Translated from *Uspekhi Khimii*, **56**, 1753-1773 (1987) U.D.C. 547.1 : 681.142.2

Computer-assisted Synthesis

N.S.Zefirov and E.V.Gordeeva

The state, problems, and prospects of the new field of computer-assisted chemisty, namely the design and search for new ways of synthesising organic compounds with the aid of a computer (computer-assisted synthesis), are described. The principal developments in programming, used directly by organic chemists are examined. The existing programs for computer-assisted synthesis are subjected to a comparative analysis and the features which more than 20 computer systems have in common, including the "FLAMINGOES" system devised by the authors, as well as the differences between them are elucidated. The bibliography includes 123 references.

CONTENTS

I. Introduction 1002
II. Formulations of the problem. "The synthesis tree" 1002
III. Principal programs for computer-assisted synthesis 1003
IV. Representation and analysis of chemical structures 1004
V. Empirical and non-empirical computer-assisted synthesis 1004
VI. Representation and analysis of chemical reactions 1004
VII. Selection criteria 1006
VIII. The universal computer system "FLAMINGOES" for the solution of problems of synthesis and for the determination of reaction mechanisms in organic chemistry 1010
IX. Prospects for the development of computer-assisted synthesis 1011

I. INTRODUCTION

One of the most important scientific developments involving the application of computers in chemistry is the search for ways of synthesising organic compounds with the aid of a computer, which has come to be referred to in the literature as "computer-assisted synthesis", abbreviated to CS. Since 1969, when the paper of Corey and Wipke,[1] giving a description of the first program for computer-assisted synthesis, was published, this field has developed fairly rapidly and at the present time one may speak of CS not as a set of individual developments but as a major field of science (see, for example, the relevant reviews[2-7]).

The aim of the present review is to cover the principal aspects of the problem of CS. At the same time we have not included in the review the general problems of the strategy and planning of synthesis, because they have already been adequately dealt with in the literature.[8,9] On the contrary, in the present review we shall as far as possible take into account the specific computer-dependent features of the search for methods of synthesis with the aid of a computer. We thought it more useful to differentiate the key factors in the creation of programs for the solution of problems in CS without going into the details of the formulation of algorithms and the machine execution of specific programs, because the number of CS programs is fairly large and is continuously growing. We believe that, within the framework of the selected structure of the description, it will be possible to define clearly the similarities and differences between the existing CS programs and to elucidate the fundamentally new features of a particular approach.

In the concluding part of the review, we shall examine the main principles of the "FLAMINGOES" computer system which we have developed and which can be used, in particular, also for the solution of problems in the planning of synthesis.

It is noteworthy that, like any new field of science, CS has no established universal terminology, which is a significant obstacle to reviewing the relevant data. In the present review, an attempt has been made to present the main aspects of CS in terms of a more unified terminology, which embraces the most apt and accurate, in our view, formulations.

II. FORMULATION OF THE PROBLEM. "THE SYNTHESIS TREE"

Suppose that a certain compound has been specified for which it is necessary to find the possible methods of synthesis from other, precursor compounds. In this case we solve the problem of planning the synthesis in the reverse (retrosynthetic) direction.[1] It is also possible to formulate the opposite problem of the forward (synthetic) planning of the synthesis, i.e. the problem of finding the possible ways in which the given compound or set of compounds reacts and the reaction products.[1] The compound specified (set of compounds) will henceforth be referred to as the chemical system (ChS); as a rule, in the retrosynthetic approach the chemical system includes only one compound while in the direct planning of synthesis there are one or several compounds.

The result of the execution of a CS program in both the forward and reverse directions of the search is the construction of the "synthesis tree", i.e. a graph whose vertices correspond to the initial, intermediate, and final compounds but whose edges represent the chemical reactions in the course of which these compounds are interconverted. If the computer search is carried out in the reverse (retrosynthetic) direction, then the "synthesis tree" is a schematic representation of the possible ways of synthesising the specified chemical system from initial compounds via the intermediate formation of the precursors of the specified chemical system. If the search is carried out in the forward (synthetic) direction, then the "synthesis tree" can be interpreted as the "reactions tree", i.e. as the schematic representation of the possible ways in which the specified chemical system can react.

"The synthesis tree" is constructed as follows in the course of the execution of the CS program. The structure of the compound incorporated in the specified chemical system is introduced into the program. The program generates a set of precursors of the specified chemical system on the basis of empirical or logical instructions inherent in the program. This procedure can be repeated for all or some of the structures generated in the first stage and so on until certain specified conditions governing the completion of the construction of the "synthesis tree" have been fulfilled.

The most important and fundamental factor in CS is the formal treatment and coding of human thought, i.e. without the ideas and reasoning with which the organic chemist operates in selecting or assessing chemical reactions. All the computer systems existing at present for the planning of synthesis constitute only the first and probably fairly tentative steps towards the solution of this fundamental problem. Indeed, so long as a rigorous general theory of organic chemistry has not been created not only an exact prediction of the results of a certain reaction but also the *a priori* assessment of the possibility of the occurrence of this reaction is impossible.[10] Nevertheless, the attempt to reproduce with the aid of the computer all the logical, empirical, and intuitive considerations by which the organic chemist is guided in planning a synthesis appears very useful as a basis for the creation of an artificial intelligence in organic chemistry.

Using the terminology introduced, it is possible to define the task of selecting and assessing the methods of synthesis as one involving the reduction of the "synthesis tree",[11] i.e. the elimination from the latter of chemical reactions and compounds which do not satisfy particular selection criteria. The selection criteria make it possible to diminish greatly the "synthesis tree" and also to organise the generation of the precursor of a given chemical system in such a way that the program generates in the first place the most probable or the most interesting methods of synthesis. The selection criteria are most effective if they operate at the early stages of the construction of the "synthesis tree".

The "synthesis tree" can be constructed under different conditions: (a) non-interactively when the program effects the search, after the input of the necessary initial information, without the interference of the user and without requiring additional information; (b) interactively when the program cannot adopt "independently" a solution in certain situations and turns to the human operator for assistance. Ideally, the CS program should generate the optimum "synthesis tree" without the involvement of the user, but the real situation in theoretical organic chemistry makes it possible to solve under automatic conditions only fairly narrow problems or it requires the introduction into the program of very rigorous limitations, which inevitably leads to the loss of interesting results.

At the present time CS is developing in the direction of increasing the "independence" of programs, although the commonest computer systems provide also for the possibility of an interactive search.

III. PRINCIPAL PROGRAMS FOR COMPUTER ASSISTED SYNTHESIS

As already stated, the aim of the present review is to cover the main aspects of the given scientific field as a whole. The specific features of the tasks to be performed by CS define the main problems on which attention should be concentrated: (a) the representation and analysis of the chemical structure in CS programs; (b) the representation and analysis of chemical reactions (transformations) in CS programs; (c) the selection criteria in CS programs. To facilitate the understanding of the data described below, we quote a brief compilation of the main CS programs with an indication of the leaders of the project, the Company, and the literature references (the most novel computer developments are designated by an asterisk).

"AHMOS"* (Automatisierte Heuristische Modellierung Organisch-chemischer Synthesen); Weise (GDR).[12–15]

"ASSOR" (Allgemeines Simulations-System Organischer Reaktionen); Schubert [Federal Republic of Germany (FRG)].[16,17]

"CAMEO"* (Computer-Assisted Mechanistic Evaluation of Organic Reactions); Jorgensen (USA).[18–26]

"CASP" (Computer-Assistierten Synthese Planung); developed by a Company (FRG, Switzerland).[6]

"CHIRP" (Chemical Engineering Investigation of Reaction Paths); Agnihotri and Motard (USA).[27]

"CICLOPS" (Computers in Chemistry, Logic Oriented Planning of Syntheses); Ugi and Gasteiger (FRG).[28]

"EROS"* (Erzeugung von Reaktionen fur die Organische Synthese); Gasteiger (FRG).[29,30]

"FLAMINGOES"* (Formal-Logical Approach to Molecular Interconversions. Non-empirical Generation, Orientation and Evaluation of Syntheses); Zefirov (USSR).[31–34]

"LHASA"* (Logic and Heuristic Applied to Synthetic Analysis); Corey (USA).[35–41]

"MASSO"; Moreau (France).[42]

"MATCHEM" (MAThematical model of constitutional CHEMistry); Ugi and Gasteiger (FRG).[28]

"OCSS" (Organic Chemical Simulation of Synthesis); Corey and Wipke (USA).[1]

"PASCOP" (Programme d'Assistance a la Synthese des Composes OrganoPhosphores); Kaufmann and Choplin (France).[43–48]

"REACT"; Carhart, Smith, and Djerassi (USA).[49,50]

"REACT" (REACTion path synthesis program for the petrochemical industry); Govind and Powers (USA).[51]

"SECS"* (Simulation and Evaluation of Chemical Synthesis); Wipke (USA).[11]

"SOS" (Simulation of Organic Synthesis); Barone and Chanon (France).[52–54]

"SYNCHEM"* (SYNthetic CHEMistry); Gelernter (USA).[55–58]

"SYNGEN"* (SYNthesis GENeration); Hendrickson (USA).[59–61]

"TOSCA"* (TOpological Synthesis design by Computer Application); Sander (FRG).[62]

Bersohn's program* (Canada).[63–65]

Whitlock's program (USA).[66]

"EPOS"* (Empirical Planning of Organic Synthesis); Koptyug, Smirnov, and Piottukh-Peletskii (USSR).[67]

Many of the programs listed are not fully novel developments, but constitute modifications of programs developed previously, specially designed for the solution of narrower problems than the prototype programs. Thus Whitlock's program and the "SOS" and "PASCOP" systems are based on the same principles as the "LHASA" and "SECS" programs but are designed for the investigation of narrower fields of organic chemistry, such as the synthesis of acyclic structures, the chemistry of heterocyclic compounds, and the chemistry of organophosphorus compounds. The "SECS" program is also the basis of the "CASP" computer system. We shall base the subsequent exposition mainly on novel developments (designated by an asterisk), omitting the nonfundamental modifications of programs and obsolete versions such as "OCSS" (see the newer programs "LHASA" and

"SECS"), "CICLOPS", "MATCHEM" (see the program "EROS"), and "MASSO" (see the program "SYNGEN").

