Computer assistance in the design of syntheses and a new generation of computer programs for the solution of chemical problems by molecular logic

Ivar K. Ugi, Johannes Bauer, Regina Baumgartner, Eric Fontain, Dietmar Forstmeyer and Severin Lohberger

Organisch-Chemisches Institut der Technischen Universität München, 8046 Garching, F.R.G.

ABSTRACT - Molecular logic, the concept and its mathematical representation, are discussed. A status report is given on the development of computer programs for the deductive solution of chemical problems by molecular logic. The program PEMCD (Program for determining the Exact Minima of Chemical Distance) has been implemented recently; preparations for an improved version have already begun. IGOR 2 (Interactive Generation of Organic Reactions), a user-friendly PC version of IGOR has evolved and is soon ready to be distributed. The implementation of RAIN 2 (Reactions And Intermediates Networks) is in its early stages. Besides its use in the prediction and mechanistic elucidation of chemical reactions, including metabolic pathways and retrosynthetic analyses, it will be incorporated as a module into a system for the bilateral design of syntheses, together with a substructure correlation module that still is under development.

INTRODUCTION

The reaction library based synthesis design programs belong to the best known type of computer assistance for organic synthesis (ref. 1). In a recent review, Zefirow (ref. 2) mentioned more than a dozen such retrosynthetic programs. However, besides these information oriented programs, there exist also some synthesis design programs that are based on molecular logic (refs. 1-10). Moreover, many chemistry oriented computer programs are useful for syntheses, although they have not been explicitly devised for this purpose (refs. 8-10).

We use the term molecular logic for the inherent logic of chemistry: While molecular logic refers to the logic that chemists use ad hoc in their reasoning and in the solution of individual chemical problems, the logical structure of chemistry is a global edifice of relations, mostly equivalence relations; it provides the general framework for molecular logic and the deductive solution of chemical problems. Molecular logic and the logical structure of chemistry are both consequences of the valence chemical properties of the chemical elements and some physical principles that determine the behaviour of chemical systems.

The discovery of the logical structure of chemistry began with the brilliant insights of Berzelius, Liebig, Wöhler, Couper, Gerhard, Frankland and Kekulé who created the "structural theory of chemistry". This theory and the underlying concept of chemical constitution was officially accepted in 1860 by the chemical world at the Congress of Chemistry (refs. 11, 12), a forerunner of the IUPAC conferences. Since then, the structural theory is an integral part of all chemical reasoning. The addition of stereochemical ideas completed the essentials of conceptual chemistry from 1874 on (ref. 13).

The development of quantum theory and new types of knowledge about mole-
cules, as obtained through physico-chemical and computational methods, are
responsible for the more recent refinements of molecular logic.

The representation of the logical structure of chemistry by qualitative
mathematical models and theories began with Cayley's (ref. 14) molecular
graphs that led to the elegant present day applications of graph theory in
chemistry by Balaban (ref. 15), Kvasnička et al. (ref. 16) and many
others. Polyá's enumeration theory (ref. 17) that was extended and perfec-
ted by Ruch (ref. 18) marks the onset of the direct qualitative mathemati-
cal treatment of stereochemistry.

THEORY OF THE BE- AND R-MATRICES

In 1973 Dugundji and Ugi (ref. 19) formulated the theory of the BE- and R-
matrices, an algebraic model of the logical structure of constitutional
chemistry. This qualitative mathematical theory of constitutional che-
mistry pertains to the chemistry of a fixed collection $A = \{A_1...,A_n\}$ of
atoms. An ensemble of molecules (EM) that is based on A contains each atom
of A once and only once.

The Family of Isomeric EM of A, FIEM (A) is the complete set of all such
EM. The FIEM(A) contains the entire conceivable chemistry of A. Since A
can be any collection of atoms, the theory of the BE- and R-matrices, a
theory of the FIEM (A) covers all of chemistry and is a universal
model of the logical structure of constitutional chemistry.

The BE-matrix B of an EM(B) $\in$ FIEM(A) is an nxn symmetric matrix whose
rows/columns are assigned to the atoms of A; the i-th row/column belongs
to A_i. The off-diagonal integer entry $b_{ij} = b_{ji}$ (i $\neq$ j) of B is the formal
order of the covalent bond between A_i and A_j. A single covalent bond A_i-A_j
corresponds to a pair of valence electrons that is shared by A_i and A_j.
The diagonal entry b_{ii} is the number of lone valence electrons at A_i. Thus
the sum over the entries of B is the total number of valence electrons in
EM(B). In contrast to an adjacency matrix or a connectivity matrix, a BE-
matrix is a genuine mathematical entity with well-defined algebraic pro-
perties.

The chemical constitution of

$$H - C \equiv N\mathsf{l}$$

is thus represented by

$$B\ (HCN)\ =\ \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 3 \\ 0 & 3 & 2 \end{bmatrix},$$

an equivalent of its constitutional formula (with $H=A_1$, $C=A_2$ and $N=A_3$).

Since n atoms can be indexed in up to n! distinct ways, a given EM can be
described by up to n! distinct but equivalent BE-matrices. Any two BE-ma-
trices B and B' represent the same EM, and are equivalent if they are
interconvertible by row/column permutations that correspond to a permuta-
tion of the atomic indices in $A = \{A_1,...,A_n\}$. One of the atomic indexings
can be defined as the canonical indexing, e.g. by the Morgan algorithm
(ref. 20) or CANON (ref. 21).

$$B = P \cdot B' \cdot P^{-1} \text{ (P is a permutation matrix!)}$$

We extend the concept of isomerism from molecules to EM. Then any stoichiometrically balanced chemical reaction EM (B) —> EM (E) is an interconversion of isomeric EM $\in$ FIEM(A). The above reaction is represented by the transformation of the BE-matrix $\mathbf{B}$ of the educt EM(B) into the BE-matrix $\mathbf{E}$ of the product EM(E) by addition of a so-called reaction matrix (R-matrix) $\mathbf{R}$, according to $\mathbf{B} + \mathbf{R} = \mathbf{E}$.

