# Organic Reaction Similarity in Information Processing

Alexander J. Lawson

Beilstein Institute, Varrentrappstrasse 40-42, D-6000 Frankfurt am Main, Germany

Received July 15, 1992

A simple model is presented for a noninteractive, computer-assisted recognition of classes of preparations and reactions ("analogous cases"). In particular, the integration and use of the system in the production of the *Beilstein Handbook* is discussed, and the economic advantages of the resulting tabular representation are outlined.

## INTRODUCTION

This paper was presented in a symposium concerned with "Similarity in Organic Chemistry" and deals ultimately with the question of recognition of similarity in organic reactions. The distinguishing feature of the subject as described here is nevertheless subtly different from the wealth of excellent research already reported[1] on this subject, since (as indicated in the title) the prime motor for the model described here is process driven: i.e., the application is part of a solution in the context of economic production factors. This aspect of information handling is rarely presented as such since rapid progress on many fronts usually reduces the value of the results significantly even before final publication takes place. With this risk clearly recognized, it is nevertheless of interest first to examine the background leading to the development and use of a reaction-similarity coder before discussing the nature of the coder itself. This should enable the reader to separate the more particular from the more general aspects of the topic. As in all applied research, it is important clearly to document the context of the application.

## PRIMARY CONTEXT: ORIGINAL PUBLICATION OF CHEMICAL INFORMATION

As is well-known, research organic chemists periodically publish their results, mostly in primary journals and patents. The publication generally centers around a concept and is only rarely a bare description of observations. This is a natural result of the scientific method. In many cases (I would insist in *all* cases), the concept is concerned directly or indirectly with an effect or property of an organic substance or substance class. This is a natural result of the central economic importance of applied organic chemistry. Properties and effects are invariably uniquely coded in the structure of individual chemicals. Changes in structure must bring about a change in properties. The process of changing one structure into another is a reaction, in the sense of this paper. Ultimately, therefore, the driving force for the publication of a reaction is the optimization of achieving an effect or property, although this is generally masked by the use of structural descriptions as a synonym for properties.

The above analysis is certainly an oversimplified model, but it is in good agreement with the historical development of the communication of information in organic chemistry. In particular, a strong dichotomy is evident: Structures (i.e., properties) are the essential facts of our science, while concepts (i.e., generalizations) are the currency of individual researchers. The special case of reactions is an overlap of both.

In accordance with this model, one generally finds the format of a typical publication split into two main sections, the first dedicated to the concept (Discussion of the Results), the second to the structures and properties (Experimental Section), while the description of reactions often occurs in both, depending on the *extent of generality* of the reaction concept described. The higher the degree of generality, the more likely the reaction is to be classified as a concept (and accordingly becomes part of the author's intellectual currency). In order to establish the degree of generality, reactions are often described for many analogs, often in the form of a table.

## SECONDARY CONTEXT: COLLECTIVE PUBLICATION OF CHEMICAL INFORMATION

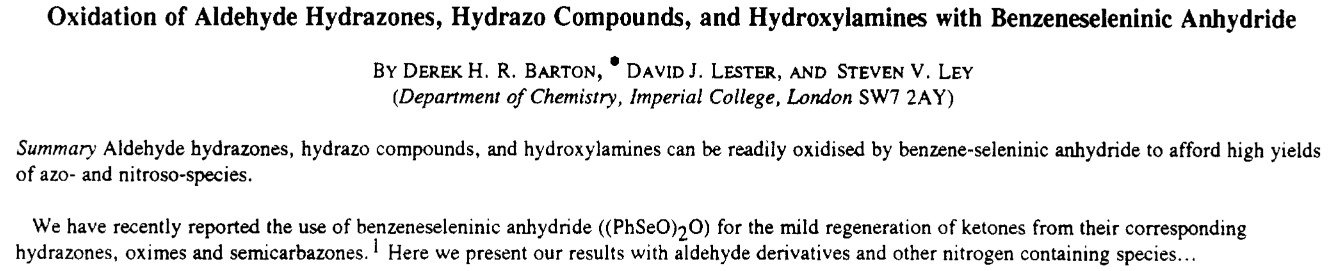
Further consequences of the model can be discerned in this area; the dichotomy outlined above has its clear expression in the historical development of secondary information. Concept-based systems have developed in parallel to property-based systems, although both have attempted to integrate the other aspect, and both have problems in the integration of reaction data. The most obvious example is the comparison between Beilstein and Chemical Abstracts. Beilstein is the more senior of the two and is a collection of structures and properties, ordered according to these structures, and pointing to the original publications. Chemical Abstracts is a collection of abstracts of the original publications, ordered (basically) according to the concept area and chronological publication date and pointing to the individual structures. Both systems have their distinct advantages, depending on the nature of the information sought.

