Chemical Similarity, Chemical Distance, and its Exact Determination

Micaela Wochner, Josef Brandt, Annette von Scholley, and Ivar Ugi*

Dedicated to Professor Ulrich Schöllkopf on the occasion of his 60th birthday

The principle of minimum chemical distance, a modern quantitative version of the classical principle of minimum structure change, can now be applied with precision. – In this article we describe for the first time the PEMCD, a computer program for the determination of the exact minima of chemical distance. The PEMCD solves the combinatorial n!-problem for the chemically relevant cases through an approach that is based on chemical and graph theoretical reasoning. The underlying theory and algorithms of PEMCD are discussed. PEMCD does not have the flaws of the previous PMCD-program that finds the minima of chemical distance through an algebraic approximation. In some cases the prerequisites of this approximation are not met, and then the PMCD-program may find false minima of chemical distance (i.e. local minima instead of global).

Micaela Wochner: Born 1958 in Nürnberg. She obtained Dipl.-Chem. 1982 at the Institute of Organic Chemistry, Technische Universität München, and also there the Dr. rer. nat. in 1985 with I. Ugi and J. Brandt, with a Thesis on «Die exakte Lösung der Zuordnungsaufgabe für chemische Reaktionen», whose results are an essential part of the present article (PEMCD). Since 1985 she does research in computer chemistry at the aforementioned institute.

Annette von Scholley-Pfab: Born 1953 in Stuttgart. Dipl.-Chem. 1978 at the Technische Universität München; Dr. rer. nat. in 1981 with I. Ugi and J. Brandt, Thesis on «Hierarchische Klassifizierung und Dokumentation chemischer Reaktionen». 1982/83 one year postdoctoral associate of M.F. Lynch, University of Sheffield (England); 1984/85 at the Beilstein-Institut, Frankfurt am Main (FRG).

Josef Brandt: Born 1935 in Stolberg (Rheinland). Dipl.-Chem. 1962 at the Technische Hochschule Aachen; graduated (Dr. rer. nat.) 1963 with a Thesis «Über das Assoziationsverhalten und bevorzugte Molekülformen bei gemischten Alkylalkanen». Since 1960 research work at the Max-Planck-Institut (MPI) für Kohlenforschung, Mülheim a.d. Ruhr, and from 1974 at the aforementioned institute of I. Ugi, where the habilitation took place in 1982 with a Thesis on «Ein mathematisch begründetes hierarchisches Ordnungssystem für chemische Reaktionen und dessen theoretische und praktische Anwendung».

Ivar Ugi: Biographical information is found in Chimia 39 (1985) 43 and 40 (1986) 340.

1. Introduction

Chemical distance (CD) is an elementary concept with almost ubiquitous implications and applications in chemistry. Although CD, as a notion, is independent of computer assistance in chemistry, its rife importance was primarily recognized through computer-assisted chemistry.

With CD we have for the first time a well-defined quantitative measure of chemical similarity, and the notion of CD is indispensable for the systematic classification and documentation of chemical reactions.

The intuitive principle of minimum structure change that has guided chemists since the middle of the nineteenth century can now be formulated with precision and in quantitative terms as the principle of minimal chemical distance (PMCD). The PMCD is a versatile and powerful device for chemical reasoning, including automated reasoning, and it is one of the basic rules of the game in logic-oriented computer assistance in chemistry.

The development of problem-solving computer programs started about twenty years ago, and until the last decade there was unlimited optimism about the prospects of computer assistance in chemistry. Some chemists even feared that computers will take over many of the intellectual activities in chemistry, and that computers will usurp some jobs of chemists, or, at least, deprive chemists of their prestige.

In the meantime the chemical community is somewhat disappointed with computer assistance in chemistry. There is some «Katzenjammer» among those who have expected miracles. We still do not have any batch programs that solve chemical problems in some astounding way, not even in a fully satisfactory manner, neither in the field of synthesis design, nor in automated structure elucidation from spectroscopic data, nor in other areas. It seems that no dramatic progress can be expected for the foreseeable future.

Nevertheless, there has been steady progress in the development of computer programs for chemistry, and in the near future a variety of interactive computer programs for the solution of complex chemical problems by small computers will be available.

These programs will be expert systems in the sense that their users must be experts. These programs will do the logic and combinatorial work in solving chemical problems, and they will also make decisions that can be reached by formal procedures, but all decisions that require chemical knowledge, experience, and intuition will be left to the expert user.

The usefulness of computer programs for the solution of chemical problems depends very much on effective selection procedures that are capable of picking in a non-arbitrary manner the few desirable solutions of a given problem from the immense number of conceivable solutions that a computer may generate. This holds particularly for the mathematically-based computer programs for chemistry[1]. Due to new ways of classifying the results, these logic-oriented computer programs are now maturing into widely usable rou-

* *Correspondence*: Prof. Dr. I. Ugi
Organisch-chemisches Institut
Technische Universität München
Lichtenbergstrasse 4, D-8046 Garching
(Bundesrepublik Deutschland)

tine tools of chemists[2]. The interactive operating mode with a clear division of duties between the computer and its user as here definite advantages. Thus the capabilities of man and machine are best exploited. Heuristic rules and selection procedures are avoided through the use of hierarchic classification and purely formalistic selection procedures[3–6]. Since there is no heuristic rule that is equally valid throughout all of chemistry, the fully automated application of heuristics to computer-assisted chemistry may often lead to arbitrary decisions.

The principle of minimum chemical distance[7–9] is the foundation of increasingly important formalized selection procedures. In logic-oriented computer chemistry the PMCD provides effective formal procedures for obtaining realistic results.

The PMCD, a computer-oriented version of the classical principle of minimum structure change, follows from the theory of the BE- and R-matrices[9–11]. This mathematical model of the logical structure of constitutional chemistry was published in 1973, and it is now the theoretical foundation of a great variety of computer programs for the deductive solution of chemical problems[1, 2].

The PMCD says that the result of a chemical reaction is generally achieved by the redistribution of a minimum number of valence electrons. A minimum number of covalent bonds are generally broken/made during a chemical reaction that follows the PMCD; a maximum set of maximum substructures is thus preserved[8].

The classical principle of minimum structure change during chemical reactions was enunciated by *Kolbe* in 1850[12]. This heuristic, intuitive, and somewhat vague principle has been modified and reformulated many times[13]. In the traditional educt/product oriented graph theoretical treatment of chemical reactions, the prinicple of minimum structure change has been stated as the principle of the maximum common subgraph in the educts and products of a chemical reaction[14] (see Scheme 1).

A chemical reaction, or sequence of reactions, is the conversion of an ensemble of molecules (EM) into an isomeric EM. The chemical distance (CD) between any two isomeric EM depends on the atom-by-atom correlation of the EM; the number of valence electrons that are redistributed during a chemical reaction is a function of the mechanism of the reaction[8] (see Schemes 3–7).

If the mapping of the EM is not specified, a redistribution of the minimum number of valence electrons is assumed for the interconversion of the EM, and it is used as the CD between these EM, i.e. their minimum CD (MCD) is then their CD.

The CD is a quantitative measure of similarity for isomeric molecules and ensembles of molecules. The concept of CD and the PMCD are useful, sometimes even indispensable, for many purposes in chemistry, e.g. the detailed description and systematic documentation of chemical reactions[15], the elucidation of reaction mechanisms and metabolic pathways[7], and the design of syntheses, be it by the retro-synthetic[16], or the bilateral approach[17].

In general, the application of the PMCD involves the solution of a formidable combinatorial problem. In order to find the atom-by-atom mapping of an EM with n atoms onto an isomeric EM under the condition of MCD, one must solve a linear assignment problem that corresponds to a combinatorial $n!$-problem.

In this article we describe for the first time a *p*rogram for the determination of the *e*xact *m*inimum of *c*hemical *d*istance, PEMCD. The PEMCD solves the MCD problem for the majority of the chemically relevant cases through an approach that is based on chemical and graph theoretical considerations.