The programs "REACT" and the "CHIRP", "ASSOR", "AHMOS", and "CAMEO" systems occupy a special place among the computer systems enumerated. The program "REACT" (Carhart, Smith, and Djerassi) is an auxiliary component of the "CONGEN" system designed to establish the structure of an unknown compound from spectroscopic and chemical data. The "REACT" system (Govind and Powers) is in many respects similar to the "LHASA" and "EROS" programs but is specially designed for the investigation of chemical engineering processes.

The "AHMOS", "CAMEO", and "ASSOR" programs represent the "mechanistic" tendency in CS. The "AHMOS" and "CAMEO" systems search exclusively in the forward (synthetic) direction generating not the methods of synthesis of chemical compounds but the possible ways in which the specified chemical system can react taking into account the reaction mechanisms and the laws of physical organic chemistry. The "ASSOR" program has many features in common with the "EROS" system (they sprang from a common root—the "MATCHEM"–"CICLOPS" systems), but "ASSOR" is distinguished by a more detailed stage-by-stage examination of the transformations generated in the specified chemical system. The machine search in the "ASSOR" systems can be prosecuted in both the synthetic and retrosynthetic directions, in contrast to the "AHMOS" and "CAMEO" programs. For the description of the "FLAMINGOES" and "EPOS" programs, see Sections VIII and VI respectively.

IV. REPRESENTATION AND ANALYSIS OF CHEMICAL STRUCTURES

As already stated, the functioning of any CS program begins with the input of certain initial data based on information about the structure of the specified chemical system. Virtually all CS programs have at the present time the facility for the input of the structure of a specified CS in the form of a figure with the aid of special graphical devices (graphical display, light pen, etc.).[68] The figure is input as a rule in the form to which the user is most accustomed. The graphical information obtained is transformed into an internal representation of this structure in the program. We shall consider the main principles of the coding of the molecular structure in computer programs, because even at this stage it is possible to determine which particular aspects of molecular structures are the most important for the solution of CS problems.

Any CS program takes into account in the first place the types of atoms in the specified chemical system and the bonds between them, including the multiplicity of the bonds. In other words, a common feature of all the programs is the representation of the structure by a chemical multigraph and its Adjoint matrix.[69] In a number of CS programs, the so called "connection tables", based on the Adjoint matrix of the corresponding multigraph, are used[1,70] for the description of the structure of a chemical compound. The connection tables can also contain additional information about the specified chemical system, for example, they can describe the stereochemical characteristics of the structure[11,24,40,71–75] and atomic charges.[18,39,40,53] In the "SYNGEN" program, the description of the specified chemical system includes only the atoms of the skeleton and the bond between them. Furthermore, for each atom of the skeleton provision is made in the connection table for a "functionality" column, indicating the character of the bonds between the given atom and particular functional groups (FG).[61,76]

The "EROS" program retains the description of the structure in the form of a *BE* matrix, which indicates the bond orders between the atoms of the specified chemical system (non-diagonal elements) and the number of free electrons in the outer valence shell of each atom (the matrix diagonal).[29]

The "SYNCHEM" program[55] uses both the usual representation of structures in the form of connection tables and in the form of the Wiswesser line-formula notation (WLN). The description of the specified chemical system in the form of a connection table is used in the analysis and transformations of structures, while the Wiswesser notation is convenient for the storage and compilation of structural data banks and libraries. Furthermore, the programs are associated with the library of available substances, compiled on the basis of the catalogue of the Aldrich Company in which the reagents are described precisely with the aid of the Wiswesser notation.

In the course of the analysis of the structures forming part of the composition of the specified chemical system, the CS programs extract additional information necessary for further work, namely the characteristic features of the structure of the skeleton (the number and interlinking of the rings and chains), the number and form of the FG, the symmetry of the structure, stereochemistry, etc. The structure is analysed as a rule under automatic conditions in a special recognition unit.[77–87]

V. EMPIRICAL AND NON-EMPIRICAL COMPUTER-ASSISTED SYNTHESIS

Before proceeding with the description of various ways of representing chemical reactions in CS programs, one should note that there are various types of access to these formalised descriptions. At this stage it is in fact possible to differentiate two fundamentally different approaches: the empirical CS, where the transformations of a specified chemical system are effected on the basis of coded information about known organic reactions, and the non-empirical CS, where the transformations are generated by logical-combinatorial procedures without resorting to factual information. In the former case, chemical reactions must be compiled beforehand in the form of a library, while in the latter a set of logical instructions and/or a combinatorial algorithm, which directly generates the transformations, is used to search for the transformations in the specified chemical system.

The advantage of the empirical approach is that the program predicts as a rule entirely probable methods of synthesis and the majority of the selection criteria are contained implicitly in the description of each specific reaction. On the other hand, such programs are linked to a specific library of reactions and are incapable of suggesting a fundamentally new synthetic pathway or finding a new reaction. The non-empirical CS programs are free from this disadvantage, but they require the inclusion in the program of rigorous selection criteria in order to avoid the generation of unrealistic or chemically uninteresting results. At the present time many non-empirical CS programs include also the interaction with data banks by resorting to which it is possible to assess chemically the results obtained, i.e. it is possible to realise an empirical selection criterion.[13,18,61]

VI. REPRESENTATION AND ANALYSIS OF CHEMICAL REACTIONS

The main stage in the computer-assisted planning of synthesis is the search for reactions effecting transformations in specified chemical systems. It is significant that, for

the solution of the problem of retrosynthesis, it is essential to consider retroreactions ("antithetical reactions" or "transformations"[1]), i.e. reactions which are the opposite of those which result in the formation of a specified chemical system. Different terminology is used in the literature for the designation of such reactions. Strictly speaking, the term "redistribution of bonds" reflects most accurately the formalised character of the representation of chemical reactions in CS programs, but in the present review we shall employ the term "transformation" as the most widely used.

The majority of CS programs refer to the empirical approach and use in their execution libraries of transformations (LT).[1,11,40,44,50,51,53,57,64,66] It is possible to differentiate the following common features of all LT. Firstly, the description of the transformation includes the enumeration of the structural fragments which must be present in the specified chemical system in order that the given transformation can be effected. Secondly, the description contains a set of different tests, which determine the possibility, in principle, of effecting the given transformation in the specific chemical system and/or determine the priority ("rating") of the transformation described. Furthermore, the description also includes a list of the structural changes which must be carried out in order to generate the structures corresponding to the result of the application of the given transformation to the specified chemical system. Finally, an identifier must be present for each transformation.

The aspects listed above are characteristic of all LT described in the literature,[11,40,55,88] but, depending on the specific features of the concrete CS program, the description of the transformations can be supplemented by additional information. For example, in the "LHASA" and "SECS" programs, a "character" is attributed to each transformation, i.e. an abbreviated specification of the structural changes in the specified chemical system induced by the given transformation (the formation or dissociation of a bond, ring opening or closure, the introduction, removal or modification of FG, etc.).[11,40] Furthermore, the LT of the "LHASA" and "SECS" programs contain information about the reaction conditions (temperature and auxiliary reagents) and also literature references. In the "REACT" system (Govind and Powers),[51] much attention has also been devoted to the description of the technological conditions for the successful occurrence of specific processes.

The description of transformations in Bersohn's program includes the indication of the yield of the corresponding synthetic reaction and additional tests analysing the correctness of the structures obtained as a result of the given transformation.[88]

The "LHASA"[39-41] and "SECS"[11] programs have the largest and most highly organised LT. The transformations are recorded in these libraries in special languages (CHMTRN in the "LHASA" program and ALCHEM in the "SECS" program), which combine the rigorous logical constructions of the algorithmic machine language and an English vocabulary comprehensible to the chemist.

Gelernter uses a different form of description of transformations, which is closer to the machine interpretation.[55] The set of tests which must be performed on the structure of the specified chemical system is represented by Boolean functions.[89] As a result of the successive operation of these functions, special codes are produced, which enables the program to reach a conclusion about the applicability of the transformation, to attribute to it the appropriate rating, or to report that another transformation must be resorted to in a preliminary step.

the "EPOS" system, the first empirical CS program in the USSR, is being developed at the Scientific-Information Centre for Molecular Spectroscopy of the Novosibirsk Institute of Organic Chemistry of the Siberian Division of the USSR Academy of Sciences.[67] The program is designed for retrosynthetic planning of organic synthesis using a library of transformations; in terms of the form of access to the LT, the system is similar to Gelernter's program.