The off-diagonal entry $r_{ij} = r_{ji}$ of the nxn symmetric R-matrix $\mathbf{R} = \mathbf{E} - \mathbf{B}$ indicates the change in the formal order of the covalent bond between A_i and A_j, and the diagonal entry A_i is the change in the number of lone valence electrons at A_i. Accordingly an R-matrix corresponds to the change in the distribution of the valence electrons during the reaction EM(B) —> EM(E), i.e. an R-matrix describes a redistribution pattern of valence electrons, a pattern of "electron pushing arrows". An irreducible R-matrix is obtained from an R-matrix by removing all rows/columns that contain only zeros. The rows/columns of an irreducible R-matrix belong to the reactive centers that form the core of the reaction. An irreducible R-matrix corresponds to a category of reactions, that have in common some electron redistributions scheme as represented by $\mathbf{R}$.

The equation $\mathbf{B} + \mathbf{R} = \mathbf{E}$ describes only a real chemical reaction if it complies with (a) the mathematical fitting condition and (b) the valence chemical boundary conditions.

  (a) The negative entries $r_{ij} > 0$ of $\mathbf{R}$ must coincide with entries $b_{ij} \geq |r_{ij}|$ of $\mathbf{B}$, in order to ensure $e_{ij} = b_{ij} + r_{ij} \geq 0$. In essence, this means one cannot break bonds that do not exist, and one cannot push electrons from where there are none.

  (b) Each row/column of a BE-matrix must correspond to one of the allowable valence schemes of the chemical element that has been assigned that row/column. The allowable valence schemes may differ from case to case and may be chosen ad hoc. The sets of valence schemes of stable molecules are generally different from those of reactive intermediates (ref. 22).

In our most recent generation of computer programs for the deductive solution of chemical problems the valence chemical boundary conditions are contained in the so-called transition tables that are part of the reaction generators (ref. 23). A transition table contains the allowed valence schemes of the respective chemical element for the educts and the products of the reactions considered, as well as the forbidden, or the permitted transitions during chemical reactions. For each considered chemical element, a transition table is provided, either as a standard transition table, or as an ad hoc transition table, formulated by the user to suit a given problem.

Since EM(B) and EM(E) contain the same number of valence electrons, we have

$$\sum_{ij} e_{ij} - \sum_{ij} e_{ij} = \sum_{ij} (e_{ij} - b_{ij}) = \sum_{ij} r_{ij} = 0.$$

However, for chemical reactions,

$$d(E,B) = \sum_{ij} |r_{ij}| \geq 0$$

is twice the number of valence electrons that are redistributed during the interconversion EM(B) $\rightleftharpoons$ EM(E). Since $d(\mathbf{E,B})$ has the mathematical properties of a distance, and has immediate chemical meaning, we call $d(E,B)$ the chemical distance (CD) between $\mathbf{B}$ and $\mathbf{E}$.

With a fixed atomic indexing of EM(B), i.e. a fixed assigment of the rows/columns of B to the atoms of $A = \{A_1...,A_n\}$, the CD between B and E depends on the indexing of the atoms in EM(E). Thus the chemical distance $d(B,P \cdot E \cdot P^{-1})$ is a function of the permutation matrix P, that represents the reindexing of A. The theory of the BE- and R-matrices can also be expressed in geometric terms. Then CD does not only have a mathematical meaning but can also be visualized as a so-called "city-block distance" (refs. 19, 24,25). The nxn BE-matrices B and E of EM(B) and EM(E) correspond to the BE-points $P(B)$ and $P(E)$ in an n^2-dimensional Euclidean space, and the R-matrices are there vectors from $P(B)$ to $P(E)$. EM(B) is representable by up to n! equivalent BE-matrices $B' = P \cdot B \cdot P^{-1'}$. Accordingly, we have a cluster of BE-points $P(P \cdot B \cdot P^{-1})$. For each point $P(E)$ there is a point $P(B)$ in the cluster of EM (B) at $d(B,E)$, the minimal CD (MCD) between the clusters of B and E. Without specifying a particular atom-by-atom mapping of EM(B) onto EM(E), the term CD means the MCD.

An R-matrix is representable as a sum of its components.

$$R = R_1 + \ldots + R_m$$

The successive transformation of B by the components of R corresponds to a sequence of chemical reactions, or mechanistic steps. This is important for step-wise elaboration of reactions.

$$B + R_1 = E_1; \; E_1 + R_2 = E_2; \; E_2 + R_2 = E_2;\ldots; \; E_m + R_m = E$$

Under (a) and (b) these additive transformations commute only if they are independent of each other. For example the reaction mechanism

$$H-C\equiv N| \longrightarrow \overset{\oplus}{H} + |\overset{\ominus}{C}\equiv N| \longrightarrow |\overset{\ominus}{C}\equiv\overset{\oplus}{N}-H$$
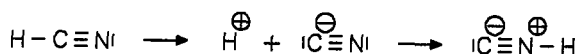
is represented by

$$B(HCN)+R_1=E_1(H^{\oplus}+CN^{\ominus}), \; E_1(H^{\oplus}+CN^{\ominus})+R_2=E(HNC), \text{ and } B(HCN)+R=E(HNC),$$

with $E_1(H^{\oplus}+CN^{\ominus}) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 2 & 3 \\ 0 & 3 & 2 \end{bmatrix}$, $\quad E = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 2 & 3 \\ 1 & 3 & 0 \end{bmatrix}$,

$$R_1 = \begin{bmatrix} 0 & -1 & 0 \\ -1 & +2 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad R_2 = \begin{bmatrix} 0 & 0 & +1 \\ 0 & 0 & 0 \\ +1 & 0 & -2 \end{bmatrix}, \quad R = \begin{bmatrix} 0 & -1 & +1 \\ -1 & +2 & 0 \\ +1 & 0 & -2 \end{bmatrix}$$

Note that $B(HCN) + R_2 = E_2$ would violate (b) because E_2 represents

$$E_2 = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 3 & 0 \end{bmatrix}$$

$$\overset{\ominus}{H}\diagdown \atop C\,\vdots\,N^{\oplus}$$

a molecule with a forbidden valence scheme of the hydrogen atom.