Equally, both systems have their distinct disadvantages, as should be clear from the above analysis. Finding what a particular author said about a particular concept (for instance nonclassical carbonium ions) is well-nigh impossible in Beilstein, while the factual data to classes of structures (for instance the property p$K_a$ for a series of monounsaturated acyclic hydrocarbon carboxylic acids) is well-nigh impossible in Chemical Abstracts.

In the case of reactions and preparation data, Chemical Abstracts has distinct problems with the individual description, while Beilstein (traditionally very strong in that respect) conversely has distinct problems with the treatment of the conceptual aspects of reaction documentation.

The introduction of electronic information services over the past few decades has enabled some of these drawbacks to be reduced in their effect for both systems, but the fundamental problems remain. This paper deals with the Beilstein approach to the problem of summarizing conceptual reaction information in the production of the printed *Handbook.*

0095-2338/92/1632-0675$03.00/0



**Oxidation of Aldehyde Hydrazones, Hydrazo Compounds, and Hydroxylamines with Benzeneseleninic Anhydride**

By Derek H. R. Barton, * David J. Lester, and Steven V. Ley
(*Department of Chemistry, Imperial College, London* SW7 2AY)

*Summary* Aldehyde hydrazones, hydrazo compounds, and hydroxylamines can be readily oxidised by benzene-seleninic anhydride to afford high yields of azo- and nitroso-species.

We have recently reported the use of benzeneseleninic anhydride ($(PhSeO)_2O$) for the mild regeneration of ketones from their corresponding hydrazones, oximes and semicarbazones.[1] Here we present our results with aldehyde derivatives and other nitrogen containing species...

**Figure 1.** Reprinted by permission of the Royal Society of Chemistry (from *J. Chem. Soc., Chem. Commun.* **1978**, 276).