A previous computer program for the approximation of the minimum of chemical distance, the PMCD-program[7, 8], relied on an algebraic approach, and the approximation standpoint that the linear assignment problem can be replaced by a more easily solved quadratic assignment problem. We developed the present PEMCD, since we noticed that the PMCD-program may not perform in a satisfactory manner, whenever the changes in bond order do not all have the same value, e.g. in the case of the Streith reaction[18, 19].

2. PMCD and the Classification and Representation of Chemical Reactions

In a reaction, or sequence of reactions along a pathway of MCD a minimum number of covalent bonds is broken/made and a minimum number of lone valence electrons changes its placement; we have no redundant bond breaking/making, i.e. a chemical bond is either broken, or made, but not broken, remade and broken again etc.[8].

In some cases (e.g. the Streith reaction[18, 19]), the energetically preferred reaction mechanism is «a little longer» than the MCD pathway.

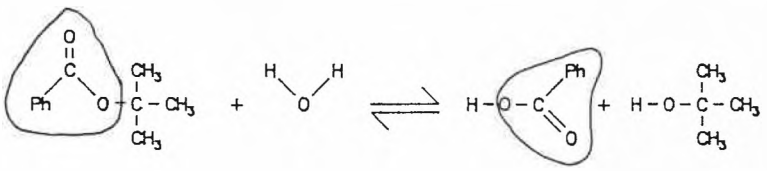
During a reaction, or sequence of reactions according to MCD *maximum sets of maximum common subgraphs* are preserved (e.g. in Scheme 2).

The maximum common subgraph versus the maximum set of maximum common subgraphs is illustrated by the hydrolysis of *tert*-butyl benzoate in Schemes 1 and 2.

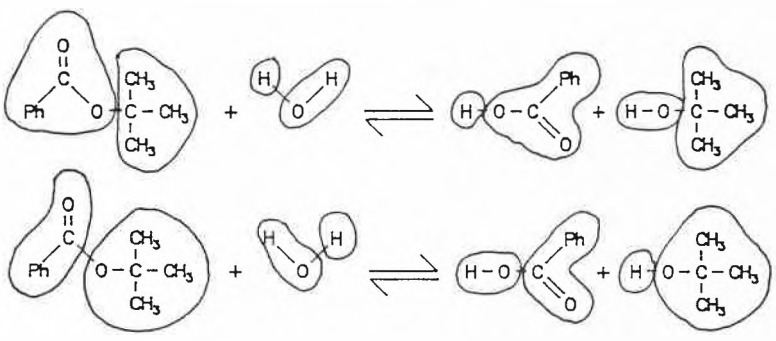
The example of the Streith reaction in Schemes 3–7*) demonstrates that in many cases the unique representation of chemical reactions is not trivial matter, and that atom-by-atom mappings of the educts onto the products are definitely needed for unambiguous representations of

*) In Schemes 3–7 symbols of H atoms are omitted for the sake of simplicity.

Scheme 1: Maximum common subgraph.



Scheme 2: Maximum set of maximum common subgraphs.



Ph = Phenyl-

chemical reactions. These Schemes 3–6 are the result of a «Summer Reaction Mechanism Contest» at the University of California, Los Angeles, in 1983 that was directed by *M.E. Jung* who also confronted us with this problem, inviting a computer-assisted solution.

The contestants had to propose a plausible mechanism for a reaction that had been published by *Streith* et al.[18] in 1982 (see below). Four of the proposed reaction mechanisms (Schemes 3–6) were accepted by the jury as chemically plausible.

The mechanism of the Streith reaction was elucidated by *Streith* et al.[19] in 1985. It corresponds to Scheme 3, the mechanistic proposal by *M.E. Jung*. This reaction mechanism is also contained in a network of reaction mechanisms (Scheme 7) that has been recently generated by the computer program RAIN[6].

The solutions of Scheme 4 and Scheme 6, as well as two further solutions with an MCD of 28 were found by PEMCD (see below) as the MCD reaction pathways for the Streith reaction. This demonstrates that for a complex chemical reaction more than one mechanistic pathway of MCD may be conceivable, and compatible with the experimental evidence that is available.

When the Streith reaction was subjected to the previous PMCD-program[7, 8], it was first rejected. After the phenyl group had been represented as hyperatom, it was, however, accepted. The PMCD-program then produced the solution of Scheme 5 with CD = 32. The reasons for the discrepancy between the results of the PEMCD and the PMCD-program are due to the approximation standpoint of the PMCD-program (see below).

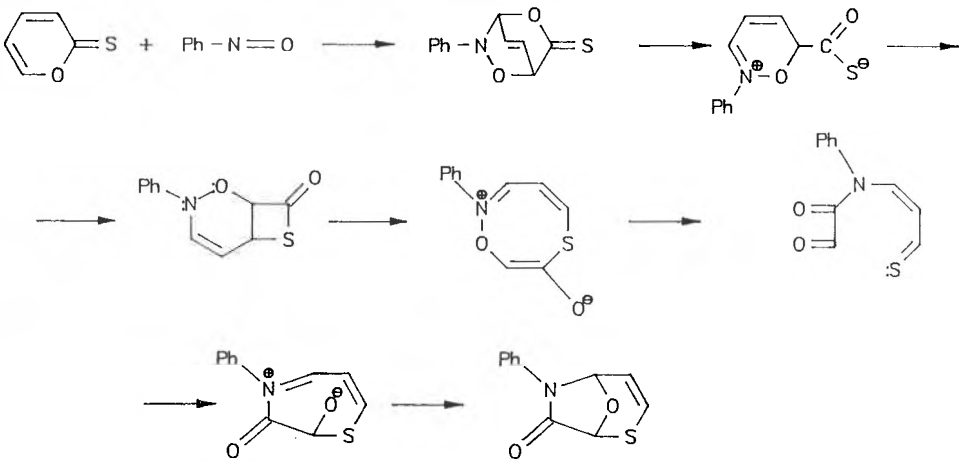
The proposal of Scheme 4 with CD = 28 is also compatible with recent experimental results of *Streith* et al.[19] yet it is less probable, because it contains an unprecedented type of hydrogen shift. A similar reaction of 3-methylthiopyrone by the mechanism of Scheme 4 would require an even less likely migration of a methyl group[19].

The preference of Scheme 3 in chemical reality shows that a reaction pathway with a CD one level above the MCD (for non-radical reactions the CD increases in increments of four units) may have energetic advantages over those with MCD.

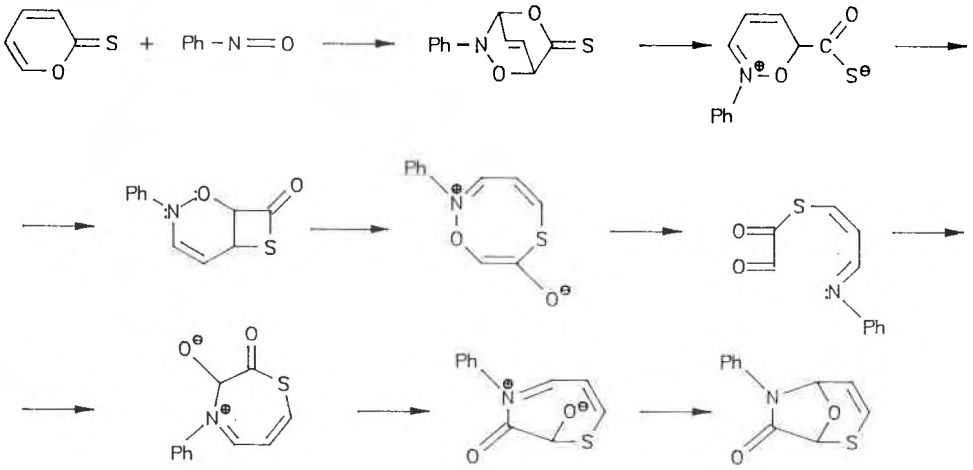
The fact that multistep reaction mechanisms and sequences of chemical reactions may sometimes slightly deviate from MCD, means that the PMCD is not strictly valid, and that when it is applied, not only the reaction pathway of MCD must be considered but also those at one, or even two levels above the MCD. The mechanistic reasons for the deviations of multistep processes from the MCD are mostly interesting.