CORRELATION, CLASSIFICATION AND UNAMBIGUOUS REPRESENTATION OF MOLECULAR SYSTEMS AND CHEMICAL REACTIONS

Substructure correlation in synthesis design

The known computer programs for the design of syntheses are based on the retrosynthetic approach. There a tree of precursors is successively elaborated from the target molecule (TM) of the synthesis by applying the retroreactions of the reactions that lead towards the target. These retroreactions are either obtained from a reaction library (ref. 2), or they are manufactured as needed by so-called reaction generators (ref. 23).

Regardless of the source of the retroreactions, the trees of synthetic routes tend to expand very rapidly and in order to avoid combinatorial explosions, the trees of synthetic routes must be heavily pruned by suitable selection procedures. Most of these selection procedures are based on thermochemical estimates, charge affinity patterns (refs. 3,26), or on heuristic rules that are hardly applicable equally well under all conditions (ref. 26). Thus there is a danger that this essentially heuristic tree pruning will sometimes be quite arbitrary.

The bilateral design of syntheses is an attractive alternative the retrosynthetic approach (ref. 7). Provided that the TM, its coproducts and the starting materials (SM) of a synthesis are all known, it is possible to design the syntheses bilaterally, i.e. to generate a network of synthetic pathways that connects the SM with the TM and its coproducts. We had thus originally idea to establish an atom-by-atom mapping of the EM of SM onto the target EM, by the principle of MCD, and to use the components of the corresponding "grand R-matrix" of the synthesis to manufacture a network of the conceivable synthetic routes (ref. 7). This plan was given up, because there are too many mappings that must be considered, also some whose CD > MCD. We now prefer a development of reaction networks with more degrees of freedom.

When a network of reactions is produced by letting trees converge bilaterally from two centers, the number of nodes that are encountered until the trees merge may be as large as in a tree that grows unilaterally from one center until it reaches the other center. This is true for trees, that are growing randomly. Since a synthesis can be imbedded in the chemistry of an FIEM, and since the this corresponds to a lattice of points in a high-dimensional Euclidean space synthetic pathways have "length and direction". By virtue of this metric aspect the combinatorial problem of synthesis design is substantially reduced by a bilateral approach with the theory of the BE- and R-matrices as a foundation (refs. 7, 27-29).

The SM of a synthesis are generally not known when a given TM must be synthetized. In order to apply the bilateral deductive approach to synthesis design, not only the complete EM of SM must be known, but also the coproducts of the TM. Then the problem of the synthesis is reduced to converting the EM of SM into an isomeric EM that contains the TM. The suitable SM for the synthesis of a given TM can be found from a list of available SM by substructure correlation. One first looks for the SM that has the largest substructure in common with the TM. Then a SM is picked that has a maximum substructure in common with the remainder of the TM etc. etc., until a collection of SM is reached whose union of eligible substructures covers the TM. It is like the assembling of jigsaw puzzle. This seems to be a straightforward concept, but in practice the aforementioned search for SM by substructure correlation belongs to the most difficult problems in the field of computer assistance to synthesis.

Substructure correlation is of great importance in many areas, that are related to chemistry, in particular in structure/activity analysis. Sear-

ching a list of molecules that contain a given substructure is relatively
easy to achieve and belongs to the retrieval routines of chemical documen-
tation, whereas in contrast, the correlation of molecules by substructures
that are not known beforehand is a much more difficult task. An early
analysis showed that substructure correlation by the known methods is an
np-complete problem (ref. 32) and is not suitable for most cases of inte-
rest. However, this analysis led also to a then new way of correlating
substructures. It was found that the required memory space and computer
time will only increase polynomially (of degree $\leq$ 5), if the
substructures of all considered molecules are imbedded in a common, conti-
guous oriented network of father/son relations. Then the largest common
substructures of any two molecules are the common nodes that are closest
to both parent molecules. This concept of substructure correlation of
CORREL (ref. 30) belongs now to the state of the art (loc. cit. ref. 8).

Although, in general, the size of the substructure networks does not
increase much by the addition of further molecules when a list of 1800-
2000 molecules is reached, we have found that the number of substructures
will besome too large if a complete network of substructures is generated
from a comprehensive list of available starting materials for organic
synthesis. We are now developing a substructure correlation system for
bilateral synthesis design that is confined to the indispensable SM and to
substructures that are synthon-related. Thus some empirical chemistry
enters logic oriented synthesis design through the back door.

We   have   done extensive model studies   towards   synthesis-oriented
substructure correlation. An already partly executed plan has resulted:

   (a) Guided by synthon considerations, a list of available SM is as-
       sembled from the preparative literature and the catalogs of the
       major supply houses.

   (b) In each SM these bonds are determined that are preferably dissec-
       ted under the aspect of syntheses.

   (c) The SM are partitioned into fragments (FR) according to (b).

   (d) An oriented network of father/son relations is manufactured from
       the FR.

   (e) The FR of the TM are analogously incorporated into the oriented FR
       network, in order to find the SM with the largest common FR and
       its optimum placement in the TM. Then the SM with the largest FR
       in common with the remainder of the target, etc. etc.

According to test runs with relatively small lists of SM, with (a)-(e) a
surprisingly large number of solutions is found. Therefore an automated
selection procedure is developed for the EM of SM. in this procedure the
TM and syntheses are e.g. classified by the chemist according to the
following properties and criteria, and evaluated accordingly:

   - polyfunctional polar, (almost) consonant TM (ref. 31)
   - complex polycyclic TM that require a strategy
   - strained polycyclic TM with constitutional symmetries (ref. 21).
   - TM with dissonant functional groups
   - TM requiring radical phenol coupling
   - TM requiring stereoselective operations

The selection of the preferred SM is accomplished by the use of a set of
evaluation criteria:

   - total CD of the synthesis
   - estimated number of functional groups involved.

