1310

IN

P-20

NASA TECHNICAL MEMORANDUM

NASA TM-88401

COMPUTER-ASSISTED DESIGN OF ORGANIC SYNTHESIS

Hiroshi Kaminaka

(NASA-TM-88401) COMPUTER-ASSISTED DESIGN OF ORGANIC SYNTHESIS (National Aeronautics and Space Administration) 20 p HC A02/MF A01 CSCL 09B

N86-27925

Unclas
G3/61 42901

Translation of "Konputa ni yoru Yuki Gosei Keiro Sekkei", IN: Kagaku to Kogyo (Osaka), vol. 59, no. 7, 1985, pp 263-268

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
WASHINGTON, D. C. 20546 APRIL 1986

STANDARD TITLE PAGE

1. Report No. NASA TM-88401	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle COMPUTER-ASSISTED DESIGN OF ORGANIC SYNTHESIS.		5. Report Date APRIL 1986 6. Performing Organization Code
7. Author(s) Hiroshi Kaminaka (Kaminaka, H.)		8. Performing Organization Report No. 10. Work Unit No.
9. Performing Organization Name and Address SCITRAN Box 5456 Santa Barbara, CA 93108		11. Contract or Grant No. NASW- 4004 13. Type of Report and Period Covered Translation
12. Sponsoring Agency Name and Address National Aeronautics and Space Administration Washington, D.C. 20546		14. Sponsoring Agency Code

15. Supplementary Notes

Translation of "Konputa ni yoru Yuki Gosei Keiro Sekkei", IN: Kagaku to Kogyo (Osaka), vol. 59, no. 7, 1985, pp 263-268

16. Abstract

The computer programs to design synthetic pathways of organic compounds have been utilized throughout the world since the first system was reported by Corey in 1969, and the LHASA was reported in 1972 to become the predominant system. Many programs have been reported mainly in the United States and Europe, and groups of corporations, especially chemical companies, have been trying to improve programs and increase the efficiency of research. In Japan, unfortunately, no concrete movement in this area has been seen.

Of course, it goes without saying that these kinds of programs are effective for efficient research, but the remarkable aspect is that these can present data which are unexpected to the researchers to stimulate them to develop new ideas.

17. Key Words (Selected by Author(s))	18. Distribution Statement Unclassified and Unlimited		
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 20	22. Price

COMPUTER-ASSISTED DESIGN OF ORGANIC SYNTHESES

Hiroshi Kaminaka

SUMMARY The computer programs to design synthetic pathways of organic compounds have been utilized throughout the world since the first system was reported by Corey in 1969, and the LHASA was reported in 1972 to become the predominant system. Many programs have been reported mainly in the United States and Europe, and groups of corporations, especially chemical companies, have been trying to improve programs and increase the efficiency of research. In Japan, unfortunately, no concrete movement in this area has been seen.

Of course, it goes without saying that these kinds of programs are effective for efficient research, but the remarkable aspect is that these can present data which are unexpected to the researchers and stimulate them to develop new ideas.

/13*

1. BEGINNING

Various phenomena observed by people since antiquity, such as green leaves of trees trembling in the wind, fresh colors of beautiful blooming of flowers, the flavor of kelp, the poison of shellfish, the fragrance and efficacy of grasses and flowers, and indigo plants and safflowers used for dyeing, have been objects of research since over a hundred years ago as the science of chemistry has steadily progressed. Many chemists were

---

* Numbers in the margin indicate foreign pagination

challenged by these problems. Some through many years of endeavor, some through brilliant talent, and some through luck and chance were successful in isolating and purifying various chemical compounds from animals and vegetables to prove that the afore-mentioned phenomena are caused by these compounds. But as one fact was clarified, another question arose. Research turned to the question of which chemical structures were responsible for the efficacies and colors imparted by these chemical compounds; gradually chemical structures were clarified one after another through various methodologies.

The human mind never ceases to inquire. As the chemical structures become clear, one tries to synthesize them. After many trials and errors, almost artful, detailed synthetic pathways were constructed, and many compounds, which had been originally considered only God's creations, were synthesized by chemists. As a result, a large amount of data has been accumulated, becoming the basis not only of today's chemistry but also of the chemical industries which produce drugs, agricultural chemicals, dyes, cosmetics, and antioxidants.