In contrast to the empirical CS programs, in the programs based on the non-empirical approach, the transformations of a specified chemical system are effected as a result of the application of certain logical instructions and not on the basis of a bond redistribution scheme obtained from the LT.[13,16,18,27,28-34,42,61,62]

The "EROS" and "TOSCA" programs employ sets of "reaction generators", i.e. instructions which describe in the most general form the bond redistribution in the course of chemical reactions.[29,30,62] In the latest version of the "EROS" program,[30] five reaction generators are used, which when combined describe the majority of organic reactions: (1) X: + I–J → I–X–J; (2) I–X–J → X: + I–J; (3) I–J + K–L → I–K + J–L; (4) I–J + K–L + M–N → N–I + J–K + L–M; (5) I–J + K–L + X:→ J: + I–K + L–X, where I, J, K, L, M, and N are reaction centres, i.e. atoms linked by bonds which undergo a change in their order by unity. The dissociation of the "bond" between the centres I and J in the reaction generator can imply both real dissociation and a decrease of the bond order in the specified chemical system: analogously the formation of a "bond" in the reaction generator can correspond both to the real formation of a bond and to an increase of the multiplicity of the bond already present. The centre X: corresponds to an atom whose valence changes by two units in the course of the reaction, for example, the carbene centre in the carbene addition reaction. The functioning of the "TOSCA" program is based on reaction generators analogous to the generators (3) and (4). An additional transformation of the type I–J–K + L–M → I–K + L–J–M is also used.[62]

Evidently the same reaction generator can give rise to different chemical reactions for different centres and bonds in the chemical system. For example, the generator (3) can correspond both to addition to the double bond and to a substitution reaction. Thus reaction generators applied to a specific chemical system can give rise, together with well known reactions, to completely new transformations.

The description of the reaction mechanism constitutes a specific feature of the "ASSOR" program,[16,17] which operates with only four transformations (principal reactions). These transformations correspond to oxidation–reduction processes ($I^{\cdot} + J \rightleftarrows I + J^{\cdot}$ and $I^- + J^+ \rightleftarrows I^+ + J^-$) and to dissociation–association processes ($I–J \rightleftarrows I^+ + J^-$ and $I–J \rightleftarrows I^- + J^+$), which are represented in the most general form. Such representation of the transformations is in many respects analogous to the reaction generator in the "EROS" program, aut the "ASSOR" system pays more attention to the reaction mechanism, which entails an increase of the number of CS stages. For example, four search stages in the "ASSOR" program correspond to the result of the operation of the generator (3) (one stage of the machine search in the "EROS" program).

The "AHMOS" system[12,13] uses a formalised representation of organic reactions, which is also in many respects analogous to the representation of reactions in the "EROS" program. The principal differences consist in the fact that, firstly, the transformations in the "AHMOS" program are more specific than the reaction generator, because they contain an indication of the type of reaction centre (electrophile, nucleophile, etc). Secondly, the transformations in the "AHMOS" program correspond in fact to the elementary stages of

organic reactions and not to the overall result of the process. Weise[13] considers the transformations corresponding to the following processes:

(1) addition: $E + Nu \rightarrow E^{-}-Nu^{+}$;

(2) substitution: $EFABG-NFABG + Nu \rightarrow EFABG-Nu^{+} + {}^{-}NFABG$;

(3) dissociation: $EFABG-NFABG \rightarrow {}^{+}EFABG + {}^{-}NFABG$;

(4) protonation: $X-H + Nu \rightarrow X^{-} + Nu^{+}-H$;

(5) sextet rearrangement; $Y-X-Z \rightarrow X^{+}-Z^{-}-Y$;

(6) polarisation: $X^{-}-Z \rightarrow X=Z^{-}$ or $X^{-}-Z=Y^{+} \rightarrow X=Z-Y$;

(7) electrophilic substitution: $E + H-C(Ar) \rightarrow E^{-}-C(Ar) + H^{+}$;

(8) elimination: $Nu + H-X-EFABG-NFABG \rightarrow X=EFABG + H-Nu^{+} + {}^{-}NFABG$.

Here E is an electrophile, Nu a nucleophile, EFABG and NFABG are respectively electrofugitive and nucleofugitive leaving groups, and X, Y, and Z are other reaction centres (Z is a sextet atom).

The non-empirical CS program "FLAMINGOES"[31-34] can operate with both multistage chemical reactions (with an arbitrary degree of complexity) and with individual stages of the mechanism. The universality of the "FLAMINGOES" computer system compared with other non-empirical CS programs is achieved by using not an individual extremely limited set of reaction generators but a general combinatorial algorithm in the course of the execution of which the most varied chemical reactions and/or reaction stages can be generated, including completely new or little investigated processes (for a more detailed description of the "FLAMINGOES" system, see Section VIII).

The program "SYNGEN"[60,61] is based on a novel classification of organic compounds by Hendrickson.[90-95] This system is also based on the claim that the majority of organic reactions can be represented as a change in the immediate environment at one or several carbon atoms on passing from the initial compounds to the products (or in retrosynthesis on passing from the specified chemical system to the precursors). It is suggested that the entire wide variety of the possible environments of a carbon atom can be expressed by four bond types:

(1) type H: σ-bond with a hydrogen atom or an atom less electronegative than hydrogen;

(2) type R: σ-bond with another carbon atom;

(3) type Π: π-bond with another carbon atom;

(4) type Z: σ- or π-bond with an atom more electronegative than hydrogen.

The number of bonds of each type is designated respectively by h, σ, π, and z; evidently, $h + \sigma + \pi + z = 4$. Hendrickson calls the sum $f = \pi + z$ the functionality of the carbon atom. In this system any reaction will be defined if the change in the environment at all the carbon atoms on passing from one chemical system to another is specified. For an individual carbon atom, this change is designated by two letters: the first characterises the type of bond formed and the second the type of bond ruptured. Thus altogether there is a possibility of $4 \times 4 = 16$ types of structural changes at the carbon atom.

Hendrickson divided the entire set of organic reactions into three classes. The first class includes constructive reactions, i.e. reactions as a result of which the carbon skeleton of the organic compound is altered. Constructive reactions involve three types of structural changes: RH, RΠ, RZ. The trans-functionalisation reactions (the introduction, elimination, or alteration of the FG), including nine types of structural changes, namely HH, HZ, ZH, ZZ, ΠH, ΠZ, HΠ, ZΠ, and ΠΠ, belong to the second class. The remaining four types of changes in the environment of the carbon atoms refer to destructive reactions, which are not used in the "SYNGEN" program.

Among all the constructive reactions, the "SYNGEN" program employs only those in the course of which a single C–C σ-bond is formed and the environment changes at not more than three carbon atoms in the chain on either side of the bond formed. For such constructive reactions and also for the trans-functionalisation reactions, lists of instructions are stored in the program in a coded form and ensure the execution of the necessary transformations in the specified chemical systems; these instructions are also referred to as half-reactions. On the one hand, the compilation of such instructions does not involve the use of factual data from organic chemistry, but, on the other, the set of possible half-reactions is stored in the computer memory like the libraries of transformations in the empirical CS programs. Consequently Hendrickson's method is to some extent intermediate between the empirical and non-empirical CS.

VII. THE SELECTION CRITERIA

In the previous Sections it was shown how the chemical information in the known computer systems is formalised and coded. However, as already stated, the most important problem in CS is the creation and formalisation of selection criteria whose operation in fact imparts to the CS programs the features of an "artificial intelligence".

In empirical CS programs, the selection criteria can be used in three main stages of the search: (1) the choice of a definite "strategy of the synthesis", which limits the number and form of the transformations available in the LT; (2) the assessment or grading of concrete transformations before their use in the specified chemical system; (3) the assessment or grading of concrete precursors of the given chemical system obtained as a result of the transformation.