- possible convergence of the synthesis
- location of the bonds broken/made, relative to the functional groups and consonant/dissonant parts of the TM
- construction of a polycyclic ring system in the TM
- stereocontrolled synthesis of acyclic TM
- exploitation of constitutional symmetries
- loss of major parts of the SM
- useless functional groups in SM
- preservation, or loss of the stereochemical characteristics of the SM
- relation of the charge affinity patterns in the SM and TM, and consequences for dealing with existent dissonances.

For example, a FR network with 426 nodes is generated from a test set of 73 SM. When the alkaloid ajmaline **1** is subjected to a subgraph correlation with this list of SM,

1
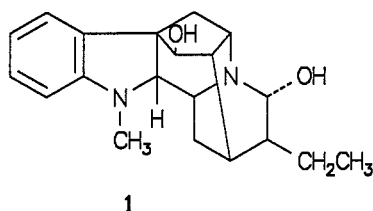
137 assignments FR/TM are found, and 850 correlations SM/TM. Under the assumption that the azole moiety of ajmaline is covered by an indole derivative, 21 mappings SM/TM result. Among these, some of the educts of published ajmaline syntheses are found. Without the azole/indole mapping five further sets of SM are found for ajmaline.

After establishing a set of SM for a given TM, its coproducts must be determined, in order to have a stoichiometrically balanced synthesis. The coproducts of the TM can be found by the difference of the substructures in the SM and the TM. However, this is by no means a trivial task. A corresponding computer program is still at the design stage.

Once the complete EM of the SM and the TM and its coproducts are known, it remains to generate a network of the conceivable syntheses. This can be accomplished by the computer program RAIN (refs. 27- 29) that is described later in this article.

CONSTITUTIONAL EQUIVALENCES AND CHEMICAL SIMILARITY

The classification and canonical representation of molecules and chemical reactions is not only essential for chemical documentation but for almost any computer assistance in chemistry, including the computer assistance des for syntheses.

The chemical constitution of molecules is nowadays preferably represented by so-called topological matrices like the connectivity matrices (ref. 32) and BE-matrices (ref. 19). The required canonical indexing of the atoms can be achieved by the Morgan algorithm (ref. 20) that is primarily based on the unlabelled graph of the molecule, or by the relaxation algorithm CANON (ref. 21) that takes both into account simultaneously, the under-lying graph of the molecule, and the chemical elements that correspond to the labels of the nodes. Thus any constitutionally equivalent atoms are detected by CANON and an indexing at the atoms results that is also suitable for computer assisted stereochemistry (refs. 33,34).

When (EM(B) ⊔ EM(E)), the union of the educts EM(B) and the products EM(E) of a chemical reaction is subjected to a suitably modified version of CANON, the reactive centers in EM(B) and EM(E) are detected as being

non-equivalent in the educts and the products, whereas the atoms that are locally equivalent in EM(B) and EM(E) belong to the parts of the reactants that are invariant under the considered reaction. The latter procedure is the decisive step in establishing the atom-by-atom mappings of EM(B) onto EM(E), with redistribution of a minimum number of valence electrons, and are in accordance with the principle of minimum chemical distance (refs. 24,25), a modern quantitative version of the classical principle of minimum structure change (ref. 35).

The determination of the reactive centers and the invariant parts of reactants, as well as the atom-by atom mappings of the educts onto the products is indispensable for any documentation system of chemical reactions beyond listing educts and products. J. Brandt et al. have recently developed a hierarchic classification and documentation system for chemical reactions (refs. 36,37) that relies on the aforementioned atom-by-atom mappings.

The concept of CD is not only useful for the classification and representation of chemical reactions, and sequences of chemical reactions, e.g. multistep syntheses. Chemical similarity (refs. 38,39) can now be clearly defined in terms of CD (ref. 25). In fact, CD is a quantitative measure of chemical similarity. Note that the precise degree of simularity of a given molecular system can only be determined to a given reference, if all considered systems are isomeric EM. Let the size of common substructures be a measure of similarity, then the non-matching parts of the molecules may still differ to varying extent, unless the systems are isomeric. Thus, similarity based order is only possible within an FIEM, otherwise, at best, there is only semi-order.

HIERARCHICAL CLASSIFICATION OF CHEMICAL REACTIONS

Once an atom-by-atom mapping, or at least the reactive centers, the electron redistribution scheme, and the invariant parts of the reactants of a chemical reaction are established, e.g. by PEMCD (ref 25), it is possible to classify chemical reactions according to a hierarchy of criteria. In a hierarchic documentation system that is based on the theory of the BE- and R-matrices, Brandt et al. (refs. 36,37) use this hierarchy:

- R-category, consisting of reactions that have in common an irreducible R-matrix and electron redistribution scheme.

- RB-class, consisting of reactions that have in common the same intact BE-matrix, i.e. invariant bonds at the core of the reaction.

- RA-class, consisting of the reactions of an RB-class having in common the same chemical elements at the reactive centers.

A somewhat different hierarchic classification of chemical reactions is used in the computer program IGOR (refs. 40,41) where the user follows interactively the hierarchically classified results. In IGOR we have essentially the hierarchy (ref. 41)

- R-category

- basis reaction, characterized by a graph of the educts and products where the nodes are the reactive centers and the edges are the covalent bonds that participate in the reaction

- RA-class

- individual chemical reactions

In the computer program RAIN (refs. 23,41,42) the R-categories can be partitioned according to the T-matrices (refs. 23,42) of the reactions.

REACTION GENERATORS

The reaction generators (RG) solve the fundamental equation $B + R = E$ of the theory of the BE- and R-matrices. The RG of type I (RGI) generates from a given BE-matrix $B = E - R$ the pairs (R,E) that comply with (as) the mathematical fitting condition and (b) the valence chemical boundary conditions. The RG I predicts the reactions that EM(B) can undergo, and they also find any EM from which EM(B) can be formed by a conceivable chemical reaction.