Note that in none of the Schemes 3–6 the atom-by-atom correlations of the educts and the products are identical with any of the others, and that in each case different sets of bonds are broken/made. A systematic documentation of the Streith reaction beyond merely stating the educts and the products, thus requires a statement about the atom-by-atom correlation of the educts and the products that simultaneously indicates the bonds broken/made.

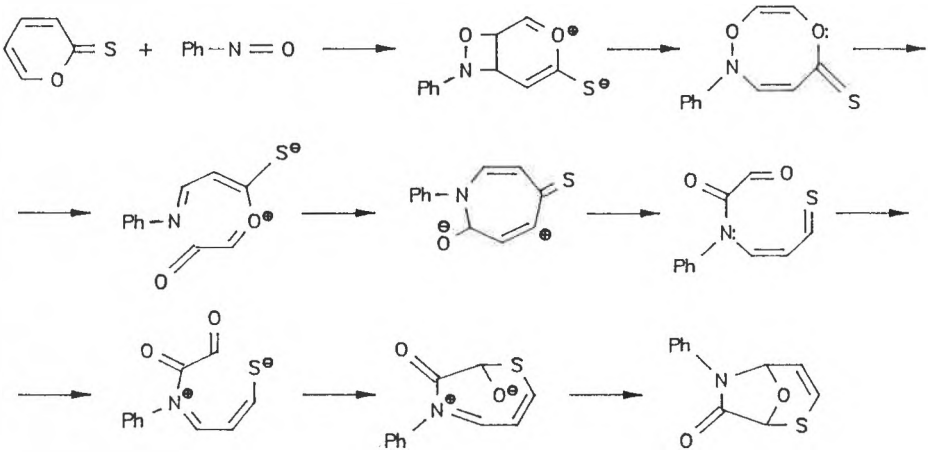
Scheme 3: Number of bonds changed: broken 8, made 8; CD = 32.



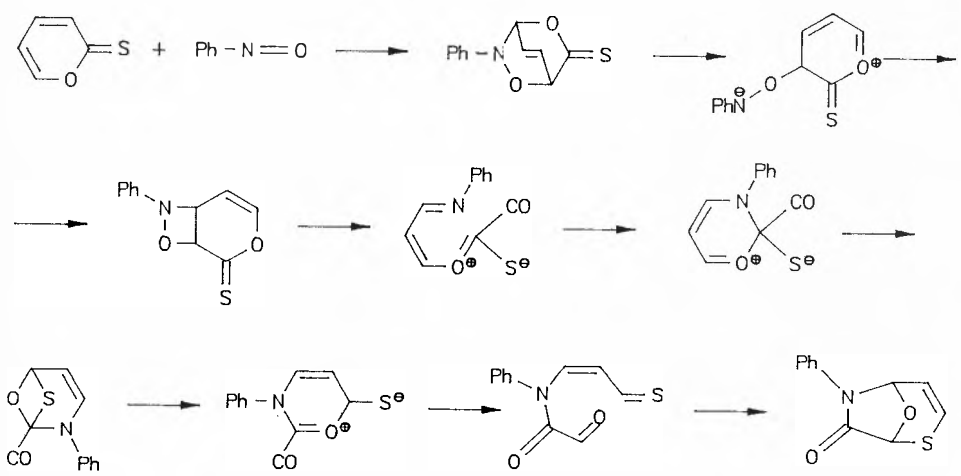
Scheme 4: Number of bonds changed: broken 7, made 7; CD = 28.



Scheme 5: Number of bonds changed: broken 8, made 8; CD = 32.



Scheme 6: Number of bonds changed: broken 7, made 7; CD = 28.



The PMCD is also useful for detecting similarities of molecular systems[23], because it enables us to find the maximum set of maximum substructures that is common to any two comparable EM.

The common biochemical precursors of related natural products could, for instance, be thus detected by PMCD (see below).

Furthermore, the PMCD is helpful in the study of reaction mechanisms, since it provides a complete numerical labeling of the atoms in the educts and the products of chemical reactions.

Isotope labeling experiments yield only a partial correlation of atoms in the educts and products of chemical reactions or their sequences. This atom-by-atom mapping also indicates the bonds that must be broken and made in order to achieve the given chemical result (see Schemes 3–6).

Scheme 8 illustrates the application of PEMCD to a biosynthetic route that has been postulated by *Battersby* and *Parry*[24]. Under the assumption that in biochemistry the economy of bond breaking/making is also generally prevalent, the plausibility of the aforementioned biosynthesis is supported by the fact that it is a pathway of MCD.

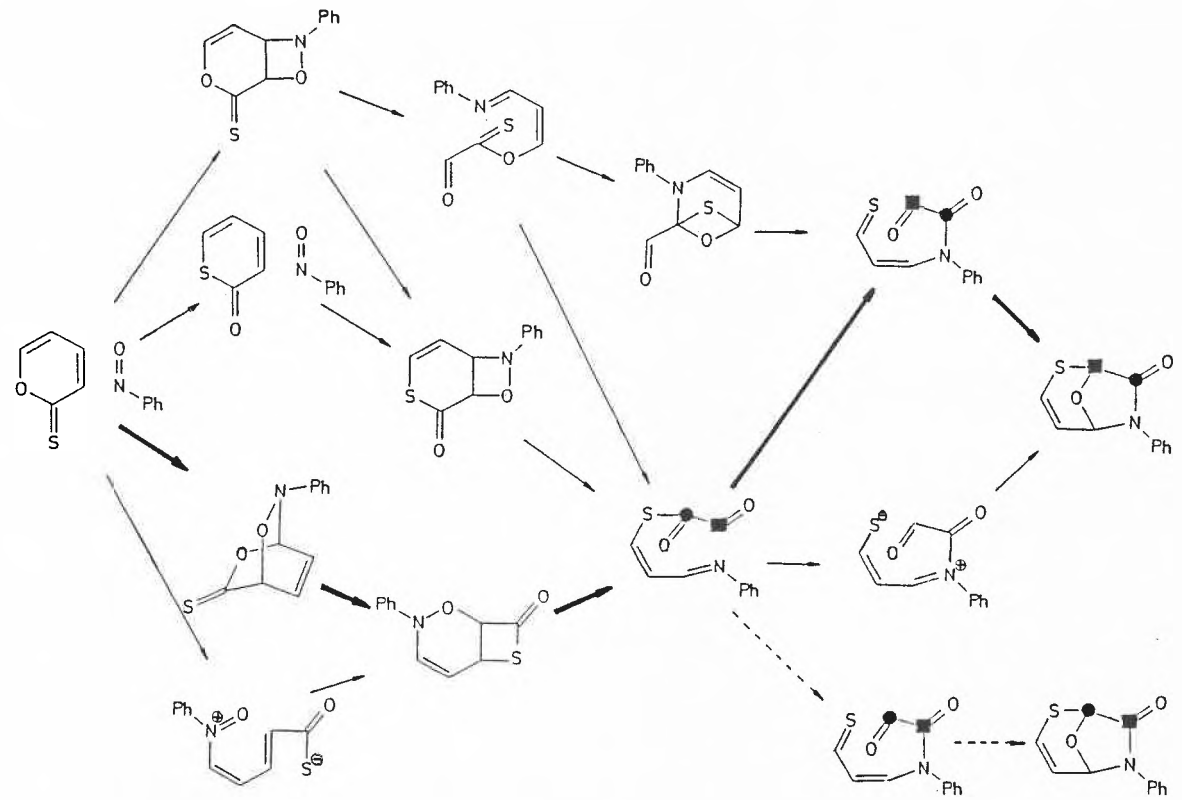
Since the PMCD can also be applied to sequences of chemical reactions, it is useful as a guiding principle for computer-assisted design of syntheses[5,6,25], and the

3. Applications of the PMCD

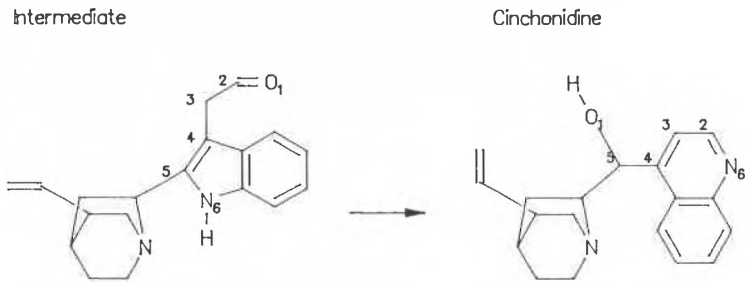
The PMCD has many applications in computer-assisted chemistry. The PMCD, or an equivalent device is indispensable in the systematic classification and documentation of chemical reactions[1,5]. Not only the PMCD, but also a hierarchic classification of chemical reactions follows immediately from the theory of the BE- and R-matrices[2].