"The strategies of the synthesis", which make it possible to direct the machine search in a specified direction, have been developed to the greatest extent in the "LHASA" and "SECS" programs. We shall list the main types of such strategies.[11,35,39-41]

1. "The strategy of applicability", i.e. the search in the LT for all the transformations which are applicable to a specified chemical system. The main disadvantage of this strategy consists in the fact that large classes of transformations, such as the introduction, removal, or modification of FG are almost always applicable, which leads to an appreciable growth of the "synthesis tree".[11,35]

2. Strategies designed to achieve the transformations, i.e. the search in the LT for known sufficiently reliable reactions with the aid of which it is possible to carry out an elegant synthesis of specified chemical systems.[41] Such reactions include, for example, the following processes: the Diels–Alder reaction, the Robinson annelation, the Aldol condensation, and sigmatropic rearrangements. Another variety of strategies designed for transformations consists in the construction with the aid of the simplest, "mechanistic" transformations (the addition of a proton, the removal of a proton, etc.) of the possible mechanisms of the formation of the specified chemical system.[41]

3. Structurally oriented strategies, i.e. the search for transformations leading precisely to precursors of the specified chemical system which contain definite structural fragments.[11,41]

4. Topological strategies, i.e. the search for transformations leading to the rupture of one or several "strategic" bonds as a result of which the structure of the initial chemical system may be appreciably simplified.[11,41,96]

5. Stereochemical strategies, i.e. the search for transformations with the aid of which it is possible to synthesise the specified chemical system with a definite configuration of all the stereocentres.[41]

6. Strategies oriented towards functional groups, i.e. the search for transformations which effect the introduction, removal, protection, or modification of FG. This strategy usually precedes the application of other strategies, whose direct application is frequently difficult precisely because of the presence in the specified chemical system of concrete functional groups.[11,41,97-99]

The selection of a definite strategy can be achieved automatically or with participation of the user. The application of several strategies makes it possible as a rule to find extremely effective and elegant methods of synthesis.[40,41]

In the programs of Bersohn[63,64] and Gelernter[55-57] access to the LT is achieved on the basis of data obtained in the course of the analysis of the structure of the specified chemical system or, in other words, "the strategy of applicability" is carried out. In Bersohn's program, the application of this strategy is rigorously ordered. Firstly, all the structural fragments found by the computer in the recognition unit receive different ratings. The highest rating is given to structural fragments in the formation of which the greatest number of bonds are formed simultaneously, as happens, for example, in the formation of a three-membered ring in the Simmons–Smith reaction and of a six-membered ring in the Diels–Alder reaction.

Secondly, the reactions in each section of the LT are also arranged in order of decreasing priority of the corresponding synthetic reactions. The highest rating is given to the reactions involving the introduction of FG; next follow the reactions leading to the construction of the molecular skeleton, the isomerisation reactions, and the reactions leading to the protection of the FG, the removal of the protection, the modification of the FG, and the elimination of the FG. Fragmentation reactions have the lowest rating, because the corresponding transformations complicate the structure of the specified chemical system.[64]

One of the most fundamental and hitherto unsolved problems of organic chemistry—the problem of the *a priori* estimation of the probability of the occurrence of the reaction—is solved in CS programs with the aid of formal and empirical selection criteria. The transformation is regarded as formally feasible if the given chemical system contains a structural fragment formally essential for the execution of the given transformation. Empirical selection criteria are based on a more profound analysis of the structure of the specified chemical system, the reaction conditions, the reactants, etc. The final assessment (rating) of the transformation is determined as a result of a series of tests, which are as a rule contained in the description of the transformation itself.[11,39,40,55,88] The information obtained in the recognition unit is most actively employed precisely at this stage. For a more detailed analysis of the transformations, additional information may be resorted to.

Additional selection criteria have been developed most in the "SECS" program.[11] Thus the applicability of a particular transformation from the LT is determined not only by the results of the tests contained in the description of the transformation but also in conformity with a specified "list of aims", which contains information about the required changes in the skeleton on the FG. In conformity with this list, the transformations of the LT are rejected if the character of the transformation does not correspond to any of the "aims" in the list. If the transformation satisfies the given criterion but the structural fragment indicated in the description of the transformation differs from the structural fragment of the specified chemical system only by the presence and/or type of FG, then, before the direct application of the transformation, the "SECS" program carries out the necessary modification of the structure (in the "LHASA" program the strategy oriented towards the FG operates analogously).

In the first version of the program "SECS" much attention was already devoted to allowance for the steric and electronic effects in assessing a specific transformation. The subprogram "SYMIN" constructs in the computer memory a three-dimensional model of the specified chemical system using calculations by the method of molecular mechanics. In order to estimate the probabilities of reactions with participation of coupled systems, the calculations of molecular orbitals by the Hückel method are used. Depending on the results obtained, the rating of the transformations selected from the LT may be altered.

An important criterion of the applicability of selected transformations is a test designed to show that the conditions favourable for the occurrence of the corresponding synthetic reaction would induce undesirable changes in the specified chemical system. The "SECS" program analyses the sensitivity of various FG under the conditions of the given reaction and selects the protecting reagent. If the reaction conditions and the structure of the specified chemical system preclude a satisfactory protection of the FG, the transformation is regarded as unsuitable.

In the absence of information about the symmetry of the specified chemical system, the application of the same transformation to a structure having several symmetry elements inevitably leads to the duplication of results. The "SECS" program determines the entire symmetry group of the three-dimensional model of the specified chemical system, which makes it possible to generate subsequently only non-equivalent results.

Bersohn's[63,64] and Gelernter's[55-57] programs operate non-interactively, generating the methods of synthesis of the specified chemical system without interference by the user. Consequently the selection criteria underlying these programs must be more rigorously formalised than the criteria in the "LHASA" and "SECS" programs. Thus in Bersohn's program the transformations within a single class of priority are assessed by a special rating which is higher the higher the yield of the corresponding synthetic reaction.[88]

In an earlier study, Bersohn and co-workers[63] used additional criteria for the assessment of individual stages of the synthesis. Each transformation received its rating, which was characterised in terms of arbitrary units by the "simplification"/"cost" ratio. The value of "simplification" increased if the transformation resulted in a decrease of the number of rings, the number of FG, and the number of non-hydrogen atoms. The "cost" depended on many factors, including the yield of the reaction.

In the "SYNCHEM" program[55-57] each transformation is assessed rigorously in conformity with a set of tests in the form of Boolean functions. These tests embody many of the heuristic rules by which the chemist is guided in assessing a chemical reaction. On the basis of the results of these tests, the initial rating of the transformation can be altered (for example, it can be increased if a conjugated activating group is present or it can be reduced if the transformation is interfered with by steric hindrance). The successive application of the Boolean functions can result in a partial modification of the very scheme of the transformation. For example, if a FG sensitive to the reagent used during the reaction is found, then the program can introduce another reagent or can resort to a procedure involving the protection of the FG.

Together with the assessment of the transformations, the precursors themselves are assessed in empirical CS programs. For example, in the "SECS" program[11] different variants are rejected with the aid of the following structural selection criteria: the breakdown of Bredt's rule; the presence of an antiaromatic system; the presence of a cumulated system; a triple bond in a small ring; a *trans*-olefin in a small ring; a cyclic system with a bridge in the *trans*-position; *trans*-condensed three-membered rings; an incorrect valence of an atom; two charges of identical sign on different atoms; an unstable combination of FG. Analogous selection criteria for structures are applied in the majority of both empirical and non-empirical CS programs.

Apart from the criteria testing the formal correctness and "chemical significance" of the structure generated, a very effective criterion, determining the completion of the construction of the synthetic chain in the specified chemical system, is used in a number of non-interactive CS programs designed for retrosynthesis.[55,56,60,61,63,64] In order to apply this criterion, the program must be supplemented by a special data bank—the library (catalogue) of the available substances (DBS). If the structure of the successive precursor generated in the course of the machine search is identical with one of the structures stored in the DBS, the construction of one of the possible synthetic pathways is regarded as completed. Bearing in mind that the reactions leading to the introduction, removal, or modification of FG play a secondary role compared with the reactions involving the construction or modification of the skeleton, one may assume that agreement to within the number and form of the FG is sufficient to satisfy the given criterion.[60,61]

The main problem encountered by non-empirical CS programs is that of the selection from the entire set of the possible synthetic pathways of the most probable and interesting from the chemical point of view. In the non-empirical programs, it is possible to differentiate subproblems in the solution of which selection criteria are employed: the limitation of the types of transformations generated in a concrete specified chemical system; the limitation of the applicability of the transformations in the specified chemical systems; the evaluation or rejection of the generated precursors of the specified chemical system. The selection criteria for the precursors[18,62] are in many respects similar to the analogous criteria in empirical CS programs, and we shall therefore transfer the centre of attention of our discussion to the specific features of the selection criteria of the first two types in non-empirical CS programs.

The possibility of selecting the types of transformations which can be generated in the specified chemical system in an explicit form is present in the "EROS" program,[29,30] where, firstly, provision is made for the situation where the user excludes from the list of reaction generators stored in the computer memory those which correspond to the processes which in his view cannot lead to the specified chemical system. With the user's implicit agreement, it is assumed that all five reaction generators can be applied to the search for the precursors of the specified chemical system. Secondly, there is a possibility of imposing definite limitations on the form and size of the "synthesis tree".[30] In particular, one can seek the maximum number of levels of the "synthesis tree" (the number of CS stages), the number of precursors at each stage, the maximum number of precursors generated for a single chemical system, and the maximum number of precursors which can be regarded as specified by the chemical system in the next CS stage.

The main principle in accordance with which the types of transformations are selected in the "SYNGEN" program[60,61] is that the planning of the synthesis must be based in the first place on the constructive reactions, i.e. reactions in the course of which the skeleton of the specified CS is formed. The trans-functionalisation reactions in the course of which the FG are modified but the molecular skeleton is unaffected are regarded as auxiliary.

As shown in Section VI, the description of any transformation in the non-empirical CS programs in a general form can be represented as a set of reaction centres with an indication how the bonds between them are distributed. Consequently the selection criterion which can limit the applicability of particular transformations consists in the selection of the potential reaction centres or reactive bonds from the entire set of atoms and bonds in the specified chemical system.