This kind of research has been done mostly with beakers and flasks, and this will not change greatly in the future. Therefore, some may have questions, such as how the computer can contribute to the synthesis of organic compounds and whether this is just another fad in the so-called computer age. The author will explain the basic ideas concerning the computer-assisted designing of organic synthetic pathways, the present status of the use of this system, its influence on chemical research, and so on, in the report.

2. COMPUTER-ASSISTED DESIGNING OF ORGANIC SYNTHETIC PATHWAYS

2.1. Basic Ideas

Corey of Harvard, one of the most noted in the field of chemistry of organic synthesis, investigated the courses taken by researchers in designing synthetic pathways of organic compounds. One of the methods was through intuition and experience, and he called this "direct association method". This method is greatly dependent on the researcher's ability and is limited in general to chemical compounds of simple structure. /14

How chemical compounds of more complex structure are dealt with is illustrated in the Figure 1, using longifolene as an example.

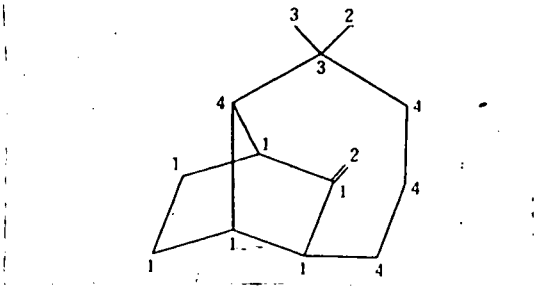
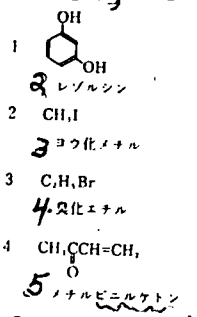


Figure 1. Synthesis of longifolene (logic-centered method).



Key: 1--Longifolene; 2--Resorcin; 3--Iodomethyl; 4--Bromoethyl; 5--Methyl vinyl ketone.

This compound has a rather complex chemical structure, but, based on past research examples, the interatomic bonds which are theoretically possible to sever are severed, gradually leading to simpler compounds and finally ending up with four compounds which are easily available in the chemical industrial circle. This compound, therefore, can be synthesized from these four compounds, namely resorcin, iodomethyl, bromoethyl, and methyl vinyl ketone, as starting compounds. The numbers attached to the structural formula of longiofolene indicate the derivations of the carbon atoms, namely 1 from resorcin, and 2,3, and 4 from respective compounds.

The method constructing logically the synthetic pathways based on past research data in this fashion is called "logic-centered method" by Corey.

In other words, this is the method of taking apart the target compounds into simpler compounds, based on past research data, making a "synthetic tree" as illustrated in the Figure 2, and constructing the synthetic pathways of chemical compounds.

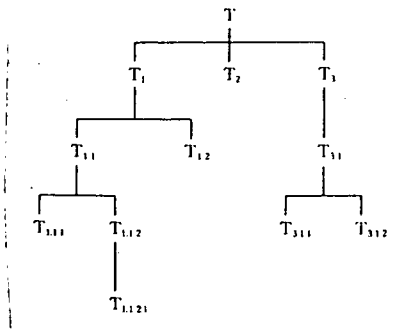


Figure 2. Synthesis tree

Also, as illustrated in Figure 3, Corey proposed the following terms to explain the above ideas, namely "synthesis" for getting the target compound from the starting compounds, "retrosynthesis" for the reversal of this, and "transform" for the reversal of the usual "reaction".

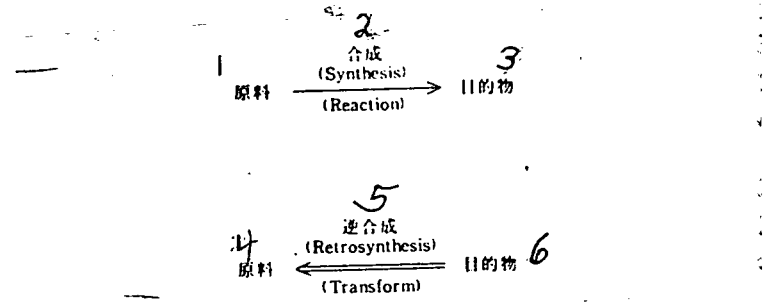


Figure 3. Synthesis and retrosynthesis

Key: 1--Starting compounds; 2--synthesis; 3--target compound; 4--starting compounds; 5--retro-synthesis: 6--target compound.

