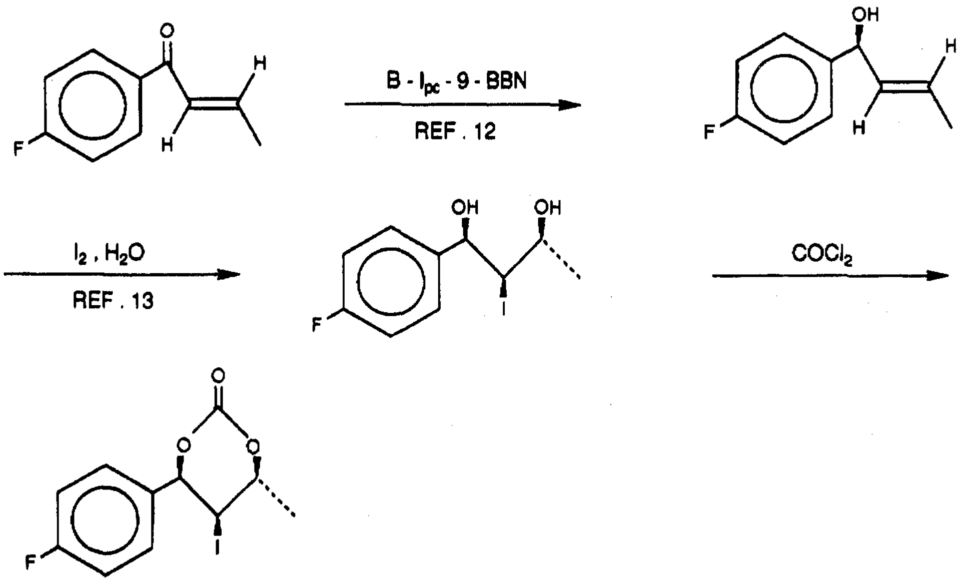

Figure 4. Three-step route to a chiral cyclic carbonate.

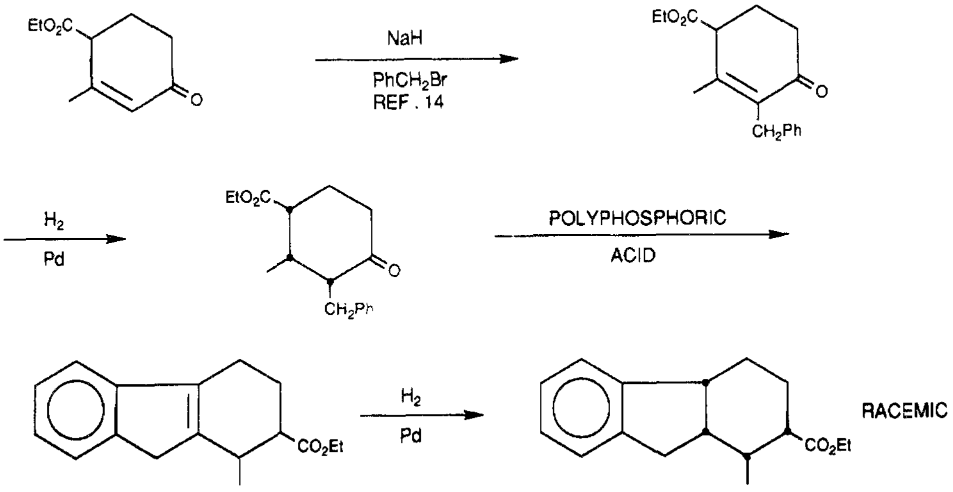

Figure 5. Four-step route to a tricyclic compound.

was recommended accordingly. The following step utilized asymmetric induction, so that three chiral centers were created to give a homochiral product in just two steps.

A Tricyclic Compound. This rather simple and known synthesis starts from Hagemann's ester which is noted as available. The route is interesting mainly for the two hydrogenations. If we allowed the program to simulate hydrogenations at any site in the molecule where two saturated carbon atoms with hydrogen atoms are neighbors, an unmanageable number of not particularly useful alkene structures would be generated. The program focuses on positions α-β to a carbonyl carbon, exocyclic to a saturated ring, one bond away from an aromatic or heterocyclic ring or in a ring where chirality is produced by the hydrogenation, under direction from some hindering substituent of an existing chiral ring atom. Another place for hydrogenations evident to the program is in a six-membered ring where an electron-withdrawing substituent, or its derivative, is placed so that a Diels–Alder adduct is the reactant of the hydrogenation. In addition, when a particular starting material is required by the user, and a double bond present in the starting compound is a single bond in the product, then the program must also note the need for hydrogenation. It is an ongoing study to note the various general situations wherein functionality is usefully removed and to incorporate this knowledge into the program. (See Figure 5.)

PREVENTION OF NONSENSE GENERATION

For each reaction the reacting atoms in the produced substructure are defined very carefully, stating the degree of unsaturation required, the atomic number, the range of number of attached hydrogen atoms allowed, and the heteroatom neighbors allowed. The same is true for all the significant neighboring atoms. This prevents the generation of an acyl lithium, for example, in a reaction to make an α-ketol. These definitions also prevent the generation of an ortho halo aryl Grignard reagent etc.