From 1977 on a documentation system for chemical reactions[1,5] that is founded on the PMCD and this hierarchical classification has been developed. Some other recently proposed graph theory based documentation systems of chemical reactions[20–22] are similar. Their underlying graph-theoretical approaches to chemical reactions are isomorphic to parts of the algebraic theory of the BE- and R-matrices. However, the latter systems are incomplete, since they do not use the PMCD or any comparable device for atom-by-atom correlation of the educts and the products, and the detection of the reactive sites in order to ensure uniqueness in representation. For computer use the graph-theoretical representations of reactions must be converted into algebraic terms.

Scheme 7: Network of reaction mechanisms generated by RAIN[6].



Scheme 8: Biosynthesis of chinchonidine.



elucidation of metabolic pathways. When trees of reaction pathways are generated, e.g. in retro-synthesis, «curling» pathways can be avoided by the use of the PMCD. When reaction networks are bilaterally generated, the application of the PMCD ensures directed growth of the pathways from both ends, and joins the tree from the educts with the tree from the products in an optimum fashion[2].

4. Theoretical Foundations of the PMCD

The PMCD belongs to the theory of the BE- and R-matrices[9–11] whose essential features are discussed in this section.

The chemical constitution of molecules or ensembles of molecules with n atoms is represented by their $n \times n$ symmetric BE-matrices (*b*ond and *e*lectron matrices) B and E. The conversion of an educt-EM(**B**) with n atoms into an isomeric product-EM(**E**) by a chemical reaction is represented by the matrix equation

$$B + R = E$$

were R is an $n \times n$ reaction matrix that describes the redistribution of the valence electrons during the process EM(**B**) $\mapsto$ EM(**E**).

The rows/columns of B and E are assigned to the atomic cores belonging to EM(**B**) and EM(**E**), the *b*eginning and the *e*nd of the reaction. The positive integer entries b_{ij} and e_{ij} of B and E indicate the covalent bond orders between the atoms A_i, A_j and the redistribution of the lone valence electrons. The reaction matrix (R-matrix) R is a symmetric matrix with positive and negative integer entries r_{ij} that indicate the changes of the formal covalent bond orders and the redistribution of lone electrons. We have

$$\sum_{ij} r_{ii} = 0$$

but

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

The function $d(\mathbf{B},\mathbf{E})$ is called the chemical distance between EM(**B**) and EM(**E**)[8,9].

In an EM with n atoms the atoms can be indexed in up to $n!$ distinct ways, and thus EM(**B**) and EM(**E**) can be represented by up to $n!$ distinct BE-matrices $B' = P \cdot B \cdot P^{-1}$ and $E' = P \cdot E \cdot P^{-1}$ with permuted rows/columns; this corresponds to a permuted assignment of the atoms of the EM to the rows/columns of the BE-matrices. P is an $n \times n$ permutation matrix. The CD is a function

$$F(\mathbf{P}) = d(\mathbf{B}, \mathbf{P} \cdot \mathbf{E} \cdot \mathbf{P}^{-1}) = \sum_{ij} |b_{ij} - e_{ij,p}|$$

of the atom-by-atom correlation of EM(**B**) and EM(**E**). The indexed atoms in EM(**B**) and EM(**E**) may be mapped onto each other in up to $(n!)^2$ different ways. Without guiding concepts, the trial-and-error search for a correlation of atoms that corresponds to the minimum of $d(\mathbf{B},\mathbf{E})$ would require the determination of up to $(n!)^2$ atom-by-atom bijections of EM(**B**) onto EM(**E**). In a preceding paper[8] proof was given that $n!$ or fewer atom-by-atom bijections suffice to find by trial and error an atom-by-atom bijection with MCD.

In a mathematical sense, the CD is a genuine distance, and it has geometric meaning: An $n \times n$ BE-matrix is representable by a BE-point P(**B**) in a n^2-dimensional euclidean space $\mathbb{R}^{n^2}$. With a given assignment of the atoms in EM(**B**) and EM(**E**) to the rows/columns of their BE-matrices B and E, $d(\mathbf{B},\mathbf{E})$ is precisely the L_1-distance («city block distance») between the points P(**B**) and P(**E**) that represent EM(**B**) and EM(**E**).

An EM(**B**) corresponds to a cluster of up to $n!$ BE-points $P(\mathbf{B}') = P(\mathbf{P} \cdot \mathbf{B} \cdot \mathbf{P}^{-1})$. The structure of these clusters is such that for each P(**B**) there exists a point in the cluster of the P(**E**) whose CD from P(**B**) is the MCD. The clusters of BE-points behave like parallel planes in which each point of a plane has a closest point on a parallel plane.

5. The Approximate Determination of MCD by Quadratic Assignment

An exhaustive calculation of the up to $n!$ distinct mappings, EM(**B**) $\mapsto$ EM(**E**), and the selection of those with MCD, i.e. determination of isomorphic subgraphs, is, in general, not feasible.

In contrast to graph isomorphism, subgraph isomorphism has not yet been explored to any appreciable extent[26,27], due to the *np*-complete[28] nature of the problem.

In order to solve this *np*-complete problem the so-called PMCD-program[7,8] (program for *m*inimal *c*hemical *d*istance) was developed. The PMCD-program is based on the approximation standpoint that the minimum of $d(\mathbf{B},\mathbf{E})$ corresponds to the minimum of $D(\mathbf{B},\mathbf{E})$, the euclidean L_2-distance, and to the maximum value of $B \times E$, the inner product of the BE-matrices[8]:

$$\min d(\mathbf{B},\mathbf{E}) = \min \sum_{ij} |r_{ij}| \equiv \min \sum_{ij} (r_{ij})^2$$

$$= \min D(\mathbf{B},\mathbf{E})$$

This is true, if all non-zero r_{ij} have the same absolute value $|r_{ij}|$, but not otherwise. There exists, however, a wide variety of reactions whose reaction matrices require non-zero entries $r_{ij} = 1, 2, \ldots$ (see Schemes 3–6). In such cases the minima of CD and euclidean distance do not coincide.

In the PMCD-program the linear assignment problem is replaced by a quadratic assignment problem and the minimum of euclidean distance is determined with the aid of the well-known branch-and-bond algorithms of *Burckard* et al.[29].

6. The Exact Solution of the Minimum Chemical Distance Problem

Our new PEMCD-algorithm relies on a combination of chemical and mathematical considerations. The problem is reduced, and partitioned into smaller subproblems according to chemical and graph-theoretical considerations. These subproblems are solved by exhaustive permutation of the local atom-by-atom mappings; the solutions that comply with the PMCD are selected.

Some solutions that are close to the MCD may also be of interest (see Scheme 7). The PEMCD affords generally the exact solutions to the MCD problem and the atom-by-atom bijection problem for chemical reactions, and sequences thereof.