Although this is the basic idea, it can be quite complicated. For instance, if there are more than ten compounds involved in one retrosynthesis and four stages of retrosyntheses are involved, there are more than 10^4 = 10000 pathways and it is quite obvious, therefore, that the researcher cannot handle the task easily. This is why the idea of using the computer was born with this large amount of data and the need to retrieve the necessary data instantaneously.

2.2. The Contents of the Programs to Design Organic Synthetic Pathways

Based on the above-mentioned ideas, Corey, and Wipke, a computer scientist, developed the system of computer-assisted designing of organic synthetic pathways and reported the first such program, named the "organic chemical simulation of synthesis" (OCSS), in 1969 [1].

The contents were further refined, resulting in the improved LHASA "logic and heuristics applied to synthetic analysis" in 1972, which has become the basis of today's programs of organic synthetic pathways [2].

The following is the explanation of the programs of synthetic pathways of organic compounds, exemplified in the LHASA.

First of all, in order to handle this with the computer, the system has to be constructed completely logically.

In order to satisfy this requirement , Corey, limiting his research to lipids which exist naturally, accumulated a large amount of data, classified the data into several groups from a logical viewpoint, analyzed the accelerations, inhibitions, and terminations of substituent reactions, constructed the formulas assessing the values of retrosyntheses, or transform, and thus came up with the data base of the LHASA.

The LHASA is explained next according to its operational sequence. First, when the targeted chemical compound is written with a stylus, the structural formula of the compound appears on the screen. The fact that chemists can use, at the time of input, chemical structural formulas familiar to them rather than computer terms makes this a very useful system. This is one of the reasons why this system has been the first candidate for /115 the computer-assisted designing of synthetic pathways of organic compounds.

Once the input is made, the computer memorizes the chemical structure of the compound, ranging from what kind of atoms the end product is composed of, whether the interatomic bonds have any ring formations, to what kind of substituents it has.

The next stage is retrosynthesis. One characteristic of this program is that it enables the researcher to choose one of the following three strategies [3].

1) Retrosynthesis is pursued by concentrating on the substituents and partial structures.

2) Retrosynthesis is pursued by concentrating on special bonds.

3) Retrosynthesis is pursued by concentrating on Diels-Alder reaction, Robinson ring reaction, etc.

Which strategy to choose is totally dependent on the discretion of the researcher, and, if one strategy does not work, he can change to another strategy.

The researcher indicates to the computer to start the retrosynthesis following the decision of which strategy to choose. Inside the computer, the retrosynthesis proceeds according to the indicated

strategy and program and in reference to the previously mentioned data base. As the data base contains the function of evaluation, the compounds appear on the screen in a descending of evaluation scores. Those below certain cut-off scores are eliminated.

If the strategy of 1) or 2) is selected, the retrosynthesis terminates after each step, and if the strategy of 3) is selected, it terminates after the indicated reaction is completed. Then the researcher pursues a further retrosynthetic process of the remaining compounds after investigating the results shown on the screen, eliminating the compounds which do not appear to be appropriate because of the difficulty in obtaining the starting compounds or problems in actually performing the reactions. Therefore, the researcher proceeds with his work while conversing with the computer at each step of the process. In general, this kind of program is called "conversational type". The strength of this type is that it is possible to obtain a synthetic pathway with a high probability of actual implementation because of the elimination of pathways with lesser possibilities.

However, if the researcher using the computer lacks ability or experience, or if the researcher, even with experience and talent, is presented with too novel reactions, he may eliminate brilliant ideas without realizing their true value.

The chemical structural formulas or synthesis trees which appear on the screen while using the computer can be printed out if necessary.

The hardware for this program was originally limited to PDP-1 produced by DEC, but, because of its limited capability, the main hardware has changed to VAX-11/750 through PDP-10.

Figure 4 illustrates an actual example of the synthetic pathway of patulin and its synthesis tree.

