ABSTRACT

Computer programs that synthesize organic chemicals have existed for about 20 years. All of those for which information is available operate on large computers and can map out the synthesis of complex organic compounds.

This project is an attempt to synthesize compounds by using the same general techniques as the large programs do but to limit the scope to certain simple compounds so the program will run on a personal computer. Another objective was to learn how to write a synthesis program, since not much specific information is available in the literature. The types of compounds chosen for this project are simple fluorocarbons.

The program is written in Turbo C. It accepts as input nearly any fluorocarbon compound containing from one to about six carbon atoms that the user wants to synthesize. It then attempts to work backwards from the entered target molecule toward a hydrocarbon starting material from which the target can be synthesized in the laboratory. It operates by applying a library of reactions, which are C functions, to the target so that the internal representation of the target is changed into new forms, according to the rules of chemistry, that represent precursor molecules. After development of a set of precursors, user interaction is necessary to choose one for further reaction toward a starting material.

The program will successfully convert most entered fluorocarbons into hydrocarbon starting materials.
TABLE OF CONTENTS

CHAPTER 1  Introduction

1a. About this program 1

1b. Historical Development of Chemical Synthesis by Computer 2

Influence of graph theory 6

1c. How to Operate the Program 8

1c1. Starting Out 8

1c2. User Options 9

1c3. Entering the Formula 9

1c4. Types of Molecules Allowed 10

1c5. The Reaction Product Tree 10

1c6. The Screen Displays of the Reaction Products 13

CHAPTER 2  Operating Environment 16

2a. Turbo C 16

2b. Separately Compiled Project Files and what they do 17

CHAPTER 3  Chemical Synthesis on a Personal Computer 19

3a. Literature References vs D. Pensak 19

3b. General 21
CHAPTER 8 Special Topics
8a. Molecular Descriptors 64
8b. Errors 66

CHAPTER 9 Summary and Conclusions 68
9a. Results 68
9b. Conclusions 69
9c. Suggestions for Future Work 69

CHAPTER 10 Bibliography 71

APPENDIX A Reaction Descriptions 75

APPENDIX B Descriptions of the Code 85
B1. Module Win 85
B2. Module Step2 90
B3. Module Call react 92
B4. The Reactions 93
B5. Module Perceive 101
B6. Module Getf 104
B7. Module Goal 111
B8. Module Show 114
B9. Module Canon 115
B10. Module Ustring 118
1. INTRODUCTION

IA. About This Program.

This project is a chemical synthesis program for relatively small fluorinated organic molecules. It accepts as input the formula of the molecule the user wants to synthesize. The program then attempts to work backwards from the entered target molecule toward a simple starting material by applying its library of reactions to the target. To reach a simple starting material, a goal state, the process is divided into steps. At the end of each step, user interaction is necessary for the process to continue. For example, if the user entered the molecule CF3CF2CF3, the program would create three new molecules: CF3CCl2CF3, CF3COCF3, and CF3CF2COOH. It would then display them on the screen and ask the user, who is expected to understand some organic fluorine chemistry, to select one of them to pursue to the goal state. The program selects suitable reactions for each choice, runs them if possible, depending on the type of molecule it is, and again displays all the products on the screen and waits for the user's choice. The user can select any molecule produced by the latest round of reactions or produced by any earlier round. This process can be continued until the user is satisfied that an acceptable goal state has been reached.

The pursuit of a starting material from a molecule to be synthesized is called reverse synthesis. User interaction during a reverse synthesis was employed by the first major chemical synthesis program ever written. This program, called LHASA, was begun in 1969 at Harvard University and is still in use (1). This
follow much the same type of analysis as a problem as a chemist
methods. In this task, his aim was to enable the computer to
have computers aid chemists in their search for synthetic
entering the field opened by this pioneer (16). Corey's idea was
program (1). Now, about 20 years later, more and more groups are
Harvard University in developing the first major syntheses
Professor E.J. Corey began his work as recently as 1967 at
IB. History Development of Chemical Syntheses by Computer.
chemistry.