In the "EROS" program,[29,30] the given criterion is applied in the analysis of the structure of the specified chemical system in the course of which the bonds capable, in principle, of rupturing during subsequent transformations are determined. Such bonds can be recognised both automatically and with participation of the user. In the automatic mode, the multiple bonds, the C–X, H–X, and X–X bonds (X is a heteroatom), as well as the adjoining bonds are revealed as reactive. Aromatic bonds are not regarded as reactive. The user can arbitrarily correct the list of reactive bonds initially formulated by the computer. Furthermore, the user can specify the optimum enthalpy of the reaction and the range of the permissible enthalpies. For each transformation generated by the computer, the enthalpy is calculated in accordance with Allen's scheme.[100–103] Gasteiger[103] claims that this in general non-rigorous approach yields a satisfactory agreement with experimental data and can serve for a rough estimation of the probability of the reaction. If the calculated reaction enthalpy falls within the range of permissible values, the transformation is regarded as probable; otherwise the precursor found by the computer is rejected. The specified optimum enthalpy makes it possible to estimate in detail the probability of the process predicted by the computer. The closer the calculated reaction enthalpy to the specified value, the higher the rating of the transformation and the corresponding precursors will be investigated in the first place in the next CS stage. We may note that the "CHIRP" program, which is related to the "EROS" system, uses the calculated Gibbs free energies for the estimation of the probability of the transformations in the specified chemical system.[27]

Gasteiger and co-workers have also carried out studies designed to create non-empirical selection criteria for reactive bonds, criteria based on a quantitative determination of the partial atomic charges,[104,105] inductive[106] and resonance[107] effects, dipole moments,[107] and polarisability effects.[108] Iterative procedures, which make it possible to determine the distribution of partial charges in simple compounds with small numbers of atoms, containing heteroatoms, have been described.[104,105] On the basis of the data obtained, it is possible to calculate the C–X (X is a heteroatom) bond energies,[105] the pK_a,[105] and $J(^{13}C–H)$,[109] and the proton affinities.[106,108,110,111] It may be hoped that in its subsequent development the method based on the partial balancing of the orbital electronegativity[105,112] will indeed make a significant contribution to the creation of a general theory of the reactivity of organic compounds, but at the present time the selection of reactive bonds on the basis of topological and structural characteristics constitutes a more general, faster, and more reliable method.

The "TOSCA" program[62] employs an interesting and novel method for the limitation of the applicability of transformations in the specified chemical system. Like the empirical CS programs, the operation of the "TOSCA" system

is based on the selected strategy of the synthesis. The program employs two strategies: consonant and dissonant. The consonant strategy consists in the search for nucleophilic and electrophilic substitution and addition reactions. The dissonant strategy involves the search for oxidation and reduction processes—reactions leading to a change in polarity. In contrast to the empirical programs, in the "TOSCA" program the strategy is chosen precisely as the stage corresponding to the selection of the potential reaction centres and reactive bonds. All the atoms incorporated in the specified chemical system, starting with the heteroatoms, are designated by the "+" or "-" sign in accordance with their electrophilic or nucleophilic properties. If all the bonds are consonant in the specified chemical system, i.e. there is not one pair of adjacent atoms with identical charges, preference is given to the consonant strategy.

For a specified chemical system containing dissonant bonds (adjacent atoms with identical signs), it is possible to select both a dissonant and a consonant strategy. In the consonant strategy, limitations are imposed on the operation of transformations: bonds may be formed only between atoms with opposite signs while dissonant bonds are revealed as unreactive. On the other hand, the dissonant strategy permits the recombination of atoms with identical signs.

The logic of the "SYNGEN" program[60,61] is based on the idea of "ideal synthesis" as a sequence of constructive reactions.[10] For this reason, bonds in the molecular skeletons incorporated in specified chemical systems are chosen exclusively as the reactive bonds. The "SYNGEN" program initially seeks all possible ways of subdividing the skeleton into constituent components—synthons. In the general case, the number of subdivisions can be very great and additional criteria for the evaluation of the effectiveness of a particular set of synthons for the synthesis of specified chemical systems are therefore necessary. On the basis of the concept of "ideal synthesis", Hendrickson introduces the criterion of convergence.[10,60,95] In convergence synthesis, the fragments of the target structure are formed independently and are combined only at the end of the synthesis. In conformity with this requirement, the subdivision of the skeleton of a specified chemical system in the first stage should lead to two synthons with approximately equal numbers of atoms, to four such synthons in the second stage, to eight synthons in the third stage, etc.

In the course of a single subdivision of the skeleton, the rupture of one to two bonds and the closure of at most one ring are permitted. Preference is given to subdivisions for which the "reassembly" of the molecule is described by effective synthetic reactions, for example the Diels–Alder reaction. It is essential to differentiate the situation where the synthons obtained as a result of the subdivision of the skeleton are found to be identical, since the synthetic scheme simplifies greatly in this instance. Apart from the identity or similarity of the synthons, a test is made to discover if each synthon found is present in the DBS. For synthons which are absent from the DBS, the subdivision operation is repeated.[60,61]

Each sequence of subdivisions of the skeleton of the specified chemical system, leading to synthons for which correspondences in the DBS have been found, generates the so called "ordered bond set".[60] This set consists of bonds which must be broken in a specific sequence in order to obtain the specified subdivision into synthons. The bonds entering into the given set are considered in succession in the same order and all the suitable (on the basis of the presence of concrete FG in the specified chemical system) constructive half-reactions or, where necessary, transfunctionalisation reactions are applied to each bond. This results in the discovery of not only the skeletons of the initial compounds but also of the FG present in these compounds.

As already stated, the "AHMOS"[12–15] and "CAMEO" systems occupy a special place among the CS programs. In terms of the character of the representation of the structure of the transformations, these programs are close to the "EROS" and "LHASA" systems respectively. The "AHMOS" program has come to be classified as a non-empirical CS, while the "CAMEO" system, based on several small data banks, occupies an intermediate position between the non-empirical and empirical systems. We placed the special discussion of the "AHMOS" and "CAMEO" programs in the Section devoted to the selection criteria, because novel criteria based on the laws of physical organic chemistry in fact constitute an inseparable feature of these programs.

Pearson's classification of acids and bases serves as the basis of the selection criteria used by the "AHMOS" program.[12,13] Weise considers six types of reaction centres: a hard electrophile, a soft electrophile, a hard nucleophile, a soft nucleophile, a nucleofuge, and an electrofuge. The program has a small data bank in which the reactivity indices of different FG, reflecting quantitatively the nucleophilic and electrophilic properties of these groups, are stored. On the basis of the scale of evaluations introduced, all the FG found by the program in the recognition unit receive their rating—the reactivity index. Subsequently the program uses these ratings to calculate another quantitative characteristic—the "tendency to react" index for different modes of interaction of electrophiles and nucleophiles provided for by the program (see Section IV). The values obtained are compared with the specified threshold values and the transformations whose occurrence is most preferred in the specified chemical system are thus determined. The program effects priority transformations, generating thereby the possible products of the reaction of the specified chemical system.

The criteria for the evaluation and selection of the possible mechanisms of an organic reaction have been specially developed in the studies of Jorgensen and co-workers.[18–26] The "CAMEO" program for the prediction of the reaction products when the starting materials and the reaction conditions are specified has been created on the basis of these criteria. It is noteworthy that the program carries out the search in the forward (synthetic) direction and is designed not so much for the generation of whole reaction sequences as for the evaluation of the possibility of the occurrence of a particular reaction in different directions. The program has several units, each of which is designed to work with a definite class of reactions. Overall, the "CAMEO" program covers the following processes: interaction of electrophiles and nucleophiles under the conditions of base catalysis[18] including certain reactions of organometallic compounds,[19] the reactions of organosilicon compounds,[19] pericyclic reactions,[21,22] and nucleophilic and electrophilic substitution in aromatic compounds.[23–25] For each of these classes of reactions, special criteria estimating the probability of the process in a particular direction have been developed.

We shall consider in greater detail the criteria for the selection of transformations interpreted in the "CAMEO" program as the reactions of electrophiles and nucleophiles under the conditions of base catalysis.[18] The potential reaction centres are selected as follows: with the aid of a small data bank and a special algorithm,[26] the program determines the most acid protons and generates transformations corresponding to the transfer of protons to the base. The atoms to which the protons are linked are regarded as potential reaction centres—nucleophiles. For organometallic

compounds,[19] account is also taken of the possibility of the occurrence of competing addition and exchange reactions. The processes leading to a weaker base are regarded as dominant.

The potential electrophilic centres are determined from structural characteristics (for example, a multiple bond) and also from the pK_a of the leaving group taking into account the specified reaction conditions. With the aid of a set of heuristic rules and data on the structure of the reactants and the character of the FG obtained in the recognition unit, the possibility of the occurrence of competing addition, substitution, and elimination processes with participation of nucleophilic and electrophilic reaction centres is determined.