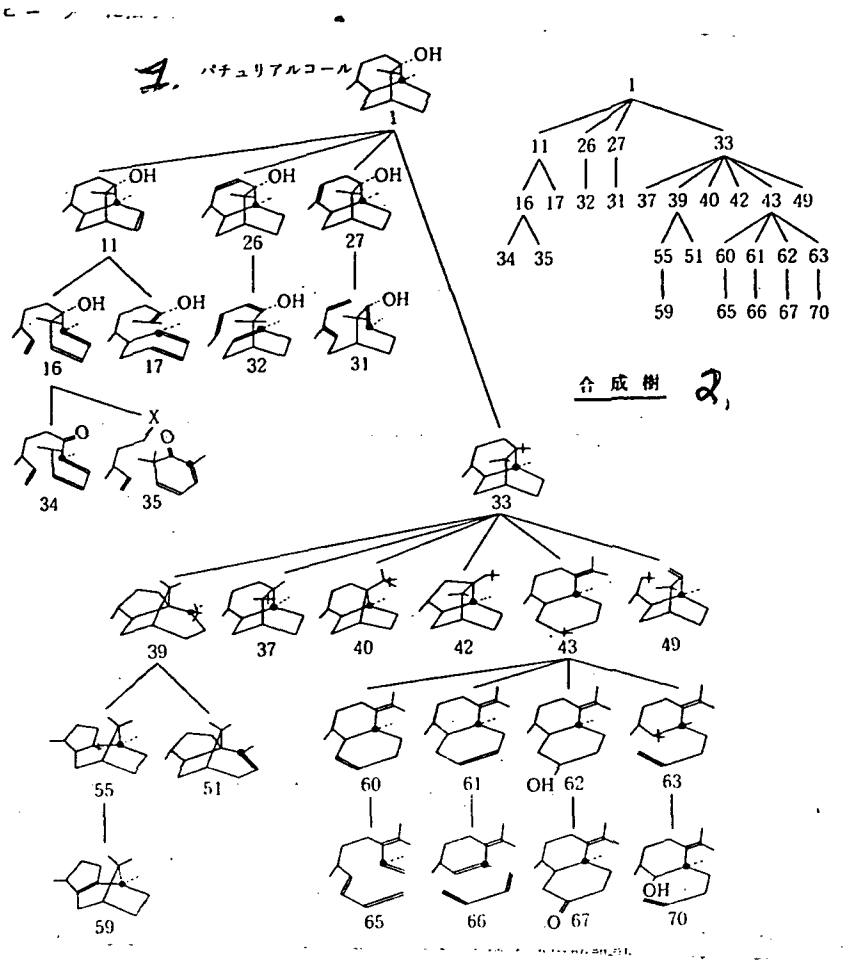


Figure 4. Organic synthetic pathway of patulin.

Key: 1--patulin; 2-- synthesis; .

Especially the reactions indicated by thick lines are those not reported in the literature. Clearly, Figure 4 is an example where the strategy 3) was utilized with Diels-Alder reaction in the designing of the synthetic pathway.

In the development of the LHASA, Johnson and his group at /16
Leeds University in England collaborated with Corey. In the actual application, fine chemical makers such as pharmaceutical companies complained about the limited applicability because of the program's limitation to lipids, which resulted in the expansion of the program to aromatic groups and complex ringed compounds, conducted

by Johnson and others, with the development of the eventually more substantial contents of the LHASA. /16

2.3. Representative Programs to Design Organic Synthetic Pathways and Their Characteristics

When Corey reported the OCSS and subsequently the LHASA, the unprecedentedness of such works perplexed some chemists but, at the same time, stimulated other chemists and their groups to develop newer programs.

Wipke, originally belonging to the Corey school but later diverting from it, reported the program called "SECS", incorporating the element of three-dimensional chemistry into the LHASA [4]. Of course, this program is based on the data base of reactions and is conversational.

Gelernter, noted for his research on artificial intelligence, also reported a program based on data base, named Syschem-II [5]. This program does not belong to the conversational type, and, once the input of the end product is done and a retrosynthesis is indicated, the computer proceeds with the retrosynthesis regardless of the intentions of the inputer and terminates when the compounds obtained as a result correspond to those listed in the Aldrich chemical catalog. The system which proceeds regardless of the intentions of the inputer in this fashion is called "palindromic type", or "batch type". In contrast to the conversational type, this type is characterized by a possibility that a large number of pathways without any sharp focus are presented, but there is also a possibility that a pathway which is totally unexpected can present itself.