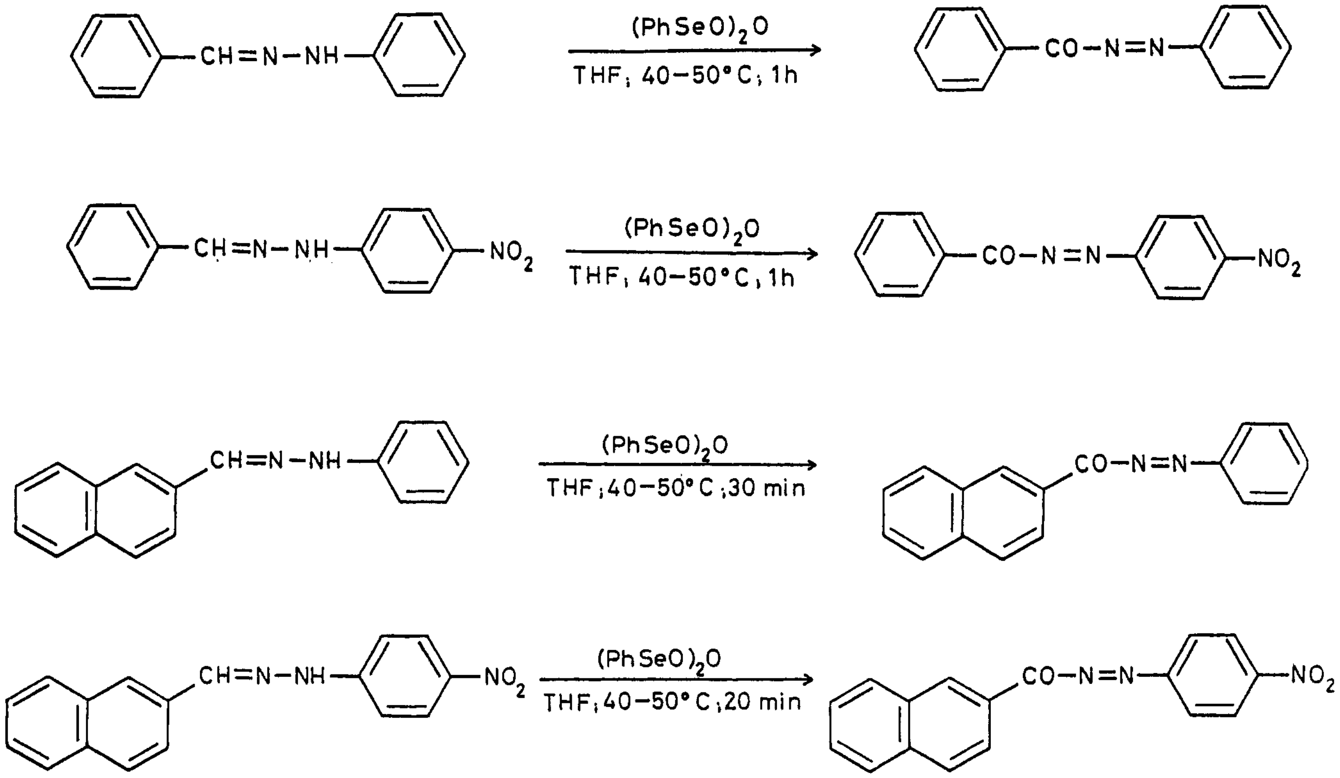


**Figure 2.** Oxidation of hydrazones.

## NATURE OF THE PROBLEM

Beilstein is produced by a twofold process:

(1) **Abstracting Step.** The original publication is read, and properties, data, and reactions are abstracted for each individual structure in the publication. At this early stage, the conceptual relationship between individual reactions is retained only in the intrinsic similarity (or otherwise) of the exact data as recorded for each individual description and the binding pointer to the same original publication. The tabular context of examples to describe generality (see above) is lost at this stage.

(2) **Editing Step.** These data are then collected for all publications and sorted according to individual structural criteria. The context of the original publication becomes subordinate to the context of the individual structure. The data are then checked in this latter context and published.

In the course of the second step, it frequently occurs that the original context must be taken into account on the desk of the editor. One example will illustrate this nicely.

Figure 1 shows the first few lines of a typical publication dealing with a reaction method. The authors have stated their concept three times in the first few lines of the paper. Obviously, one reagent is described in its use for three distinct (but related) reactions. Further reading of the paper reveals that each reaction is illustrated by several distinct examples (establishing the generality).

In the Beilstein course of events, each of these distinct examples is an entity in its own right and is recorded as such. In the interests of brevity, we will follow only one of these three reactions (the oxidation of hydrazones) in its further path.

There were four individual examples, as shown in Figure 2. The four entities then enter a fully automated process in which the structures are registered, the data are recorded, and systematic nomenclature[2] is assigned where possible. Sorting is then automatically carried out on the basis of the Beilstein System,[3] a text-generator performs the first draft of the compound entries, and these are then downloaded to the editor for inspection and checking on the basis of the original publication. Note that the entities have now a completely separate existence. They are now collated in the first instance together with data on the same and similar compounds. However, due to the rigor of the structural sorting algorithm, it often happens that these independent entities reappear consecutively in the manuscript for editing. In this case, the initial result has the appearance of Figure 3. To date, the editor must then manually bring these into a more economical form (Figure 4). This work reflects nothing more than a reestablishment of the tabular context of the original pub-

**Benzoyl-phenyl-diazene** $C_{13}H_{10}N_2O$, formula Z.
*Prep.* From benzaldehyde phenylhydrazone [$(PhSeO)_2O$; THF] (*D.H.R. Barton et.al.*, J.C.S. Chem. Commun. **1978** 276-277).

**(Naphthalene-2-carbonyl)-phenyl-diazene** $C_{17}H_{12}N_2O$, formula Z.
*Prep.* From naphthalene-2-carbaldehyde phenylhydrazone [$(PhSeO)_2O$; THF] (*D.H.R. Barton et.al.*, J.C.S. Chem. Commun. **1978** 276-277).

**Benzoyl-(4-nitro-phenyl)-diazene** $C_{13}H_9N_3O_3$, formula Z.
*Prep.* From benzaldehyde 4-nitro-phenylhydrazone [$(PhSeO)_2O$; THF] (*D.H.R. Barton et.al.*, J.C.S. Chem. Commun. **1978** 276-277).

**(Naphthalene-2-carbonyl)-(4-nitro-phenyl)-diazene** $C_{17}H_{11}N_3O_3$, formula Z.
*Prep.* From naphthalene-2-carbaldehyde 4-nitro-phenylhydrazone [$(PhSeO)_2O$; THF] (*D.H.R. Barton et.al.*, J.C.S. Chem. Commun. **1978** 276-277).