Subsequently this unit of the "CAMEO" program underwent significant modifications and includes at the present time also the recognition and investigation of reactions with participation of organosilicon reactants[20] and vinyl and aromatic electrophiles.[23]

The functioning of the unit for the study of electrophilic substitution reactions in aromatic systems[25] is based on the determination of the potential reaction centres participating in the process. The most important procedure in this unit is the calculation of the relative reactivities of different centres in the aromatic system. Depending on the reaction conditions and the type of aromatic system (substituted benzene, a heterocycle, a polycyclic system), an initial relative reactivity index is assigned to the system. Using the data obtained in the analysis of a specified chemical system in the recognition unit and the set of heuristic rules, the program takes into account the influence of substituents. An equation has been derived for the calculation of the correction to the initial index as a function of the number and type of electron-donating and electron-accepting substituents. After the calculation of the relative reactivity indices of the aromatic systems as a whole, the program proceeds to determination of the different reaction centres in each system, being guided by the set of heuristic rules. Account is taken also of the steric effect of the substituent.

The stereochemical selection criteria play a major role in the study of pericyclic reactions.[21] Thus dienes are assumed to be unreactive if steric hindrance prevents them from attaining the *S-cis*-configuration. Apart from the stereochemical criteria, various heuristic rules can be used for the selection of reactive dienes, dienophiles, 1,3-dipoles, and dipolarophiles. The main criterion for the selection of pericyclic reactions is applied in the comparison of the energies of frontier orbitals. It is assumed that the lower the difference between these energies, the easier the cycloaddition. Furthermore, the program predicts the relative regioselectivity and stereoselectivity of the process from the results of the calculation of the energies of the frontier orbitals. The energies are determined from the ionisation potentials taken from the special data bank of the "CAMEO" program.

The analysis of [2 + 2]cycloaddition reactions, [2,3]- and [3,3]-sigmatropic rearrangements, and [1,*j*] shifts of σ-bonds (j = 3,5,7)[22] was subsequently included in the program. In order to estimate the priority of a particular reaction pathway, the activation energies calculated on the basis of the analysis of a large number of experimental kinetic data for reactions of the types enumerated were used. Contrary to the authors' expectation,[21] it was found that the enormous variety of the rates of pericyclic reactions and the influence of too many factors on the rate of reaction are all causes which preclude at the present time the development of general principles for the estimation of the probabilities of pericyclic reactions.

VIII. THE UNIVERSAL COMPUTER SYSTEM "FLAMINGOES" FOR THE SOLUTION OF PROBLEMS OF SYNTHESIS AND FOR THE DETERMINATION OF REACTION MECHANISMS IN ORGANIC CHEMISTRY

The "FLAMINGOES" computer system (Formal-Logical Approach to Molecular Interconversion. Non-Empirical Generation, Orientation, and Evaluation Synthesis) constitutes a novel development in the application of computers to organic chemistry achieved by the authors of this review.[31-34] The formal-logical approach, developed previously by Zefirov and Trach,[113-119] is used for the formal specification of the chemical information used in the given system. The general treatment of the chemical processes, based on this approach, made it possible to create the given universal computer system, which is suitable for the solution of any problems of organic chemistry formalised as a constructive list of bond redistributions in chemical structures. The principal types of problems to the solution of which the "FLAMINGOES" system can be applied are, firstly, computer-assisted synthesis, secondly, the study of rearrangements, and, thirdly, the prediction of the mechanisms of complex multistage reactions.

We shall consider briefly the representation and analysis of chemical reactions in the "FLAMINGOES" system. The most important types of organic reactions are formally described as a result of a cyclic bond redistribution (CBR) in the initial chemical system, which leads to the formation of the final chemical system.[113] Consequently, an alternative approach to the solution of the CS problem (alternative to the use of a LT or a set of reaction generators) is a systematic unrepeated generation of all possible CBR in the specified chemical system, which permits the finding of a complete set of the possible precursors of the specified structure (in retrosynthesis) or the complete set of products of the possible reactions in the initial chemical system (in forward CS). If the results obtained are promising for further study, the procedure involving the generation of transformations (in the form of CBR) can be extended to several stages.

The "FLAMINGOES" system, which is based on a systematic unrepeated generation of CBR, has no direct analogies among foreign programs. Retaining all the advantages of the non-empirical approach (the generation of transformations independently of data banks, the possibility of operating with new or little investigated reactions), the "FLAMINGOES" system approaches empirical programs as regards the breadth of coverage of chemical data, since ~90% of the known chemical reactions can be described as processes with CBR and hence such processes can be generated by the "FLAMINGOES" system in the course of the solution of problems of computer-assisted synthesis.

We shall consider the principal types of selection criteria used in the "FLAMINGOES" system. Firstly, the formal criterion designed to test that the correct valences of the atoms and bond orders are used is introduced.

Secondly, the program uses the following empirical criterion: in order to limit the number of formally correct but chemically uninteresting solutions, the "FLAMINGOES" system can be designed to generate not all the CBR possible in the specified chemical system but only those which correspond to the specified limitations to the size of the CBR and also the limitations to the number of bonds formed and the number of bonds ruptured. The "FLAMINGOES" system is interactive and the chemist can introduce all the limitations into the program in the dialogue mode. It is noteworthy that the permissible dimensions of the CBR and also the numbers of formed and ruptured bonds define the set of permissible transformations in the specified chemical system.[113]

Thirdly, the "FLAMINGOES" computer system uses the equivalence criterion. The generation of the CBR in the chemical system for which the corresponding graph has several automorphisms inevitably leads to repeated results. The operation of the equivalence criterion has been organised in the "FLAMINGOES" system in such a way that one can avoid the duplication of the results already in the CBR generation stage; in order to apply this selection criterion, use is made of the information about the group of automorphisms of the graph corresponding to the initial chemical system. It must be emphasised that all the most important selection criteria in the "FLAMINGOES" program operate at the earliest possible generation stages long before the explicit construction of the final chemical system.

The "FLAMINGOES" program is written in the BASIC language and is designed for the USSR produced personal computer "Iskra 226" (64 K). The present version of the "FLAMINGOES" system includes the main program and nine principal subprograms and occupies 60 kbyte of the computer memory. The average generation time is 3–5 s per result. We shall now consider the principal program units comprising the "FLAMINGOES" system.

1. The input of the initial chemical system. In this unit the chemist introduces the structure of the initial chemical system in the form of a figure on the graphical display. The program uses the Adjoint matrix of the multigraph corresponding to the initial chemical system in order to represent the structure. In this unit the user also specifies the number of stages of the computer-assisted synthesis.

2. Construction of the group of automorphisms. In the course of the execution of the given subprogram, all the automorphic features of the graph corresponding to the initial chemical system are automatically generated. Information about the symmetry of the graph is essential for the generation of only non-equivalent CRB, i.e. for the application of the corresponding selection criterion. The unit designed to construct the group of automorphic features is based on the "AVTOGRAF" [?AUTOGRAPH. Ed. of Translation] FORTRAN package.[120]

3. The input of information necessary for the application of the empirical selection criteria:

(a) The limitation of the applicability of transformations in the specified chemical system (see above): in the present version of the "FLAMINGOES" program, heteroatoms and atoms linked by a multiple bond are automatically revealed as potential reaction centres. The chemist is then able to correct arbitrarily, in the dialogue mode, the set of centres put forward by the computer. In addition, the chemist selects from the set of potential reaction centres those which can undergo a change in their valence. For each selected centre, it is necessary to indicate the permissible variant of the change in valence.

(b) The limitation of the types of transformations generated in the specified chemical system (see above). The chemist is able to introduce the limitations on the permissible size of the CBR and on the number of bonds formed and ruptured.

It is noteworthy that provision is made in the program for the "search without limitations" mode in the course of which all formally possible CBR are generated (with ring sizes from three-membered to eight-membered). This mode makes it possible to discover all the strategic pathways to the synthesis of the specified compound which are possible in principle, especially for structures with few atoms for which the number of solutions will not be unduly great.

4. Unrepeated generation of CBR. This subprogram is the most important combinatorial unit in the "FLAMINGOES" computer system. The rapidity of the operation of this unit determines the rapidity of the operation of the program as a whole. A lexicographic inspection of the "sorting tree", in the course of which all possible CBR are generated taking into account the formal and empirical selection criteria, is effected in module A. Module B applies the equivalence criterion on the basis of information about the group of automorphisms of the graph.

5. The output of the final chemical systems. The structures of the final chemical systems can ae output (if the user so wishes) by the printing device (in the form of digital codes for the corresponding CBR), on the graphical display, and on the printer-plotter (in the form of figures).

6. The selection of results for investigation in the next stage (multi-stage CS). The chemist can select from the set of generated results the final chemical systems which are of interest for further study. The selected chemical systems are stored in a special block within the program itself; when the possibilities for the unrepeated generation of CBR in the specified chemical system have been exhausted, the program selects in succession structures encoded in the above block and introduces them as the initial chemical system in the next stage of the computer search.