First, the educts EM(**B**) and the products EM(**E**) are subjected to a preliminary search for constitutionally equivalent atoms[30–32] (a). This procedure provides some prima vista – possibly incomplete – information on the reactive sites that form the core of the reaction[15] and the invariant parts of EM(**B**) and EM(**E**).

With this information, the terminal subsets of connected reaction-invariant

atoms are determined, and then treated as hyperatoms (b).

Subsequently the *m* terminal univalent (hyper-)atoms of the reacting EM are represented as distinct *m*-valent subunits, the so-called residues (c).

These considerations that are based on chemical structure lead to a partitioning of the atoms, hyperatoms, and residues of the reacting EM into intra- and intermolecular equivalence classes.

If any equivalence classes of atoms in EM(**B**) and EM(**E**) have the same cardinality, they are mapped onto each other and subjected to exhaustive permutation (d) in order to find all of the conceiveable chemically meaningful mappings. At the same time, we establish the atom-by-atom mappings of the remaining atoms in EM(**B**) and EM(**E**), with due consideration of their atomic numbers. The combinations of these mappings yield the complete set of all chemically plausible solutions to the atom-by-atom mapping problem. Their CD is determined in order to detect the solutions of MCD.

In this section the essential algorithms of PEMCD are described in greater detail. The hypothetical disproportionation of dimethyl phosphite (**1**) into trimethyl phosphite (**2**) and monomethyl phosphite (**3**), or the reverse reaction (Scheme 9) is used as an example in the explanation of the algorithms.

7. Essential Features of the PEMCD-Algorithms

(a) Recognition of the Equivalence Classes of Atoms

In order to detect inter- and intramolecular constitutional equivalences of atoms, and the local isomorphisms of subgraphs in the formulas of EM(**B**) and EM(**E**), the union of EM(**E**) and EM(**B**) is subjected to a so-called relaxation algorithm, a modified CANON-algorithm[30–32]. Note that in contrast to CANON that operates on individual molecules, all atoms in the union of EM(**B**) and EM(**E**) are included in the treatment by this modification of CANON.

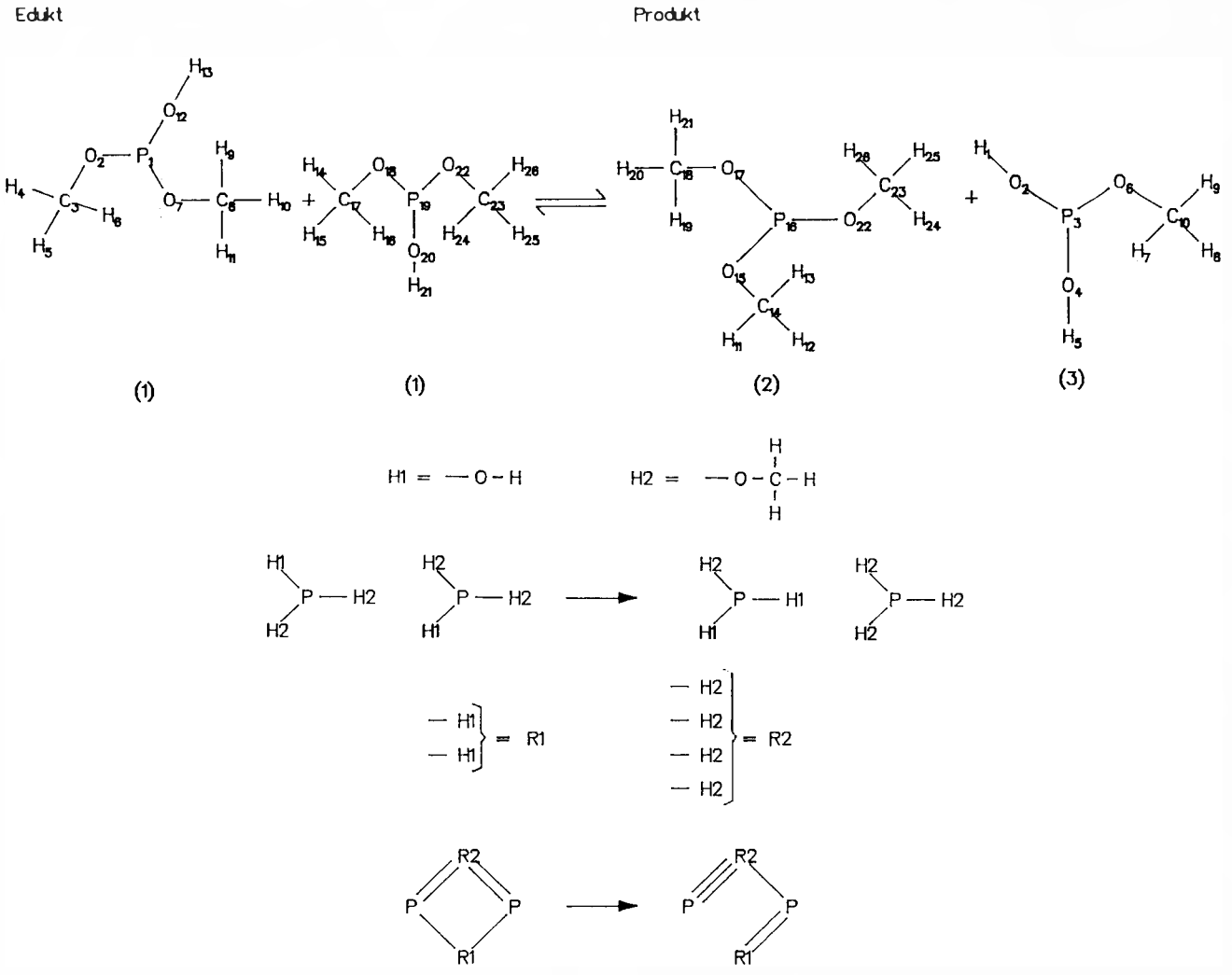
CANON takes into account simultaneously and with equal emphasis the chemical nature of the atoms and the graph of the molecule. CANON recognizes any constitutional symmetries and equivalent atoms that are present. In contrast to the original algorithm CANON, the formal bond orders of the covalent bonds of the atoms are also considered by the modified algorithm. The bonds are important for characterizing the effect of the reaction on the individual atoms.

Moreover, the termination criterion of CANON must be changed for solving the subgraph isomorphism problem. The computer-assisted detection of isomorphic subgraphs in chemical formulas has been studied in a different context by *Sussenguth*[33] and *Figueras*[34]. They compared local graph-theoretical properties in chemical graphs by means of an iterative Morgan-type algorithm[35]. They used a termination criterion that enables the algorithm to detect local equivalencies of nodes, and whether or not they are entirely equivalent within the whole graph.

First, all atoms in EM(**B**) ∪ EM(**E**) are characterized by a permanent identifying label (IL) and an atomic index (AI) that changes under iteration. The identifying labels are assigned *ad libitum* and are kept throughout the algorithm. The AI of an atom is derived from its atomic number (AN); the highest AN corresponds to AI = 1, and the AI increases with decreasing AN, in analogy to the CIP-priorities[36].

The primary atomic descriptor (PAD) of an atom is a vector that consists of its AI, followed by the AI of its covalently bound immediate neighbors in their natu-

*Scheme 9: Hypothetical disproportionation of dimethyl phosphite (**1**) into trimethyl phosphite (**2**) and monomethyl phosphite (**3**).*



ral order, then followed by the formal orders of their covalent bonds as is shown in Tables 1B and 1E. In analogy to the algorithm CANON, the atoms are ordered and reindexed by 2.AI according to their PAD.

Whenever the relaxation algorithm[37] generates an atomic descriptor (AD) for a member of EM(**B**) (or (EM(**E**)) without a corresponding AD for any member of EM(**E**) (or EM(**B**)) this member of EM(**B**) (or EM(**E**)) belongs to the core of the reaction, and not to an invariant part of the EM. The atoms in the core of the reaction are given an AI of «−1», as soon as they are recognized at any stage of the iteration, enabling the above algorithm to detect those parts of EM(**B**) and EM(**E**) that do not participate directly in the reaction, i.e. the so-called invariant parts of the reactants.