**Figure 3.** Beilstein text before analysis of the reaction context.

**Benzoyl-phenyl-diazene** $C_{13}H_{10}N_2O$, formula Z.
*Prep.* From benzaldehyde phenylhydrazone [$(PhSeO)_2O$; THF] (*D.H.R. Barton et.al.*, J.C.S. Chem. Commun. **1978** 276-277).
*Similarly prepared :*
**(Naphthalene-2-carbonyl)-phenyl-diazene** $C_{17}H_{12}N_2O$. (*Ba. et. al.*)
**Benzoyl-(4-nitro-phenyl)-diazene** $C_{13}H_9N_3O_3$. (*Ba. et. al.*)
**(Naphthalene-2-carbonyl)-(4-nitro-phenyl)-diazene** $C_{17}H_{11}N_3O_3$. (*Ba. et. al.*)

**Figure 4.** Beilstein text after analysis of the reaction context.

lication, a context which (as noted above) has been lost as a natural consequence of the process.

The driving force for this production step lies clearly on the user side; it is a question of presenting the user of the system with a concise and accurate representation of the data with a minimum of redundancy. The effect of this data compression may appear to be small for the example cited here, but the overall effect on print volume, up-to-dateness, and transparency to the user should not be underestimated. In economic terms, this one item in the production process has a savings of many millions of dollars per year when all costs (including library shelf space, user time, type setting, and materials) are taken into account. Nevertheless, the internal costs at the Institute are also considerable, and the compression of the data is not only labor-intensive but also boring for highly trained chemists, whose true vocation is a critical assessment of the chemical content of the data.

## REACTION-SIMILARITY CODER

The solution to this problem clearly lies in the automatic recognition of the reaction context which binds these examples together. This difficult task is further complicated by the boundary conditions imposed by the process itself. The human eye recognizes similarity instantly, but the interpretation of similarity is notoriously context dependent; for this reason most similarity-driven processes find their expression at the interactive user interface. Here, however, the problem calls for a type of classification with no interactive component. Furthermore, due to the largely decentralized nature of the Beilstein data collation process,[4] is would be most advantageous to carry out the context recognition as near to the source as possible (to facilitate checking of the data input at an early stage). This includes the computing environment of the personal computer rather than the mainframe. Thus, the resources available for the coder are limited. In particular, the processing time must be short, on the order of a second for each reaction input.

At the point of time of the abstracting step, the data available includes structure connection tables (CTs) and reaction conditions (alpha-numeric data of various sorts). The reaction conditions are useful in many ways for a general classification of reactions (e.g., the Birch reduction in liquid ammonia), but from the above discussion, it should be clear that the major aim of the coder here described must be structural in nature.

The structural input consists of two linked structure diagrams, in the form of CTs. The first diagram contains educts (starting materials), and the second diagram contains products. Both educt and product diagrams may contain many discrete structures. There is no guarantee of an overall explicit atom balance, nor is the stoichiometry of the individual reaction participants in any way explicit.

The first step in the analysis concentrates on the isolation of extended reaction centers. This is carried out by the use of the fragmentation algorithm of the Beilstein System, which isolates carbon-complete fragments using a formal hydrolysis scheme.[5] The fragments are coded using the RABBIT[6] procedure, which relies on an extension of the LN algorithm[7] to generate an extremely specific hash code (with disregard to stereochemistry) for each fragment. Fragments identical in educt and product assemblies are then removed from the matrix to leave elements which (by definition) undergo some change in the course of reaction. These elements are then mapped onto each other on the basis of common surviving features (skeletal and functional), and a recursive introduction of multiple amounts of the appropriate educts is carried out until a plausible account of the product is achieved.

Each fragment pair now defines an extended reaction center, whereby the limiting condition for pair mapping is that at least one carbon atom of the educt fragment is mapped to at least one carbon atom of the product fragment. Normally, the mapping produces one or two extended reaction centers, although three and higher are possible in complex reactions involving C–C-fusion or C–C-fission reactions (A + B + C → D; A → B + C + D).