The "FLAMINGOES" system has extensive possibilities for the solution of a wide variety of problems in synthetic organic chemistry and for the determination of the mechanisms of organic reactions. The initial chemical system, specified by the chemist in the dialogue mode (see paragraph 1), can contain up to 32 centres, the centres being individual atoms or whole atomic groups, introduced as a single whole. Taking into account the possibilities of the "Iskra-226" personal computer, the number of stages in computer-assisted synthesis which can be stored simultaneously in the operating computer memory does not exceed six, but, when external memory devices are used (for example a magnetic floppy disc), the number of CS stages can be arbitrarily large. The "FLAMINGOES" system makes it possible to operate with any organic structures, including complex polyatomic compounds, heterocyclic and organometallic compounds, and ionic and radical species.

In order to demonstrate the possibilities of the "FLAMINGOES" system, a computer search was carried out and a study was made of the pathways leading to the synthesis of a series of skeletal compounds and rearrangements in families of isomers and of the possible mechanisms of organic reactions with participation of ionic and radical species.[31–34] The investigation demonstrated a satisfactory level of the predictive power of the "FLAMINGOES" system. In the course of the operation of the program, not only were virtually all the literature data on the transformations investigated reproduced but also new novel methods of synthesis, rearrangements, and mechanisms of the reactions of the organic compounds investigated were predicted.[121,122]

IX. PROSPECTS FOR THE DEVELOPMENT OF COMPUTER-ASSISTED SYNTHESIS

At the present time there is no doubt whatsoever that the application of computers is essential for the solution of problems in the planning of synthesis, in the prediction of reaction pathways, and in the study of rearrangements. Furthermore, the formalised representation of chemical information in the computer permits strategic planning of a chemical experiment using new or little investigated processes.

As stated above, computer-assisted synthesis is developing in two directions, namely as empirical and non-empirical CS. At the beginning of their development, the programs

for computer-assisted synthesis must fulfill the role not only of the apparatus for the planning of the synthesis, but, to a greater extent, they must serve as an automated reference device for synthetic methods in organic chemistry. Naturally, this role can be managed only by empirical CS systems. However, highly organised data banks for chemical reactions, for example "REACCS",[123] specially designed to be used in the search for known reactions on the basis of specified structural fragments, reaction conditions, etc., appeared in the 1980's. Therefore, the problem of the search for a novel strategy for synthesis, the problem whose exhaustive solution is within the capacity of precisely the non-empirical programs—the "artificial intelligence" systems—has come to the fore at the present time in connection with the solution of problems associated with the planning of syntheses. Therefore, despite the fact that the empirical procedures have been developed to an incomparably greater extent, we believe that the future belongs to a combination of these approaches or in other words to the "semiempirical" CS programs. The enrichment of the non-empirical programs by empirical selection criteria permitting a combination of the advantages of the exhaustive unbiased combinatorial sorting with a system of expert evaluations, developed as a result of the entire experience gained in organic chemistry, appears to be especially promising from this point of view.

REFERENCES

1. E.J.Corey and W.T.Wipke, Science, 1969, **166**, 178.
2. M.Bersohn and A.Esack, Chem.Rev., 1976, **76**, 269.
3. J.Gasteiger, Chim.Ind., 1982, **64**, 714.
4. R.R.Kostikov, T.V.Mandel'shtam, and V.V.Razin, in "Sovremennye Problemy Organicheskoi Khimii" (Current Problems in Organic Chemistry), 1982, No.7, p.4.
5. J.Haggin, Chem.Eng.News, 1983, **61**(9), 7.
6. E.Zeigler, in "Computer in der Chemie", Springer, Berlin, 1984, p.280.
7. A.A.Ovchinnikov and A.I.Boldyrev, Uspekhi Khim., 1986, **55**(4), 539 [Russ.Chem.Rev., 1986(4)].
8. J.H.Winter, in "Chemische Syntheseplanung", Springer, Berlin, 1982, p.215.
9. R.K.Mackie and D.M.Smith, "Guidebook to Organic Synthesis" (Translated into Russian), Izd.Mir, Moscow, 1985, p.352.
10. J.B.Hendrickson, J.Amer.Chem.Soc., 1977, **99**, 5439.
11. W.T.Wipke, H.Braun, G.Smith, et al., ACS Symp. Ser., 1977, **61**, 97.
12. A.Weise, Z.Chem., 1973, **13**, 155.
13. A.Weise, Z.Chem., 1975, **15**, 333.
14. G.Westphal, A.Klebsch, A.Weise, et al., Z.Chem., 1977, **17**, 295.
15. A.Weise and H.G.Scharnow, Z.Chem., 1979, **19**, 49.
16. W.Schubert, MATCH, 1979, **6**, 213.
17. W.Schubert and I.Ugi, Chimia, 1979, **33**, 183.
18. T.D.Salatin and W.L.Jorgensen, J.Org.Chem., 1980, **45**, 2043.
19. T.D.Salatin, D.McLaughlin, and W.L.Jorgensen, J.Org.Chem., 1981, **46**, 5284.
20. C.E.Peishoff and W.L.Jorgensen, J.Org.Chem., 1983, **48**, 1970.
21. J.S.Burnier and W.L.Jorgensen, J.Org.Chem., 1983, **48**, 3923.
22. J.S.Burnier and W.L.Jorgensen, J.Org.Chem., 1984, **49**, 3001.
23. C.E.Peishoff and W.L.Jorgensen, J.Org.Chem., 1985, **50**, 1056.
24. C.E.Peishoff and W.L.Jorgensen, J.Org.Chem., 1985, **50**, 3174.
25. M.G.Bures, R.L.Roos-Kozel, and W.L.Jorgensen, J.Org.Chem., 1985, **50**, 4490.
26. A.Gushurst and W.L.Jorgensen, J.Org.Chem., 1986, **51**, 3513.
27. R.B.Agnihotri and R.L.Motard, ACS Symp.Ser., 1980, **124**, 193.
28. J.Brandt, J.Friedrich, J.Gasteiger, et al., ACS Symp. Ser., 1977, **61**, 33.
29. J.Gateiger and C.Jochum, Top.Curr.Chem., 1978, **74**, 93.
30. M.Marsili, J.Gasteiger, and R.E.Carter, Chim.Oggi., 1984(9), 11.
31. E.V.Gordeeva, in "Tezisy Dokladov Vsesoyuznoi Konferentsii po Molekulyarnym Peregruppirovkam" (Abstracts of Reports at the All-Union Conference on Molecular Rearrangements), Erevan, 1985, p.25.
32. E.V.Gordeeva, in "Materialy Konferenstii Molodykh Uchenykh Khimfaka MGU" (Proceedings of the Conference of Young Scientists in the Faculty of Chemistry of Moscow State University), Moscow, 1985, Part 1, pp.53–56; Deposited at the All-Union Institute of Scientific and Technical Information (VINITI), 1985, No.8373-V.
33. N.S.Zefirov, E.V.Gordeeva, and S.S.Trach, in "Tezisy Dokladov VII Vsesoyuznoi Konferentsii 'Ispol'zovanie Vychislitel'nykh Mashin v Khimicheskikh Issledovaniyakh i Spektroskopii Molekul'" (Abstracts of Reports at at the VIIth All-Union Conference on 'The Use of Computers in Chemical Research and Molecular Spectroscopy'), Riga, 1986, p.19.
34. E.V.Gordeeva, Candidate's Thesis in Chemical Sciences, Lomonosov Moscow State University, Moscow, 1986.
35. E.J.Corey, Quart.Rev.Chem.Soc., 1971, **25**, 455.
36. E.J.Corey, R.D.Cramer, III, and W.J.Howe, J.Amer. Chem.Soc., 1972, **94**, 440.
37. E.J.Corey, W.J.Howe, and D.A.Pensak, J.Amer. Chem.Soc., 1974, **96**, 7724.
38. R.D.Stolow and L.J.Joncas, J.Chem.Educ., 1980, **57**, 868.
39. E.J.Corey and D.A.Pensak, ACS Symp.Ser., 1977, **61**, 1.
40. A.K.Long, S.D.Rubenstein, and L.J.Joncas, Chem. Eng.News, 1983, **61**(9), 22.
41. E.J.Corey, A.K.Long, and S.D.Rubenstein, Science, 1985, **228**, 418.
42. G.Moreau, Nouv.J.Chim., 1978, **2**, 187.
43. F.Choplin, C.Laurenco, R.Marc, et al., Nouv.J.Chim., 1978, **2**, 285.
44. F.Choplin, P.Bonnet, M.H.Zimmer, and G.Kaufmann, Nouv.J.Chim., 1979, **3**, 223.
45. M.H.Zimmer, F.Choplin, P.Bonnet, and G.Kaufmann, J.Chem.Inf.Comp.Sci., 1979, **19**, 235.
46. F.Choplin, S.Goundiam, and G.Kaufmann, J.Chem. Inf.Comp.Sci., 1983, **23**, 26.
47. C.Laurenco, L.Villien, and G.Kaufmann, Tetrahedron, 1984, **40**, 2721.
48. C.Laurenco, L.Villen, and G.Kaufmann, Tetrahedron, 1984, **40**, 2731.
49. T.H.Varkony, D.H.Smith, and C.Djerassi, Tetrahedron, 1978, **34**, 841.