Furthermore, Bersohn of the University of Toronto also reported a program based on data base [6]. Although it is a palindromic system, it can present pathways of high applicatility by indicating the starting products or numbers of reactions.

On the other hand, without using any data base of past reactions, Ugi of Munich Technical College, believing that a reaction is a rearrangement of bonds between the starting compounds and end product, reported a program called "EROS", which can present possible reactions including unknown ones after forming determinants with interatomic bonds as elements [7]. Because there is a good possibility that the program left as such may present too many reaction pathways, the number can be limited within a reasonable range by adding thermodynamic considerations.

Yoneda, professor emeritus at the University of Tokyo and currently professor at Tokai University, reported a program called GRACE, based on an idea similar to that of Ugi [8]. Its contents are described in "Chemogram" published by Maruzen.

Hendrickson of Brandeis University, recognizing organic synthetic reactions as changes of skeletal structures of carbon atoms, reported a program to design synthetic reactions based on the idea of designating four characteristics to a carbon atom and classifying the reactions between carbon atoms into 4x4=16 categories [19].

These three programs, which proceed with reaction designs based on mathematical handling rather than on the past data of reactions, are called "logical type". In contrast, those based on reaction data are called "empirical type" or "data base type". Including these as well as the conversational and palindromic types, the programs can be classified as in Figure 5.

Although the strengths and deficiencies of each type have already been partially mentioned, these are summarized here again.

		経験型	論理型
3	会話型	LHASA SECS	
4	回分型	SYNCHEM-II Bersohn 法	EROS GRACE Hendrickson法

Figure 5. Classification of programs of designing reaction pathways.

Key: 1--Empirical type; 2--logical type; 3--conversational type; 4--palindromic type

Although the empirical type, being based on the data of past reactions, presents highly realizable reaction pathways, it has only a small possibility of coming up with totally new reactions. Although the conversational type enables the researcher to focus on concrete, more advantageous, pathways, it depends greatly on the ability of the researcher, and there is a good possibility that he may eliminate good suggestions during the procedure. Therefore, the empirical type depends on the ability of the researcher using the system, and, if the results of the experiments to test the validity of the suggestions are added to the data base, the contents of the program become more thorough. It can be said that the effectiveness of the empirical type not only depends on the ability of the researcher but also grows with the researcher.

At any rate, addition of reaction data is essential in the empirical type, and, in order to improve and enrich it and make it an even more useful system, participation and collaboration of chemistry-related industries with large numbers of researchers and groups of corporations are necessary.

On the other hand, as the logical type presents all the logically possible pathways, it cannot be denied that the responsibility of choosing the truly effective one from these pathways rests on the researcher. At the same time, however, there is a great possibility that a totally unexpected reaction can be obtained.

In order to construct logical type programs, one has to classify and elucidate many reactions according to one consistent theory. Because such programs can provide organic chemistry, hitherto called a discipline of experience and memory, with a new basic idea that many reactions can be understood according to one constant theory, the future expansion of logical type programs will be quite interesting.

It is necessary to add that logical type programs can be developed by groups with a small number of people.

3. THE STATUS OF THE USAGE OF COMPUTER-ASSISTED DESIGNS OF ORGANIC SYNTHETIC PATHWAYS

The responses to the LHASA and SECS by chemistry-related industries in the United States and Europe promptly made research more efficient, and concrete actions were set in motion around 1974. These activities in the United States, England and continental European countries will be described next. Figure 6 summarizes this:

3.1 The United States

Merck Corporation introduced SECS in 1974, investigated it within the company, added some improvements, promoted the usage of this system and, finally in 1979, started to use it in its daily business [10].

Similarly, Eli Lilly, a pharmaceutical company, adapted the LHASA, strove to expand its data base and in June of 1983, founded the LHASA-USA in collaboration with DuPont, Smith Kline, Kodak and ICI-USA, expanding further the data base and the range of application as a collaborative effort.

It is also rumored that Kodak is developing the Sychem-II in collaboration with Gelernter.