Our first synthesis design program CICLOPS (ref. 43) as well as the synthesis design program EROS (ref. 3) contain some fixed set of R-matrices that are subjected to row/column permutations and are then added to B. The synthesis design program ASSOR (ref. 44) is endowed with an RG I that generates the R-matrices from their basis elements (ref. 19). A similar RG I has recently been incorporated into EROS. Brandt and Stadler (ref. 45) have recently introduced a recursive RGI that has been developed from the system CASTOR (ref. 46). Presently the most efficient RG are the transition table guided RG, the TRG. The development of the TRG began with the TRG II, a TRG of type II of IGOR (refs. 40,42) that generates the pairs (B,E) from a given R-matrix.

A TRG I is the centerpiece of RAIN (ref. 23). This TRG I determines first the allowable valence schemes of the atoms in EM(E) by applying the allowed transitions to the valence schemes of the respective atoms in the given EM(B). For each combination of valence schemes the EM(E) are generated that comply with the theory of the BE- and R-matrices and any other conditions that may be imposed by the user.

The TRG II of IGOR selects for each row/column of a given R-matrix R a set of chemical elements whose transition tables are in agreement with the corresponding row/column of R. For each row/column of (B,E) the union of the applicable transition tables is formed. From this union the valence schemes, their transitions and the usable chemical elements are found for the row/column pairs of (B,E). Subsequently all the BE-pairs (B,E) are produced by exhaustive combination of the valence schemes of the rows/columns. Then chemical elements are assigned to the rows/columns of (B,E), and the unused valences of the reactive centers are saturated as specified by the user.

DEDUCTIVE SOLUTION OF CHEMICAL PROBLEMS AND COMPUTER ASSISTED DESIGN OF SYNTHESES

After a long period of feasibility studies and experience gathering we have arrived at computer programs that deduce the solutions to individual problems from the general principles contained in the theory of the BE- and R-matrices. In general, the unknowns "X" of the problems that can be thus solved are molecular systems, or chemical reactions. The design of syntheses belongs for instance, to that category of problems. In essence, we have two problem-solving computer programs, IGOR and RAIN, that are complementary (ref. 23). The centerpieces of IGOR and RAIN are the two complementary RG, TRGI and TRGII. Also IGOR and RAIN solve complementary sets of problems.

IGOR was first written in PL/I for a CYBER 175. Recently IGOR 2 with a user friendly menu has been implemented for PC in FORTRAN 77 and Assembler. The purpose of IGOR is to produce molecular structures and chemical reactions de novo from an input irreducible R-matrix. IGOR, like RAIN, has an interactive graphic input/output system, such that the user can follow IGOR as it proceeds along the hierarchic classification of its results.

When a zero R-matrix is subjected to IGOR, we have $B = E$, and IGOR puts out constitutional formulas of molecules, according to specifications by the user (ref. 41). Thus the complete set of the 23 isomers $(CH)_8$ and

their interconversions by the 6.6 (six center/six electron) basis reactions (ref. 47), the 17 1.3-dipoles of the elements C,N and O (ref. 41), sydnone and its 51 analogs with the empirical formula $C_2H_2N_2O_2$ (ref. 41), as well as the 278 conceivable five membered cyclic phosphorylating reagents (ref. 48) have been obtained by IGOR from zero R-matrices.

The first computer program for inventing new reactions was a program by Stevens and Brownscombe (ref. 49) that generated analogs of the extrusion reaction, a 6.6 basis reaction, by permuting the chemical elements at the reactive centers of the reactants. It produced the roughly 70000 representatives of the above 6.6 basis reaction. To our knowledge, no unprecedented, experimentally verified reaction resulted from this project.

The attemps to use IGOR to discover unprecedented reactions began with Herges' (ref. 50) computer assisted experimental of study the $(\sigma_2\pi_2 + \pi_2)$-cycloadditions of homofurane, homothiophene and N-ethoxycarbonyl homopyrrole and double bond systems, forming of seven-membered rings (ref. 41). At the time of these investigations we believed that the cycloadditions obtained were unprecedented, and thus they were the first new reactions that had been discovered with computer assistance. However, after publishing our results, H. Prinzbach pointed out that Fowler (ref. 51) had already observed an analogous reaction of N-ethoxycarbonyl-homopyrrole with an acetylene derivative, a reaction that had escaped our first literature search. Our computer programs for the deductive solution often reincent the wheel.

About the same time Herges found the thermal conversion of the α-formyloxyketones into ketones and carbon dioxide, an unprecedented type of reaction that formally is an extrusion reaction. This is the first unprecedented reaction that has been "proposed" by IGOR and realized in the laboratory (ref. 41). The hierarchical classification of chemical reactions permits to judge the degree of novelty of unprecedented reactions (ref. 52). By this classification, the aforementioned thermal decomposition of the α-formyloxy ketones belongs to the known extrusion type 6.6 basis reaction. Thus it is novel, up to the level of basis reactions.