Then the 2.AI are used to manufacture 2.AD. The latter step is repeated in order to obtain the 3.AI and the 3.AD etc. etc. until the termination criterion of the algorithm is fulfilled (see below).

The relaxation algorithm is terminated when no new AD are generated by further iteration. This criterion ensures detection of locally isomorphic subgraphs in the formulas of EM(**B**) and EM(**E**)[33, 34].

In general the relaxation algorithm produces a preliminary atom-by-atom assignment of the EM, and a preliminary identification of the reactive core and the invariant part of the reactants. Within the subgraphs of the intervariant parts in EM(**B**) and EM(**E**), it detects the corresponding isomorphic counterparts.

(b) and (c) Hyperatoms and Residues

Hyperatoms are univalent polyatomic groups at the periphery of the molecules. They are not directly affected by the reaction (Scheme 9). The hyperatoms correspond to equivalence classes of atoms

Table 1B: The IL, AI, and AD of EM(**B**) = {1,1}.

	PAD			2.AD		3.AD		4.AD		
IL	1.AI	Neighbor atoms P O C H	Bond-order	2.AI	Neighbor atoms 2.AI	3.AI	Neighbor atoms 3.AI	4.AI	Neighbor atoms 4.AI	5.AI
1	1	o 3 o o	1 1 1	1	2 3 3	2(=-1)				
2	2	1 o 1 o	1 1	3	1 4	5	-1 6	2	-1 3	2
3	3	o 1 o 3	1 1 1 1	4	3 6 6 6	6	5 8 8 8	3	2 5 5 5	3
4	4	o o 1 o	1	6	4	8	6	5	3	5
5	4	o o 1 o	1	6	4	8	6	5	3	5
6	4	o o 1 o	1	6	4	8	6	5	3	5
7	2	1 o 1 o	1 1	3	1 4	5	-1 6	2	-1 3	2
8	3	o 1 o 3	1 1 1 1	4	3 6 6 6	6	5 8 8 8	3	2 5 5 5	3
9	4	o o 1 o	1	6	4	8	6	5	3	5
1o	4	o o 1 o	1	6	4	8	6	5	3	5
11	4	o o 1 o	1	6	4	8	6	5	3	5
12	2	1 1 o o	1 1	2	1 5	4	-1 7	1	-1 4	1
13	4	o 1 o o	1	5	2	7	4	4	1	4
14	4	o o 1 o	1	6	4	8	6	5	3	5
15	4	o o 1 o	1	6	4	8	6	5	3	5
16	4	o o 1 o	1	6	4	8	6	5	3	5
17	3	o 1 o 3	1 1 1 1	4	3 6 6 6	6	5 8 8 8	3	2 5 5 5	3
18	2	1 o 1 o	1 1	3	1 4	5	-1 6	2	-1 3	2
19	1	o 3 o o	1 1 1	1	2 3 3	2(=-1)				
2o	2	1 1 o o	1 1	2	1 5	4	-1 7	1	-1 4	1
21	4	o 1 o o	1	5	2	7	4	4	1	4
22	2	1 o 1 o	1 1	3	1 4	5	-1 6	2	-1 3	2
23	3	o 1 o 3	1 1 1 1	4	3 6 6 6	6	5 8 8 8	3	2 5 5 5	3
24	4	o o 1 o	1	6	4	8	6	5	3	5
25	4	o o 1 o	1	6	4	8	6	5	3	5
26	4	o o 1 o	1	6	4	8	6	5	3	5

Table 1E: The IL, AI, and AD of EM(**E**) = {2,3}.

	PAD			2.AD		3.AD		4.AD		
IL	1.AI	Neighbor atoms P O C H	Bond-order	2.AI	Neighbor atoms 2.AI	3.AI	Neighbor atoms 3.AI	4.AI	Neighbor atoms 4.AI	5.AI
1	4	o 1 o o	1	5	2	7	4	4	1	4
2	2	1 1 o o	1 1	2	1 5	4	-1 7	1	-1 4	1
3	1	o 3 o o	1 1 1	1	2 2 3	1(=-1)				
4	2	1 1 o o	1 1	2	1 5	4	-1 7	1	-1 4	1
5	4	o 1 o o	1	5	2	7	4	4	1	4
6	2	1 o 1 o	1 1	3	1 4	5	-1 6	2	-1 3	2
7	4	o o 1 o	1	6	4	8	6	5	3	5
8	4	o o 1 o	1	6	4	8	6	5	3	5
9	4	o o 1 o	1	6	4	8	6	5	3	5
1o	3	o 1 o 3	1 1 1 1	4	3 6 6 6	6	5 8 8 8	3	2 5 5 5	3
11	4	o o 1 o	1	6	4	8	6	5	3	5
12	4	o o 1 o	1	6	4	8	6	5	3	5
13	4	o o 1 o	1	6	4	8	6	5	3	5
14	3	o 1 o 3	1 1 1 1	4	3 6 6 6	6	5 8 8 8	3	2 5 5 5	3
15	2	1 o 1 o	1 1	3	1 4	5	-1 6	2	-1 3	2
16	1	o 3 o o	1 1 1	1	3 3 3	3(=-1)				
17	2	1 o 1 o	1 1	3	1 4	5	-1 6	2	-1 3	2
18	3	o 1 o 3	1 1 1 1	4	3 6 6 6	6	5 8 8 8	3	2 5 5 5	3
19	4	o o 1 o	1	6	4	8	6	5	3	5
2o	4	o o 1 o	1	6	4	8	6	5	3	5
21	4	o o 1 o	1	6	4	8	6	5	3	5
22	2	1 o 1 o	1 1	3	1 4	5	-1 6	2	-1 3	2
23	3	o 1 o 3	1 1 1 1	4	3 6 6 6	6	5 8 8 8	3	2 5 5 5	3
24	4	o o 1 o	1	6	4	8	6	5	3	5
25	4	o o 1 o	1	6	4	8	6	5	3	5
26	4	o o 1 o	1	6	4	8	6	5	3	5

that occur with the same cardinality in the educts and products. Within the molecular graph the hyperatoms are detected by following the pathways that begin at the invariant terminal atoms.

The dissection of molecules into hyperatoms and residues leads to a further reduction of the problem. When we have m univalent atoms or hyperatoms that do not participate in the reaction, these are represented as a single m-valent unit (Scheme 9). In chemistry this approach has been used since 1878[38]; the widely practiced neglect of hydrogen atoms in the representation of molecules by «short-hand» formulas is, in essence, such a treatment. It should be noted that the relations of the atoms belonging to the residues and their intramolecular vicinities are not lost in this approach, if the intramolecular and the intermolecular constitutional equivalencies of the atoms outside the hyperatoms and the residues are registered.

(d) Exhaustive Permutation of Atoms within Substructures

The aforementioned algorithms leave two types of uncertainties unresolved: In some cases no clear separation of the reaction core and the invariant part of the reactants is accomplished; some atoms may remain unclassified. In other cases the relaxation algorithm detects isomorphisms of more than two subgraphs, without being able to determine the correct intermolecular assignment of subgraphs. These uncertainties are removed by exhaustive permutation of the IL of the atoms in certain subsets of EM(**E**) and comparison of the respective chemical distances.

The aforementioned subsets are analyzed as follows: Hyperatoms and residues in EM(**B**) and EM(**E**) are directly mapped onto each other. The remaining atoms are screened for invariancy of the final AI under the reaction. What is left, are the atoms that belong either to the reactive core, or to the «uncertainties» of the invariant part. They are classified according to their atomic numbers.

8. Conclusion and Outlook

About 20 years ago the development of problem-solving computer programs began with the early information oriented synthesis design programs[16], and the essentially graph theory based DENDRAL project[39].

Since then we have two approaches to computer assistance in chemistry, one relying mainly on empirical information and heuristics, the other primarily based on formal logic. The two distinct philosophies still determine the development of two distinct categories of computer programs for chemistry.