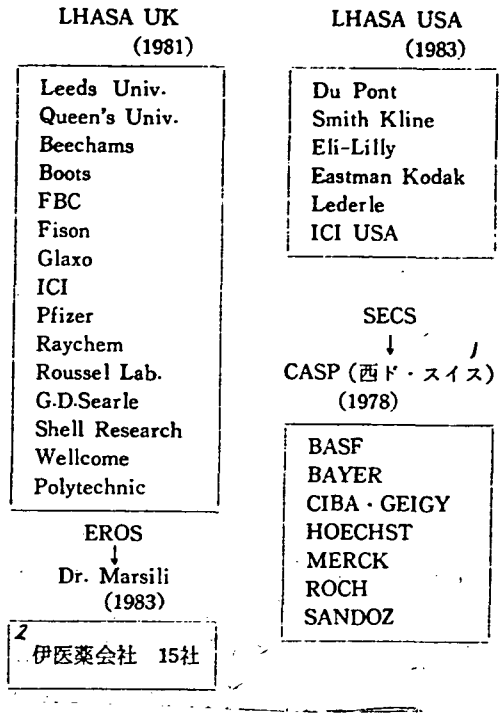


Figure 6. The status of the usage of computer-assisted designs of organic synthetic pathways by chemistry related industries in the United States and Europe.
1--West Germany, Switzerland; 2--Italian pharmaceutical companies

3.2 England

As described previously, Johnson of Leeds University collaborated with Corey in the development of LHASA and developed an organization for the effective utilization of the system. However, because the members complained that the LHASA centered around lipids had little application, the expansion of the program into aromatic groups and complex ringed compounds was promptly initiated [11]. This organization became a non-profit organization called LHASA U. K., Ltd., in 1981 and is operating with subsidies from the government. It is composed of 15 members, including universities and pharmaceutical companies, and is proceeding with the expansion and improvement of the data base. It is possible for non-members to use this system if fees are paid.

Although being a member of this organization, ICI, a major chemical company in England, introduced on its own the EROS and accepts requests for designs of synthetic pathways from outside the company [12].

3.3 West Germany, Switzerland

In West Germany, the SECS became the targeted system because a person there had worked in the laboratory of Wipke. Further perfection of the data base and expansion of the application range are essential for this system to work effectively in industries. Because it was considered that this was beyond the capability of any one company, BASF, SANDOZ, Bayer and Merck formed an industrial ally and began to develop a program called CASP (computer assisted synthesis planning) based on the SECS in 1978. In 1981, CIBA-GEIGY, HOECHST and ROCH also joined the organization with the result that all the big chemical companies in West Germany are now members. They have bi-annual meetings, striving to expand and improve the program.

3.4 Italy

Marsisli, who had participated in the development of EROS under Ugi in Munich, came home with the program, received governmental support, formed an organization composed of 15 mid-sized pharmaceutical companies, developed new methods of producing pharmaceutical drugs and is now conducting the research of in-vivo metabolism of drugs.

3.5 Japan

In contrast to the United States and Europe, where companies have formed organizations and have proceeded with perfecting and improving the programs, some with gonvermental support, the status in Japan is such that each company has introduced programs

individually and has been developing them. They were already behind at the start, and the difference has been increasing. Under these circumstances, the Ministry of International Trade has been investigating the idea of producing its own program with the collaboration of industries, bureaucracy and universities under the project name of ICIC (innovative chemical industry by computerization).since 1984; however, nothing concrete has come from this endeavor.

Now is the time that an earnest effort has to begin with experts in the fields of organic chemistry and computers and with leaders and interested companies as sponsors; people are only waiting for the right moment.

4. HOW DOES THE COMPUTER-ASSISTED DESIGNING OF ORGANIC SYNTHETIC PATHWAYS AFFECT RESEARCH

The present programs are far from always giving 100 percent perfect answers, but it is possible to shorten the research time by presenting fairly focused pathways.

Also, the computer is free from overlooks, misunderstandings and prejudices, and is capable of giving the answers appropriate for the requirements based on the given data base. In other words, these programs cannot only shorten the research time, but also can function as research assistants improving the reliability of the research and, as a consequence, the efficiency per each researcher can be increased greatly.