Our IGOR assisted search for unprecedented reactions of a higher degree of novelty led to the formation of 7 from 2 + 3 that takes place at 180° C in almost quantitative yield (refs. 52-54). The underlying basis reaction 8 + 9 -> 10 is, to our best knowledge, not populated by any known reactions. Thus novelty at the level of basis reactions has been achieved. When subjected to RAIN, the network 2 + 3 -> {4,5,6} -> 7 of reaction pathways resulted. Quantum chemical computations suggest that 2 + 3 -> 4 -> 7 is the most likely reaction pathway (ref. 52).
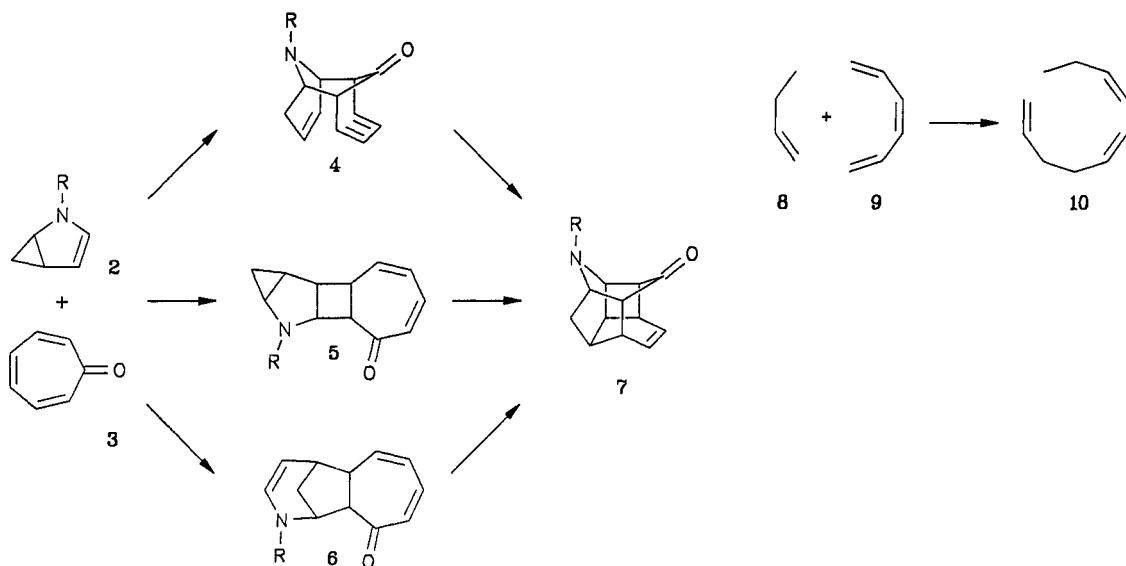
Another reaction that was suggested by IGOR is the formation of the phosphorane derivative **15** from **13** and **14**.
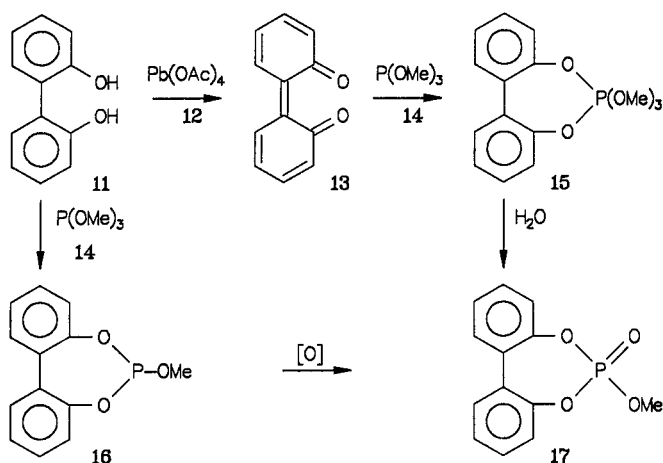

In an attempt to realize this reaction experimentally, **11** was oxidized by **12**, and subsequently **14** was added. On work-up **17** was isolated. It was first believed to have been formed from **15** by hydrolysis, which would have confirmed **13** -> **15**. However, a more detailed investigation revealed that **17** had been formed from unoxidized **11** via **16** by transesterification and oxidation.

RAIN 1 was directly developed on a PC, the COMPAQ 386 in FORTRAN 77 (refs. 27-29). In the meantime RAIN 2 is being implemented. RAIN 2 will contain an improved TRG I, and some additional routines, e.g. recognition of tautomers. The system RAIN generates a network of reaction pathways that connect two isomeric EM. The nodes of the network may be reactive intermediates, or stable molecules only, depending on the transition tables (ref. 23) that are used. The "engine" of RAIN is its TRG I, that produces the networks of conceivable reaction pathways by generating trees of reactions from the educts EM(B) and the products EM(E). Since all encountered molecules are CANONically represented (ref. 21), any mergers of the educt tree and the product tree are immediately noticed. Unless otherwise specified, RAIN selects reaction pathways with a minimum number of nodes. There are many options that the user of RAIN can apply, e.g. diverse lower and upper bounds, for the R- and T-matrices, for CD, and for formal electrical charge on the atoms, and rules that can be used on the chemical constitution and the substructures of molecules (ref. 42). Before being displayed, the network of reaction pathways is subjected to an optimization routine that minimizes the number of intersecting edges in the graph of the network. RAIN is useful for the bilateral design of multistep syntheses from known starting materials and the target EM, for the elucidation of prebiotic biosynthetic and metabolic routes, as well as the analysis of reaction mechanisms.

Besides the bilateral operating mode, RAIN can also be operated in a unilateral mode to predict the conceivable chemical reactions that an EM may undergo, and it is also usable for retrosynthetic studies.