50. T.H.Varkony, R.E.Carhart, D.H.Smith, and C.Djerassi, J.Chem.Inf.Comp.Sci., 1978, **18**, 168.
51. R.Govind and G.Powers, ACS Symp.Ser., 1977, **61**, 81.
52. R.Barone, M.Chanon, and J.Metzger, Tetrahedron Lett., 1974, 2761.
53. R.Barone, M.Chanon, P.Cadiot, and J.M.Cense, Bull.Chem.Soc.Belg., 1982, **91**, 333.
54. R.Barone, M.Chanon, and M.L.Contreras, Nouv. J.Chim., 1984, **8**, 311.
55. H.L.Gelernter, N.S.Sridharan, A.J.Hart, et al., Top.Curr.Chem., 1973, **41**, 113.
56. H.L.Gelernter, A.F.Sanders, D.L.Larsen, et al., Science, 1977, **197**, 1041.
57. K.K.Agarwal, D.L.Larsen, and H.L.Gelernter, Comput.and Chem., 1978, **2**, 75.
58. H.L.Gelernter, S.S.Bhagwat, D.L.Larsen, and G.A.Miller, Anal.Chem.Symp.Ser., 1983, **15**, 35.
59. L.B.Hendrickson and E.Braun-Keller, J.Comput. Chem., 1980, **1**, 323.
60. J.B.Hendrickson, E.Braun-Keller, and A.G.Toczko, Tetrahedron, 1981, **37**, Suppl.No.1, 359.
61. J.B.Hendrickson, D.L.Grier, and A.G.Toczko, J. Amer.Chem.Soc., 1985, **107**, 5228.
62. R.Doenges, B.Groebel, H.Nickelsen, and J.Sander, J.Chem.Inf.Comp.Sci., 1985, **25**, 425.
63. M.Bersohn, Bull.Chem.Soc.Japan, 1972, **45**, 1897.
64. M.Bersohn, E.Esack, and J.Luchini, Comput.and Chem., 1978, **2**, 105.
65. M.Bersohn, ACS Symp.Ser., 1979, **112**, 341.
66. P.E.Blower and H.W.Whitlock, J.Amer.Chem.Soc., 1976, **98**, 1499.
67. V.N.Piottukh-Peletskii, V.I.Smirnov, and V.A.Koptyug, in "Tezisy Dokladov VII Vsesoyuznoi Konferentsii 'Ispol'zovanie Vychislitel'nykh Mashin v Khimicheskikh Issledovaniyakh i Spektroskopii Molekul'" (Abstracts of Reports at the VIIth All-Union Conference on 'The Use of Computers in Chemical Research and Molecular Spectroscopy'), Riga, 1986, p.3.
68. E.J.Corey, W.T.Wipke, R.D.Cramer, III, and W.J.Howe, J.Amer.Chem.Soc., 1972, **94**, 421.
69. L.M.Masinter, J.Lederberg, N.S.Sridharan, and D.H.Smith, J.Amer.Chem.Soc., 1974, **96**, 7702.
70. M.Bersohn and A.Esack, Chem.Scripta, 1974, **6**, 122.
71. A.Esack and M.Bersohn, J.Chem.Soc., Perkin Trans. I, 1975(12), 1124.
72. M.Bersohn, Comput.and Chem., 1978, **2**, 113.
73. M.Bersohn, J.Chem.Soc., Perkin Trans.I, 1979(8), 1975.
74. W.T.Wipke and T.M.Dyott, J.Amer.Chem.Soc., 1974, **96**, 4825.
75. W.T.Wipke and T.M.Dyott, J.Amer.Chem.Soc., 1974, **96**, 4834.
76. J.B.Hendrickson and A.G.Toczko, J.Chem.Inf. Comp.Sci., 1983, **23**, 171.
77. E.J.Corey and G.A.Peterson, J.Amer.Chem.Soc., 1972, **94**, 450.
78. E.J.Corey, W.T.Wipke, R.D.Cramer, III, and W.J.Howe, J.Amer.Chem.Soc., 1972, **94**, 431.
79. M.Bersohn, J.Chem.Soc., Perkin Trans.I, 1973(12), 1239.
80. A.Esack and M.Bersohn, J.Chem.Soc., Perkin Trans. I, 1974(21), 2463.
81. A.Esack, J.Chem.Soc., Perkin Trans.I, 1975(12), 1120.
82. M.Bersohn and A.Esack, Chem.Scripta, 1976, **9**, 211.
83. M.Bersohn, J.Chem.Soc., Perkin Trans.I, 1982(2), 631.
84. C.Jochum and J.Gasteiger, J.Chem.Inf.Comp.Sci., 1979, **19**, 43.
85. C.Jochum and J.Gasteiger, J.Chem.Inf.Comp.Sci., 1979, **19**, 49.
86. B.L.Roos-Kozel and W.L.Jorgensen, J.Chem.Inf. Comp.Sci., 1981, **21**, 101.
87. J.B.Hendrickson, D.L.Grier, and A.Toczko, J.Chem. Inf.Comp.Sci., 1984, **24**, 195.
88. M.Bersohn and A.Esack, Comput.and Chem., 1977, **1**, 103.
89. F.Harary and E.M.Palmer, in "Graphical Enumeration" (Translated into Russian), Izd.Mir, Moscow, 1977, p.324.
90. B.J.Hendrickson, J.Amer.Chem.Soc., 1971, **93**, 6847.
91. J.B.Hendrickson, J.Amer.Chem.Soc., 1975, **97**, 5763.
92. J.B.Hendrickson, J.Amer.Chem.Soc., 1975, **97**, 5784.
93. J.B.Hendrickson, Top.Curr.Chem., 1976, **62**, 49.
94. J.B.Hendrickson, J.Chem.Inf.Comp.Sci., 1979, **19**, 129.
95. J.B.Hendrickson, Acc.Chem.Res., 1986, **19**, 274.
96. E.J.Corey and W.L.Jorgensen, J.Amer.Chem.Soc., 1976, **98**, 189.
97. E.J.Corey, H.W.Orf, and D.A.Pensak, J.Amer. Chem.Soc., 1976, **98**, 210.
98. E.J.Corey and W.L.Jorgensen, J.Amer.Chem.Soc., 1976, **98**, 203.
99. E.J.Corey, A.K.Long, T.W.Greene, and J.W.Miller, J.Org.Chem., 1985, **50**, 1920.
100. J.Gasteiger, Comput. and Chem., 1978, **2**, 85.
101. J.Gasteiger and O.Dammer, Tetrahedron, 1978, **34**, 2939.
102. J.Gasteiger, P.Jacob, and U.Strauss, Tetrahedron, 1979, **35**, 139.
103. J.Gasteiger, Tetrahedron, 1979, **35**, 1419.
104. J.Gasteiger and M.Marsili, Tetrahedron Lett., 1978(34), 3181.
105. J.Gasteiger and M.Marsili, Tetrahedron, 1980, **36**, 3219.
106. J.Gasteiger and M.G.Hutchings, Tetrahedron Lett., 1983, **24**, 2541.
107. J.Gasteiger and H.Saller, Angew Chem., 1985, **97**, 699.
108. J.Gasteiger and M.G.Hutchings, Tetrahedron Lett., 1983, **24**, 2537.
109. M.D.Guillen and J.Gasteiger, Tetrahedron, 1983, **39**, 1331.
110. J.Gasteiger and M.G.Hutchings, J.Chem.Soc., Perkin Trans.II, 1984(3), 559.
111. J.Gasteiger and M.G.Hutchings, J.Amer.Chem.Soc., 1984, **106**, 6489.
112. W.J.Mortier, K.van Genechten, and J.Gasteiger, J.Amer.Chem.Soc., 1985, **107**, 829.
113. N.S.Zefirov and S.S.Tratch, Zhur.Org.Khim., 1981, **17**, 2465.
114. N.S.Zefirov and S.S.Tratch, Zhur.Org.Khim., 1975, **11**, 225.
115. N.S.Zefirov and S.S.Tratch, Zhur.Org.Khim., 1975 **11**, 1785.
116. N.S.Zefirov and S.S.Tratch, Zhur.Org.Khim., 1976, **12**, 7.
117. N.S.Zefirov and S.S.Tratch, Zhur.Org.Khim., 1976, **12**, 697.
118. N.S.Zefirov and S.S.Tratch, Chem.Scripta, 1980, **15**, 4.
119. S.S.Tratch and N.S.Zefirov, Zhur.Org.Khim., 1982, **18**, 1561.

120. N.P.Zhidkov, N.S.Zefirov, A.I.Popov, et al., in "Matematicheskie Voprosy Strukturnogo Analiza i Algoritmy Mashinnykh Eksperimentov v Organicheskoi Khimii" (Mathematical Problems in Structural Analysis and Algorithms for Machine Experiments in Organic Chemistry) (Edited by N.P.Zhidkov et al.), Izd. Moskov.Gos.Univ., Moscow, 1979, p.59.
121. N.S.Zefirov, E.V.Gordeeva, G.S.Kuznetseva, and S.I.Kozhushkov, Zhur.Org.Khim.(in press).
122. N.S.Zefirov and E.V.Gordeeva, J.Org.Chem.(in press).
123. R.V.Venkataratnam, G.Thyagarajan, and A.Husain, CEW, Chem.Eng.World, 1985, 20(6), 61.

Zelinskii Institute of
Organic Chemistry,
USSR Academy of Sciences,
Moscow