Furthermore, it must b e added that the researcher tends to be satisfied after coming up with a good idea, however small it may be. Yet, the computer presents relentlessly every piece of /18
accumulated data and every combination thereof according to the program. This gives the researcher a chance to remember what he has forgotton or to correct his misunderstandings; the most

important aspect is that, while human beings tend to develop prejudices, the computer without prejudice can surprise the researcher with unexpected output results, expand his scope and often stimulate him to come up with better ideas.

In short, it can be said that the computer programs of designing synthetic pathways have the capability to shorten research time, to increase the research efficiency by heightening reliability, and to give researchers fresh stimuli into creative activities.

5. CONCLUDING REMARKS

The development of the capabilities of the computer is truly remarkable, and the expansion of the data base of programs to design organic synthetic pathways has been pursued vigorously. As a result, there is no room for doubting the prospect of the ever increasing applicability of such a system..

The only way to cope with competition from giant chemical industries in the United States and Europe and also petroleum chemical industries of resource-rich countries, such as Saudi Arabia, is to increase the efficiency of researchers and concentrate on research development with a smaller number of bright researchers. The application of computer-assisted designing of synthetic pathways, which can improve research efficiency, is considered to be an urgent task in this country.

Whatever organization is established in this country for developing the system, the sophistication of the system is an important project. This will be probably pursued at universities. The empirical type will require an input of a large amount of data, and detailed classification of such data will be essential. On the other hand, even the logical type requires classification of reaction data from a physical point of view; actually, pursuing both systems will result in an indistinct border and they will be

unified in several years. This may be the reason why this country, which has been lagging behind other countries, has still room to develop its original programs.

In any case, it is essential to classify a large amount of reaction data and input it. As this will require a large number of personnel, it appears most realistic and efficient that the industries will provide such personnel when this country plunges into the development of such a system. This is the place where a new form of collaboration between industries and universities can be created. In order to form such an organization, a courageous decision by each corporation to discard small differences and reach a common goal is required.

However, it must not be forgotten that, however sophisticated the contents of the programs to design organic synthetic pathways become, they are only a means to make research more efficient and the freedom generated by such an efficient operation of research has to be aimed at original research.

The output results of the conversational type, as described before, depend greatly on the ability of the user. If one considers that these output results have to be confirmed with experiments and that the data has to be added to the data base, eventually improving the contents of the program, it must be said that everything is dependent on the ability of the researcher.

It is said that today is the computer age. The designing of organic synthetic pathways using the computer, which is actually in the vanguard of such an age, makes research more efficient and has the capability to stimulate researchers into more original research. However, we have to come to the conclusion, age-old and ever so ordinary, that what are necessary to make these programs truly useful are the incessant endeavors, open mindedness to the truth, humble attitudes and abundant imagination of the chemists as the users.

(Received on April 30, 1985)

REFERENCES

1) E.J.Corey, W.T.Wipke, *Science* **166**, 178 (1969)
2) E.J.Corey, W.T.Wipke, R.D.Cramer, W.J.Howe, *J. Am. Chem. Soc.* **94**, 421 (1972) ; 431 (1972)
3) A.K.Long, S.D.Rubenstein, L.J.Joncas, *Chem. Eng. News* 22, May, 9 (1983)
4) W.T.Wipke, H.Braun, G.Smith, F.Choplin, W. Siefer, *A C S Symp.* **61**, 97 (1977)
5) K.K.Agarwal, D.L.Larsen, H.L.Gelernter, *Comput. Chem.* **2**, 75 (1978)
6) M.Bersohn, *Bull. Chem. Soc. Jpn.* **45**, 1877 (1972)
7) I.Ugi, J.Brandt, J.Friedrich, J.Gasteiger, C.Jochum, P.Lemmen, W.Schubert, *Pure Appl. Chem.* **50**, 1303 (1978)
8) Y.Yoneda, *Bull. Chem. Soc. Jpn.* **52**, 8 (1979)
9) J.B.Hendrickson, E.Braun-Keller, G.A.Toczks, *Tetrahedron* **37**, 359 (1981)
10) P.Gund, E.J.J.Grabowski, D.R.Hoff, G.M.Smith, *J. Chem. Ind. Comput. Sci.* **20**, 88 (1980)
11) A.P.Johnson, *Chem. Br.* 59, Jan. (1985)
12) *Ibid.* 238, Mar. (1985)