It took almost 15 years until the mathematically based chemical computer programs have become generally usable tools of chemists. The new generation of interactive computer programs for the solution of chemical problems with the aid of small computers rests on a few pillars like the PEMCD. A fixed set of a few modules is used for the execution of tasks that occur again and again in the deductive solution of a variety of chemical problems.

These closely interdependent units are:

- CANON[30–32] that is an algorithm and a computer program for the detection of constitutionally equivalent atoms and the unique indexing of the atoms in a molecule, with due consideration of constitutional symmetries. CANON operates on the first sphere of covalently bound neighbors of the atoms in a molecule, the α-atoms. CANON and its modifications are used in almost all of our programs.
- Various modifications of the program CORREL[40] serve for substructure analysis and correlation. CORREL is based on a hierarchic network of all substructures that are considered. This approach avoids *np*-completeness of substructure analysis that is otherwise encountered. Hierarchic networks are now used in many computer programs that correlate substructures, e.g. HTSS[41], KOWIST[42], RESY[43], and the structure-activity program of *Klopman*[44].

Presently, we develop a new version of CORREL that will operate with the substructures that comprize the α, β, γ-spheres of covalent neighbors of the individual atom. The systems DARC[45] and HTSS work with two-sphere vicinities of covalent α,β-neighbors. For substructure retrieval two spheres suffice, whereas the three spheres approach has advantages for substructure correlation.

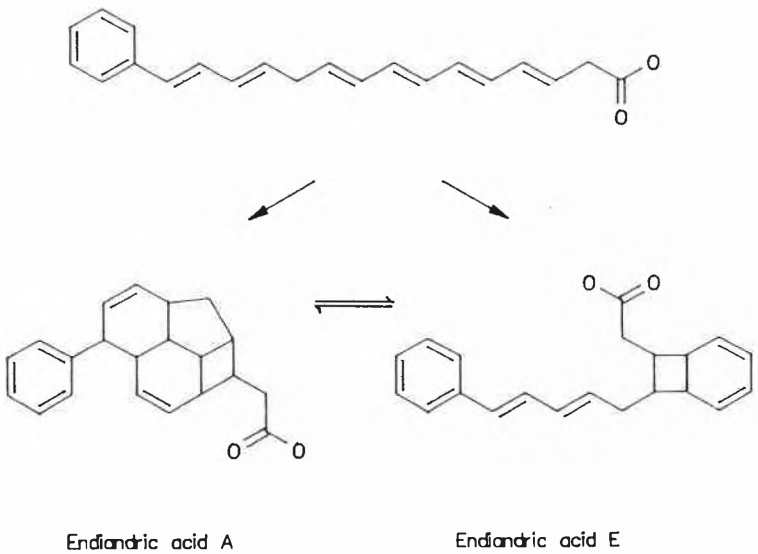
Substructure correlation is done in its own right, e.g. for structure-activity relations, but it is also used in the bilateral design of syntheses[17], in order to find the suitable starting materials for a given target, and for the suppression of forbidden substructures that are otherwise generated by IGOR[3, 4] and RAIN[5, 6].

- Hierarchic classification of chemical reactions is used within IGOR, and in the systematic documentation of chemical reactions[15].
- The *t*ransition table guided *r*eaction *g*enerators TRG I and TRG II[2] generate chemical reactions from a given BE-matrix, or a given R-matrix. TRG I is an essential component of RAIN, while TRG II is the backbone of IGOR.
- PEMCD correlates the atoms in the educts and products of chemical reactions under the aspect of MCD. PEMCD can be used in its own right for many purposes (see above), but PEMCD is also an indispensable component of the hierarchic documentation system for chemical reactions, and it will play an important role in the bilateral design of syntheses, and in RAIN.

The correlation of substructures and the determination of the minima of CD are the most difficult problems among those to be solved by the aforementioned modular subunits of logic-oriented chemical computer programs.

When we studied the hypothetical interconversion of endiandric acids[46] (Scheme 10) with PEMCD in order to find the common biosynthetic precursor, we noticed that PEMCD is unable to process very large contiguous reaction cores. This means that we must improve PEMCD: Our universal substructure module[47] will be used to compare the educt and product EM in order to detect and account for subgraph isomorphisms. In the non-isomorphic subunits the reactive centers are recognized by a modified version of CANON. The bonds that belong to the reactive centers are eliminated; thus the graphs of the educt and

Scheme 10: Synthesis and hypothetical interconversion of endiandric acids.



product EM become identical. Now the equivalent nodes in the latter common graphs of the educts and products are determined by CANON. Since these now equivalent nodes were not all equivalent in the original graphs of the educts and products, some permutations of the corresponding atomic indices must be analyzed under the aspect of MCD.

Just like recognition of the inability of the previous PMCD-program to deal adequately with the Streith reaction[18,19] has led us into PEMCD, the above study of the endiandric acids cascade will lead to improved versions of CORREL and PEMCD.

Chemical distance is a well-defined and simple concept; it has already become a household word. It implies rightfully the very esthetic metric topological nature of chemistry and relations within chemistry. «Chemical distance» was not a popular chemical notation much earlier, because the quantitative determination and practical use of chemical distance, and thus PMCD, requires sophisticated algorithms and complex computer programs whose development will still stay open-ended for a long time to come.

Acknowledgement: We acknowledge gratefully the financial support of our work by the Office of the European Community, Deutsche Forschungsgemeinschaft, and Fonds der Chemischen Industrie. We are grateful to Prof. M.E. Jung for pointing out the example of the Streith reaction, to Prof. C. Collins and Prof. C. McKenna for valuable suggestions, and to Ms. R. Karl for drawing the chemical formulas with the molecular graphics program GRECO that has been developed at our institute by E. Fontain and J. Bauer.

Received: September 30, 1987 [FR 55]