possibility, change it into a new molecule by applying the rules of
in C language that examine the molecule to sent to them and, if
up and applied to the target. Rather, the reactions are functions
no title of known reactions of specific compounds that are looked
should be noted that they do not constitute a data base. There is
The reactions will be described in detail later on, but it
Table 1, a term used often in the discussion that follows (29).
molecules are represented in the computer by a connection
much like a reverse synthesis program would do.
devise this synthesis, he or she would probably work backwards
target compound. However, during the time the chemist was
begin with a starting material and eventually and up with the
a chemist would perform the synthesis the other way around and
bring it closer to a simple starting material. In the Laboratory,
reverse reactions which attempt to simplify a molecule and to
programs is the reaction title. The reactions, of course, are all
techniques pioneered at Harvard. The core of both of these
project, while much smaller than LHASA, uses many of the
would perform (2). An important feature of his work therefore concerned the problem of how a computer can form a basic perception of a molecule the way a chemist would.

Except for a very different approach first investigated by Hendrickson, most computer synthesis programs adopt strategies similar to Corey's but each one differs somewhat in the approach to efficient recognition of structural features, and the strategy used in devising a synthesis. A persistent problem for all investigators is the need to reduce the possible paths to a goal, because present programs develop many irrelevant ones. Thus many of the programs have their own methods of coping with this problem.

The modes of representation of molecular structures are usually classified into two groups. One is an essentially linear representation, such as the Wiswesser Notation and the other is based essentially on graph theoretical exercises of which the main example is the connection table or a matrix representation of it. Different research groups often use different variations.

In Table 1-1, most of the major synthesis programs are listed. In SYNCHEN, WLN is used for fast retrieval of chemical structures from a data base, and for manipulating chemicals, a matrix-connection table is used (3). The connection tables are the primary form of representation and are often derived from a model reported by Wipke (4).
<table>
<thead>
<tr>
<th>Program</th>
<th>Authors</th>
<th>1st Pub.</th>
</tr>
</thead>
<tbody>
<tr>
<td>OCSS</td>
<td>Corey-Wipke</td>
<td>1969</td>
</tr>
<tr>
<td>LHASA</td>
<td>Corey</td>
<td>1971</td>
</tr>
<tr>
<td></td>
<td>Bersohn</td>
<td>1971</td>
</tr>
<tr>
<td>SYNCHEM</td>
<td>Gerlernter</td>
<td>1973</td>
</tr>
<tr>
<td>SOS</td>
<td>Barone-Chanon</td>
<td>1973</td>
</tr>
<tr>
<td>SECS</td>
<td>Wipke</td>
<td>1974</td>
</tr>
<tr>
<td>SYNOPSIS</td>
<td>Dubois</td>
<td>1975</td>
</tr>
<tr>
<td>LHASA(Dupont)</td>
<td>Pensak</td>
<td>1977</td>
</tr>
<tr>
<td>EROS</td>
<td>Ugi-Gasteiger</td>
<td>1978</td>
</tr>
<tr>
<td>MASSO</td>
<td>Moreau</td>
<td>1978</td>
</tr>
</tbody>
</table>

Table 1-1.

The DARC group uses several connection tables but uses a topological representation for a very compact set of tables. The system also accepts a caricaturization of a compound: one may focus on one part of the structure and just neglect the rest if it is considered as not being significant. This is typically what is done in pattern recognition methods, and makes the DARC approach very efficient and versatile (5). DARC is an expert system, not a true synthesis program.

Hendrickson's description of molecules focuses on the carbon skeleton of the molecule and the functional groups that it bears (6,7,8). His method allows very rapid search methods, rivalling the WLN technique.

LHASA uses the traditional connection table (9).

There are two main techniques used in reaction descriptions. The most classical approach mimics the organic chemist's way of thinking. The other is more theoretical.

The classical approach has been the most productive in describing new approaches to synthesis. It is the method used in this project. LHASA uses a very detailed perception of the target to see exactly how a reaction will affect it (10). Bersohn's