Certain reactants which the program may generate are too unstable to be acceptable or have never been detected. Examples are propiolic aldehyde, unsubstituted cyclopentadienone, substituted and unsubstituted cyclobutadienes, etc. Each time a reactant is generated, the program routinely checks that it is not on the list of such impossible molecules. Bredt's rule violations are normally prevented at the time of considering the feasibility of the reaction, i.e., before the generation of reactants.

The program has an extensive list of which reagents destroy which functional groups. In addition, account is taken of the existence in the molecule of substructures which are not commonly thought of as functional groups but which would react with certain reagents. Examples are allylic or benzylic saturated CH, or aromatic CH. In the former case we might expect oxidation and in the latter case, sensitivity to electrophilic aromatic substitution. Knowledge at this level might be called the zeroth order knowledge of reactivity. In addition, selectivity information is indispensable. The program knows about matters such as that an unconjugated monosubstituted double bond is more reactive to epoxidation conditions than is any other kind of double bond in the molecule. It knows that acid chlorides are more reactive to nucleophiles than primary chlorides. This might be called the first order knowledge of reactivity. There is a still more sophisticated level in which one also includes conformation in the consideration. Our program does not yet think at this second order level.

The above techniques in our experience have effectively precluded the generation of nonsense routes.

LIMITING THE SEARCH

We list here the various devices which the program uses to limit the search.

I. Limits That an Acceptable Route Must Conform To:

1. A limit on the number of steps. Without this limit very long syntheses would be proposed.
2. A limit on the number of facilitating reactions in the route. A facilitating reaction is defined as one that does not make a carbon–carbon bond and does not make a heterocyclic ring that is found in the goal molecule, does not introduce functionality onto a saturated nonfunctional atom, and in esterification, amide formation, etc. only connects or removes a trivial piece. The trivial pieces are methyl, ethyl, and *tert*-butyl. When this limit is set to zero, we demand a route with no protection or deprotection, no reactions that only change functionality, etc.
3. A limit on the number of consecutive facilitating reactions at one site.
4. A limit on the number of resolutions allowed.
5. The injunction ASYMMETRIC ONLY which rules out any reactions which produce chirality randomly.
6. A limit on the number of reactions which are not regiospecific. An example would be an electrophilic aromatic substitution which produces both ortho or para products, requiring the separation of regioisomers.
7. A limit on the number of doubtful reactions. For most reactions, we can say that the forecasting of the yield is a doubtful exercise. For a few very tested

reactions, the yield is almost always over 90%, so these are classified as not doubtful. An example is the hydrogenation of a terminal double bond or the reduction of an ester with $LiAlH_4$ to a primary alcohol or the first alkylation of a β-keto ester. Of course the presence of interfering groups changes this picture.

8. A limit on the number of resolutions of enantiomers.
9. A limit on the amount of backtracking permitted. This is useful for making a very deep search for a synthesis which requires many steps.
10. Certain user-specified bonds must not be made in any reaction.
11. Certain user-specified atoms must not be reacting atoms in any reaction.
12. In advance, the program can be ordered that if a route is found, then no second route should be generated which differs from the found route only in functionality, e.g., alkyl tosylates rather than alkyl chlorides, etc.
13. Any specific reaction conditions can be ruled out, e.g., on the grounds of cost or marketability of the product. So we can, for example, rule out the use of artificial light, which is expensive for the production of ton quantities of material, or we can rule out the use of mercury in any step or rule out reactions which proceed below a certain temperature, etc.
14. Specific reactions can be ruled out. If the program proposes a route which seems only marginally acceptable, in subsequent runs of the program we can rule out a key reaction of the route, so as to concentrate the program's attention on other parts of the synthetic tree. This exclusion can apply just to the top of the tree or to anywhere in the tree.
15. A limit on the making of peripheral bonds. In order to increase the likelihood of getting a convergent synthesis, we would like to make central bonds, particularly exocyclic central bonds, rather than peripheral ones. A parameter called the centrality ratio controls this. The centrality ratio is the maximum permitted ratio of the distance of a bond to be made from a calculated center of the molecule to the distance of the most remote bond from the center. If the centrality ratio is unity, then any bond in the molecule is allowed to be made. The bond in question is found in the goal molecule. Peripheral protection is always allowed because the bond(s) from a functional atom to the protecting group are absent in the goal.

II. Limits That an Acceptable Starting Material Must Conform To:

1. A limit on the number of functional groups.
2. A limit on the number of carbon atoms.
3. A limit on the number of rings.
4. A limit on the number of chiral atoms.
5. A limit on the number of aromatic ring substituents.
6. A limit on the number of non-hydrogenic atoms in the molecule.
7. A limit on the number of types of functional groups. For example, resorcinol has only one type of functional group. An α-β unsaturated ketone has at least two types of functional groups.
8. One particular specified compound can be a required starting material.
9. The starting material may be required to be in the database of purchaseable compounds.
10. In a synthesis of an aromatic compound, we can rule out an orthosubstituted starting material.
11. We can rule out any starting material containing a specified element, e.g., fluorine or lithium, etc.