[1] I. Ugi, J. Bauer, J. Brandt, J. Friedrich, J. Gasteiger, C. Jochum, W. Schubert, *Angew. Chem. 91* (1979) 99; *Angew. Chem. Int. Ed. Engl. 18* (1979) 111.
[2] J. Bauer, E. Fontain, I. Ugi, Proceedings of the 8th ICCCRE, *Anal. Chim. Acta*, in press; I. Ugi, M. Wochner, *J. Mol. Struct. (Theochem) 165* (1988) 229.
[3] J. Bauer, I. Ugi, *J. Chem. Res. (S)* (1982) 298; *J. Chem. Res. (M)* (1982) 3101, 3201.
[4] J. Bauer, R. Herges, E. Fontain, I. Ugi, *Chimia 39* (1985) 43.
[5] E. Fontain, J. Bauer, I. Ugi, *Chem. Lett.* (1987) 37.
[6] E. Fontain, J. Bauer, I. Ugi, *Z. Naturforsch. B42* (1987) 889.
[7] C. Jochum, J. Gasteiger, I. Ugi, *Angew. Chem. 92* (1980) 503; *Angew. Chem. Int. Ed. Engl. 19* (1980) 495.
[8] C. Jochum, J. Gasteiger, I. Ugi, J. Dugundji, *Z. Naturforsch. B37* (1982) 1205.
[9] J. Dugundji, I. Ugi, *Top. Curr. Chem. 39* (1973) 19.
[10] I. Ugi, in J. Brandt, I. Ugi (Ed.): *Proceedings of the 7th ICCCRE*, Hüthig, Heidelberg, in press.
[11] V. Kvasnicka, M. Kratochvil, J. Koca, *Collect. Czech. Chem. Commun. 48* (1983) 2284; V. Kvasnicka, *ibid. 48* (1983) 2097, 2118; *49* (1984) 1090.
[12] H. Kolbe, *Liebigs Ann. Chem. 75* (1850) 211; *76* (1850) 1.
[13] W. Hückel: *Theoretische Grundlagen der Organischen Chemie I*, 8. Aufl., Akademische Verlagsgesellschaft, Leipzig (1956); J. Hine, *Adv. Phys. Org. Chem. 15* (1977) 1; F.O. Rice, E. Teller, *J. Chem. Phys. 6* (1938) 489; K.E. Schuler, *J. Chem. Phys. 21* (1953) 624; R.B. Woodward, R. Hoffmann, *Angew. Chem. 81* (1969) 797; *Angew. Chem. Int. Ed. Engl. 8* (1969) 781.
[14] M.F. Lynch, P. Willett, *J. Chem. Inf. Comput. Sci. 18* (1978) 154; P. Willett, Ph. D. Thesis, University of Sheffield (1978); D. Bawden, T.K. Devon, F.T. Jackson, S.I. Wood, M.F. Lynch, P. Willett, *J. Chem. Inf. Comput. Sci.19* (1979) 90; P. Willett, *ibid. 20* (1980) 93; J.J. McGregor, P. Willett, *ibid. 21* (1981) 139; C. Marshall, Ph. D. Thesis, University of Leeds (1984).
[15] J. Brandt, J. Bauer, R.M. Frank, A. von Scholley, *Chem. Scr. 18* (1981) 53; J. Brandt, A. von Scholley, *Comput. Chem. 7* (1983) 51; J. Brandt, Habilitationsschrift, Technische Universität München (1981); J. Brandt, A. von Scholley, M. Wochner, K. Stadler, 188th ACS National Meeting, Division of Chemical Information, Symposium «Chemical Reaction Data Bases», Philadelphia, August (1984).
[16] J.H. Winter: *Chemische Syntheseplanung in Forschung und Industrie*, Springer, Berlin (1982); M.Wochner, I. Ugi, *Chem. Ind.* (1986) 498.
[17] I. Ugi, J. Bauer, J. Brandt, J. Friedrich, J. Gasteiger, C. Jochum, W. Schubert, J. Dugundji, in J. Bargon (Ed.): *Computational Methods in Chemistry*. Plenum, New York (1980), p. 275.
[18] G. Augelman, H. Fritz, G. Rihs, J. Streith, *J. Chem. Soc. Chem. Commun.* (1982) 1112.
[19] A. Defoin, G. Augelman, H. Fritz, G. Geoffroy, C. Schmidlin, J. Streith, *Helv. Chim. Acta 68* (1985) 1998.
[20] N.S. Zefirov, *Acc. Chem. Res. 20* (1987) 237.
[21] C.S. Wilcox, R.A. Levinson, *ACS Symp. Ser. 306* (1986) 209.
[22] S. Fujita, *J. Chem. Inf. Comput. Sci. 26* (1986) 205, 212, 224, 231, 238.
[23] M. Johnson, *Czech. Math. J. 37* (1987) 112; C.-C. Tsai, M. Johnson, V. Nicholson, M. Nain, in: *Graph Theory and Topology in Chemistry*, Vol. *51*, p.231, Elsevier, Amsterdam (1987).
[24] A.R. Battersby, R.J. Parry, *J. Chem. Soc. Chem. Commun.* (1971) 31.
[25] S.H. Bertz, *J. Am. Chem. Soc. 104* (1982) 5801.
[26] R.C. Read, D.G. Corneil, *J. Graph Theory 1* (1977) 339; G. Gati, *ibid. 3* (1979) 95.
[27] J.R. Ullmann, *J. Am. Chem. Soc. 98* (1976) 31; A.K.C. Wong, F.A. Akinniyi, *IEEE CH1962-0* (1983) 197; A.T. Brint, P. Willett, *J. Mol. Graph. 5* (1987) 49.
[28] G. Tinhofer: *Methoden der angewandten Graphentheorie*, Springer, Wien (1976); M.R. Garey, D.S. Johnson: *Computer and Intractability – A Guide to the Theory of NP-Completeness*, Freeman, San Francisco (1979).
[29] R.E. Burckard, U. Derigs: *Assignment and Matching Problems: Solution Methods with FORTRAN Programs*, Springer, Berlin (1980).
[30] W. Schubert, I. Ugi, *J. Am. Chem. Soc. 100* (1978) 37.
[31] W. Schubert, I. Ugi, *Chimia 33* (1979) 183.
[32] I. Ugi, J. Dugundji, R. Kopp, D. Marquarding: *Perspectives in Theoretical Stereochemistry, Lecture Note Series*, Vol. *36*, Springer, Berlin (1984), chap. 8.
[33] E.H. Sussenguth Jr., *J. Chem. Doc. 5* (1965) 36.
[34] J. Figueras, *J. Chem. Doc. 12* (1972) 237.
[35] H.L. Morgan, *J. Chem. Doc. 5* (1965) 107.
[36] R.S. Cahn, C.K. Ingold, *J. Chem. Soc.* (1951) 612; R.S. Cahn, C.K. Ingold, V. Prelog, *Experientia 12* (1956) 81; R.S. Cahn, C.K. Ingold, V. Prelog, *Angew. Chem. 78* (1966) 413; *Angew. Chem. Int. Ed. Engl. 5* (1966) 385; V. Prelog, G. Helmchen, *ibid. 94* (1982) 614 respectively *21* (1982) 567; see also: E.F. Meyer, *J. Comput. Chem. 1* (1980) 229.
[37] A. von Scholley, *J. Chem. Inf. Comput. Sci. 24* (1984) 235; V.J. Gillet, S.M. Welford, M.F. Lynch, P. Willett, J.M. Barnard, G.M. Downs, G. Manse, J. Thomson, *ibid. 26* (1986) 118.
[38] J.J. Sylvester, *Am. J. Math. 1* (1878) 64.
[39] R. Lindsay, B.G. Buchanan, E.A. Feigenbaum, J. Lederberg: *Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project*, McGraw Hill, New York (1980).
[40] J. Friedrich, I. Ugi, *Inf. Commun. Math. Chem. 6* (1979) 201; *J. Chem. Res. (S)* (1980) 70; *J. Chem. Res. (M)* (1980) 1301, 1401, 1501.
[41] P. Bruck, Proceedings of ARCH, *Anal. Chim. Acta*, in press.
[42] E. Meyer, Proceedings of the 8th ICCCRE, *Anal. Chim. Acta*, in press.
[43] W. Donner, GDCh-Konferenz «Computer im Labor», Frankfurt am Main, October (1986).
[44] M.R. Frierson, G. Klopman, H.S. Rosenkranz, *Environ. Mutagen. 8* (1986) 283.
[45] J.E. Dubois, in R.J. Feldmann, E.Hyde (Ed.): *Computer Representation and Manipulation of Chemical Structure Information*, Wiley, New York (1974); J.E. Dubois, Proceedings of the 8th ICCCRE, *Anal. Chim. Acta*, in press.
[46] K.C. Nicolaou, N.A. Petasis, J. Uenishi, R.E. Zipkin, *J. Am. Chem. Soc. 104* (1982) 5557; K.C. Nicolaou, N.A. Petasis, R.E. Zipkin, *ibid. 104* (1982) 5560.
[47] M. Wochner, I. Ugi, unpublished.

On June 15, 1988 Prof. Ivar Ugi has received the renowned Philip Morris Research Award «Challenge Future» in recognition of the Dugundji-Ugi Model, an algebraic representation of the logical structure of chemistry. This qualitative mathematical theory was developed in a joint effort by the late Prof. James Dugundji, a prominent topologist at the University of Southern California, Los Angeles, and I. Ugi. The Dugundji-Ugi Model serves as the theoretical foundation of this paper, and it is also the basis of a wide variety of computer programs for automated reasoning in chemistry and the deductive solution of chemical problems, such as the classification and documentation of molecular structures and chemical reactions, substructure analysis and correlation, retrosynthetic and bilateral design of syntheses, prediction of chemical systems and reactions, and the elucidation of the mechanism of chemical reactions and biochemical processes. Thus the Dugundji-Ugi Model is potentially useful in many areas that range from medicinal chemistry to industrial process design and protection of the environment. – The Prize «Challenge Future» is annually awarded to four outstanding scientists by the Philip Morris Foundation, sponsored by the Philip Morris Company, Munich.