RAIN has already proven its power and usefulness in many test cases (refs. 27-29,42). The Streith-Defoin reaction (ref. 55) has played a particularly important role in the development of PEMCD (ref. 25) and of RAIN (refs. 27-29).The prebiotic synthesis of adenine from hydrogen cyanide (ref. 56) is a particularly interesting example, for which we only have preliminary results. The complexity of this problem is illustrated by one of the subnetworks that have been obtained (ref. 42).
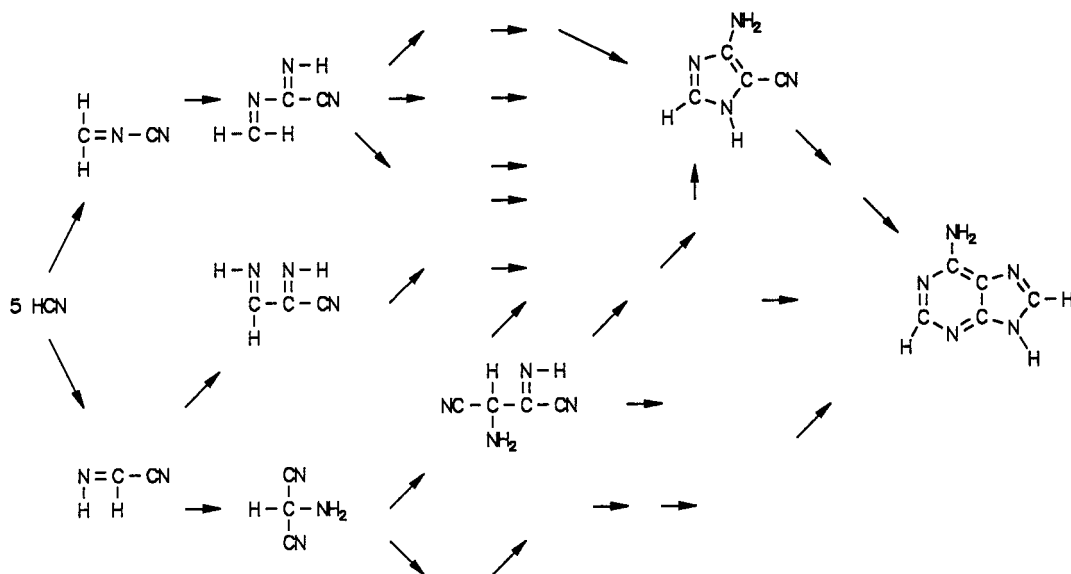
CONCLUSION

The development and acceptance of computer assistance in chemistry, and particularly in synthesis design have been slower than anticipated. Also, no dramatic progress is foreseeable for the immediate future. And yet, it will take no more than a few years until computer assistance in the design of syntheses and in the solution of other chemical problems will belong to chemists' everyday life. Very soon a great variety of computer programs for chemistry will become available, and a new generation of chemists will appear who have grown up with computers, like our generation did with bicycles.

Probably there will not be many batch programs that solve complex chemical problems fully automatically in some astounding way, but there will be many programs that will help with routine chores, and there will be many interactive programs performing the operations that are well automatable, and leave the decision that require the intelligence and imagination, knowledge and experience of an expert user.

In the field of synthesis design the reaction library based information oriented retrosynthetic systems have already demonstrated their capabilities and they will continue to have their share of the "market". The use of the logic oriented synthesis design systems, and related computer programs that do not primarily rely an detailed empirical data will soon increase significantly. It is conceivable, that in the long run, the logic oriented systems will change chemistry more profoundly, than the systems that rely on empirical data.

REFERENCES

1. M. Wochner, I. Ugi, Chem. Industrie, 1986.
2. N.S. Zefirov, E.V. Gordeva, Uspecki Khimii 56, 1753-1773 (1987); Russian Chem. Reviews 56, 1002-1014 (1987); see also:
3. J. Gasteiger, C. Jochum, Top Curr. Chem. 74, 93-126, (1978); J. Gasteiger, M.G. Hutchings, B. Christoph, L. Gann, C. Hiller, P. Löw, M. Marsili, H. Saller, K. Yuki, ibid. 137, 19-73 (1987).

4. Z. Hippe, D. Achmatowitz, R. Hippe Data processing in Chemistry, Ed.: Z. Hippe, Elsevier, Amsterdam 1986, p. 207-218.
5. V. Kvasnička, M. Kratochvil, J. Koča, Coll. Czech. Chem. Commun. 48, 2284-2303 (1983); J. Koča, M. Kratochvil, V. Kvasnička, L. Matyska, J. Pospichal, A Synthon Model of Organic Chemistry and Synthesis Design, Lecture Note Series, Springer, Heidelberg, in press.
6. J.B. Hendrickson, Top. Curr. Chem. 62, 49-172 (1976); J.B. Hendrickson, D.L. Grier, A.G. Toczko, J. Amer. Chem. Soc. 107, 5228-5238 (1985).
7. I. Ugi, J. Bauer, J. Brandt, J. Friedrich, J. Gasteiger, C. Jochum, W. Schubert, J. Dugundji, Computational Methods in Chemistry p.275-300, Ed.: J. Bargon, Plenum Press, New York, 1980.
8. I. Ugi, M. Wochner, J. Mol. Str., Theochem., 165, 229-242 (1988).
9. I. Ugi, J. Bauer, E. Fontain, Chem. Technik , (Leipzig), in press.
10. I. Ugi, J. Bauer, J. Brandt, J. Friedrich, J. Gasteiger, C. Jochum, W. Schubert Angew. Chem., 91, 99-111 (1979); Angew. Chem. Int. Ed Engl., 18, 111-123 (1979).
11. E. Hjelt, Geschichte der Organischen Chemie, Vieweg, Braunschweig 1916.
12. I. Ugi, Gießener Univ. Blätter 2, 69-79 (1978); Proc. Döbereiner Coll. Ed.: F. Bolck, Jena Univ. Press, Jena 1981, p. 57-69; Chem. Didactica, 8, 103-122 (1982).
13. J. Gasteiger, P.D. Gillespie, D. Marquarding, I. Ugi Topics Curr. Chem., 48, 1-37 (1974).
14. A. Cayley, Phil. Mag. 13, 19-27 (1857); 67, 444-458 (1974); Ber. dtsch chem. Ges. 8, 1056-1059 (1875).
15. A.T. Balaban, Chemical Applications of Graph Theory, Academic Press, London 1976.
16. V. Kvasnička, Coll. Czech. Chem. Comm. 48, 2097-2117, 2118-2129 (1983); 49, M. Kratochvil, J. Koča, V. Kvasnička, Chem. Listy, 79, 807-893, (1985); V. Kvasnička, J. Pospichal, ibid, in press.
17. G. Polyá, Acta Math. 68, 145-159 (1937).
18. E. Ruch, W. Hässelbarth, B. Richter, Theor. Chim. Acta 19, 288-300, (1970); W. Hässelbarth, E. Ruch, ibid. 29, 259-268 (1973); W. Hässelbarth, E. Ruch, D.J. Klein, T.H. Seligman, Group Theoretical Methods in Physics, Ed.: R.T. Sharp, B. Kolman, Academic Press, New York 1977, p. 617-638.
19. J. Dugundji, I. Ugi, Top Curr. Chem. 39, 19-64 (1973); I. Ugi, J. Ind. Chem. Soc. 62, 864-868 (1985); Proceedings of the 7. ICCCRE, Eds.: J. Brandt, I. Ugi, Hüthig, Stuttgart, in press.
20. H.L. Morgan, J. Chem. Doc. 5, 107-113 (1965).
21. W. Schubert, I. Ugi Chimia, 33, 1979, 183-191.
22. I. Ugi, J. Brandt, J. Friedrich. J. Gasteiger, C. Jochum, P. Lemmen, W. Schubert, Pure & Appl. Chem., 50, 1303-1318 (1978).
23. J. Bauer, E. Fontain, I. Ugi, Anal. Chim. Acta, in press.
24. C. Jochum, J. Gasteiger, I. Ugi, J. Dugundji, Z. Naturforsch., 37b, 1205-1215 (1982).
25. M. Wochner, I. Ugi, Chimia, 42, 217-225 (1988).
26. J.H. Winter, Chemische Syntheseplanung, Springer, Berlin, 1982.
27. E. Fontain, J. Bauer, I. Ugi, Chem. Letters, 1987, 37-40.
28. E. Fontain, J. Bauer, I. Ugi, Z. Naturforsch. 42 b, 889-891, 1987.
29. E. Fontain, J. Bauer, I. Ugi, PC for Chemists, Ed.: J. Zupan, Sigma Press, London, in press.
30. J. Friedrich, I. Ugi, J. Chem. Res.(S), 1980, 70; (M), 1980, 1301-1380; 1401-1497; 1501-1550.
31. D. Evans, Colloquium at UCLA, Los Angeles, CA, May 1971.
32. L. Spialter, J. Chem. Doc. 4, 261-268, 269-274 (1964).
33. I. Ugi, J. Dugundji, R. Kopp, D. Marquarding, Lecture Note Series, Vol. 36, Springer, Heidelberg, 1984, p. 1-265.
34. I. Ugi, Chimia 40, 340-350, (1986).
35. H. Kolbe, Liebigs Ann. Chem. 75, 211-235, (1850); 76, 1-19(1850).
36. J. Brandt, J. Bauer, R.M. Frank, A.v. Scholley, Chem. Scr. 18, 53-60 (1981); J. Brandt, A.v. Scholley, Comput. Chem. 7, 51-63 (1983).
37. J. Brandt, K. Stadler, A.v. Scholley-Pfab, H. Schönmann, M. Wochner, Proceeding of ACHEMA 1988, in press.