III. Limits on the Complexity of Intermediates in the Routes:

1. A descent down the synthetic tree hopefully brings us to simpler molecules than the goal molecule, e.g., there are fewer functional groups, fewer stereocenters, and/or fewer carbon atoms. We can limit the number of such items at any level in the synthesis.
2. We can limit the total number of atoms in any intermediate.

With these limits the program takes a guided tour of the synthetic tree of any molecule of interest. The nodes on the tree that are actually visited depend on the input limits. The maximum size of any "tour" is limited by storage space and the availability of relevant reactions. Our current hardware prevents us from generating reactant structure descriptions more than about half a million times. At that point we exhaust the storage capacity of the disk, and the program can no longer keep a record of what has occurred.

Naturally, the occasional user does not wish to be concerned with setting these 30 or more possible restrictions, so there are default values. It is evident that there is a new field of research here in the controlling of a noninteractive program. There are many more restrictions to be devised, some copying what human beings do and others of a novel kind. Moreover, the program must be made capable of dynamically adjusting these limits, depending on the intermediate molecule being considered and on findings in the course of the search.

CONCLUSIONS

1. A noninteractive synthesis program is feasible; nonsense is not usually produced. There is enormous saving of human time.

2. For relatively simple molecules, when the requisite reactions are in the program, the program can usually propose the same short routes as are known. At present a better route is proposed only occasionally. Especially with large complex molecules requiring syntheses of 10–30 steps, the vastness of the synthetic tree does not allow us to guarantee that the program will find any specific known route. In particular, we cannot ever be sure that we have found the best possible route. This applies both to human search and a computer search. All that we can be certain of is that any route found by the program will meet the input set of specifications.

3. Our reaction collection of just over 2000 reactions is not adequate. A typical interesting synthesis from the literature contains at least one reaction not found in our program. Our program will be widely competent only after a huge further effort.

REFERENCES AND NOTES

(1) Bersohn, M.; Esack, A. Computers and Organic Synthesis. *Chem. Rev.* **1976**, *76*, 269–286.

(2) Bersohn, M. Automatic Problem Solving Applied to Synthetic Chemistry. *Bull. Chem. Soc. Jpn.* **1972**, *45*, 1897–1903.

(3) Walser, A.; Zenchoff, G.; Fryer, R. I. Quinazolines and 1,4-Benzodiazepines. 75. 7-Hydroxyaminobenzodiazepines and Derivatives. *J. Med. Chem.* **1976**, *19*, 1378–1381.

(4) *Organic Syntheses*; Wiley: New York, 1943; Collect. Vol. II, p 130; *Organic Syntheses*; Wiley: New York, 1963; Collect. Vol. IV, p 160.

(5) *Organic Syntheses*; Wiley: New York, 1963; Collect. Vol. IV, p 829; *Organic Syntheses*; Wiley: New York, 1955; Collect. Vol. III, p 337, 658.

(6) Yasuhara, Y. *J. Org. Chem.* **1968**, *33*, 4512.

(7) Arzoumanidis, G. C.; Rauch, F. C. Aniline and Other Aromatic Amines from Carboxylic Acids and Ammonia. A Metal-Catalyzed Process. *J. Chem. Soc., Chem. Commun.* **1973**, 666–667.

(8) Barton, D. H. R.; Magnus, P. D.; Pearson, M. J. The Thermal Rearrangement of 6-Acyloxycyclohexa-2,4-dienones. *J. Chem. Soc., Chem. Commun.* **1969**, 550–551.

(9) Ross, D. S.; Hume, G. P.; Blucher, W. G. Catalysis of Aromatic Nitration by the Lower Oxides of Nitrogen. *J. Chem. Soc., Chem. Commun.* **1980**, 532–533.

(10) Hoffmann, H.; Horner, L.; Wippel, H. G.; Michael, D. Enthalogenierung von Phenolen und aromatischen Aminen, Austausch phenolichser Hydroxylgruppen gegen Chlor. *Chem. Ber.* **1962**, *95*, 523-527.

(11) Rossy, P. A.; Hoffman, W.; Miller, N. Aromatization of Dihydrothiophenes. Thiophenesaccharin: A Sweet Surprise. *J. Org. Chem.* **1980**, *45*, 617-620.

(12) Brown, H. C.; Park, W. S.; Cho, B. T.; Ramachandran, P. V. Selective Reductions. 40. A Critical Examination of the Relative Effectiveness of Various Reducing Agents for the Asymmetric Reduction of Different Classes of Ketones. *J. Org. Chem.* **1987**, *52*, 5406-5412 and references cited therein.

(13) Chamberlin, A. R.; Mulholland, R. L. Jr. Stereo- and Regioselectivity in Iodo Diol Formation from Acyclic and Allylic Alcohols. *Tetrahedron* **1984**, *40*, 2297-2302.

(14) Barnes, R. A.; Sedlak, M. *J. Org. Chem.* **1962**, *27*, 4562.

(15) *Organic Syntheses*; Wiley: New York, 1943; Collect. Vol. II, p 225.