38. C.-C. Tsai, M. Johnson, V. Niocholson, M. Nain, <u>Graph Theory and Topology in Chemistry,</u> Vol. 51, Elsevier, Amsterdam 1987.
39. R. Ponec, <u>Z. Phys. Chemie,</u> 268, 1180-1188, (1987) and preceding papers.
40. J. Bauer, I. Ugi <u>J. of Chem. Res.</u> 1982, <u>(S)</u> 298; <u>(M)</u> 1982, 3101-3196; 3201-3260; see also J. Bauer, Doctoral Thesis, Techn. Univ. München, 1981.
41. J. Bauer, R. Herges, E. Fontain, I. Ugi, <u>Chimia 39,</u> 43-53 (1985).
42. E. Fontain, Doctoral Thesis, Techn. Univ. München, 1987.
43. J. Blair, J. Gasteiger, C. Gillespie, P.D. Gillespie, I. Ugi, <u>Computer Representation and Manipulation of Chemical Information,</u> Eds.: W.T. Wipke, S.R. Heller, R.J. Feldmann, E.Hyde, John Wiley, New York 1974, p. 129-145.
44. W. Schubert, <u>MATCH Informat. Commun. Math. Chem..</u>6, 213-218, (1979).
45. J. Brandt, K. Stadler, <u>Modern Approaches to Chemical Reaction Searching,</u> Gower, Aldershot 1986, p 93-112.
46. A.v. Scholley, Doctoral Thesis, Techn. Univ. München, 1981.
47. See also: A.T. Balaban, <u>Rev. Roum Chim. 12,</u> 875-898 (1967).
48. I. Ugi, J. Bauer, E. Fontain, J. Götz, G. Hering, P. Jacob, B. Landgraf, R. Karl, P. Lemmen, R. Schneiderwind-Stöcklein, R. Schwarz, P. Sluka, N. Balgobin, J. Chattopadhyaya, T. Pathak, X.-X. Zhou, <u>Chem. Scr. 26,</u> 205-215, (1986).
49. T.F. Brownscombe, Ph. D. Thesis, Rice University, Houston, TX, USA, 1972.
50. R. Herges, Doctoral Thesis, Techn. Univ. München, 1984.
51. F.W. Fowler, <u>Angew. Chem. 83,</u> 148-149 (1971); <u>Angew. Chem. Int. Ed. Engl.</u> 10, 135-136 (1971).
52. D. Forstmeyer, J. Bauer, E. Fontain, R. Herges, R. Herrmann, I. Ugi, <u>Angew. Chem.,</u> in press.
53. D. Forstmeyer, J. Bauer, E. Fontain, I. Ugi, <u>Tetrahedron Computer Methodology,</u> in press.
54. D. Forstmeyer, Doctoral Thesis, Techn. Univ. München, 1988.
55. A. Defoin, G. Augelman, H. Fritz, G. Geoffroy, C. Schmidlin, J. Streith, <u>Helv. Chim. Acta, 68,</u> 1998-2014 (1985).
56. J. Oró, A.P. Kimball, <u>Arch. Biochem. Biophys 94,</u> 217-224 (1961); J. Oró, S.S. Kamat, <u>Nature 190,</u> 442-443 (1961); see also: R.E. Dickerson, <u>Sci. American</u> 239; 70-109 (1978); <u>Spektrum.d. Wiss.</u> 1979, 98-115.