**Table I.** Bit Mapping of Changes in Chemical Functionality.

| | | |
|---|---|---|
| 1 | hydroxy | C—O—R |
| 2 | carbonyl | C=X |
| 3 | carboxylic acid | C(=X)—X—R |
| 4 | sulfur acids | C—S(=X)—X—R |
| 5 | amine | C-N-R |
| 6 | hydroxylamine | C-N-O-R |
| 7 | azine | C-N-N..R |
| 8 | metallorganics | C-met |

**Table II.** Bit Mapping of Changes in Skeletal Features.

| | |
|---|---|
| 1 | SANDRA code |
| 2 | shape hash |
| 3 | number ring-O atoms |
| 4 | number ring-N atoms |
| 5 | ring class (acyclic, isocyclic, etc.) |
| 6 | carbon number |
| 7 | degree of unsaturation |
| 8 | functional multiplicity |

The coding of each reaction center is now carried out on the basis of eight descriptors (in the order given by the numbers below):

**Chemical Functionality** (functional groups).

(1) which groups disappear in the course of reaction

(2) which groups appear in the course of reaction

(5) which groups remain unaltered in the course of reaction

**Skeletal Features** (functional multiplicity, degree of unsaturation, carbon number, ring-heteroatom count, ring class, shape function)

(3) which features change in the course of reaction

(4) numerical sign of change of each feature

(6) discrete change in carbon count

(7) discrete change in degree of unsaturation

(8) discrete change in the shape function

Exactly 1 byte is required for each of these descriptors, as described below. Thus, each extended reaction center is coded in 8 bytes, and the coding for the total reaction is expressed as a concatenation of the sorted 8-byte strings. In practice, the use of three strings is quite sufficient for an adequate description, so that a maximum length of 24 bytes can be allocated for the reaction coding.

## BIT MAPPING OF CHEMICAL FUNCTIONALITY

Each byte is regarded as a bit string for eight parameters in any combination. The chemical function parameters used will now be briefly discussed. The key to the operation of this coder lies first and foremost in the ability to recognize chemical functionality. The Beilstein System has long stood the test of time in this respect, and the present algorithm is built on (a somewhat simplified form of) this system. In this description, all exocyclic and semicyclic bonds to heteroatoms are regarded as part of either functional groups or substituents. Substituents are formally defined as chemically inert to further derivatization (halogen atoms, nitro units), while functional groups may be further modified (or masked) by the covalent attachment of carbon moieties (alcohols, acids, amines). The Beilstein System also recognizes all divalent chalcogens (S, Se, Te) as being analogs of the corresponding oxygen functions. The process of unmasking any given molecule into its component functionalities is a relatively simple task for which a PC-based algorithm (SANDRA[8]) has long been in frequent use. The present coder (RFINESSE) uses a slight variation of the SANDRA algorithm, somewhat simplified to compress the basic set of chemical functions into exactly eight possibilities in the first instance. These are listed in Table I. Any given combination of atoms in an organic molecule will be analyzed to produce a combination of these formal functional units attached to a carbon skeleton and, hence, a byte value between 0 (no function) and 255 (all eight functions present).

The interested reader will have noticed that the term "metallorganics" is a blanket term to cover a wide variety of functionality, including carbon–phosphorus, –arsenic, and –silicon compounds (to name the most common). This detail is recorded explicitly in the SANDRA hash (see below under features) as is also other secondary information (substituents, chalcogen exchange).

## BIT MAPPING OF SKELETAL FEATURES

The second aspect of coding concerns the nature of the carbon skeleton to which the substituents and/or functional groups are attached. This is classified according to the nature of the rings (if any), the degree of C unsaturation, the multiplicity of attached functions (dihydroxy), and the general shape of the skeleton, described here as a hash function which measures the mutual proximity (along the skeleton backbone) of skeletal features, including the points of attachment of functional groups. The parameter list is shown in Table II. The essential point here is that each of these factors can be represented by an integer number (e.g., 2 ring-N, 1 ring-S) and (with the exceptions of the shape hash and the SANDRA hash) is a number which has an immediate natural relevance for the chemist, and as such, these mimic the factors which play an immediate part in the recognition of structural similarity by the human eye. One major exception is the neglect of aromaticity in this part of the coding. The differences (as recognized by the shape function) between 2-methylphenol, 6-methylenecyclohexa-2,4-dienol, and benzyl alcohol are distinct but unexceptional.

## REACTIONS WITH C–C FUSION OR FISSION

One interesting aspect of the coding is the natural result of the treatment of an extended reaction center as the composite of two fragments; in the case of reactions involving carbon–carbon fusion or fission (A + B → C or A → B + C), the reaction is coded simultaneously from the "point of view" of each of the participating carbon-based units. More simply put, the Friedel–Crafts reaction between chloroethane and benzene is coded in parallel as "ethylation of benzene" and "phenylation of ethane". [This dual representation is a key aspect of the more general RABBIT indexing,[6] which deals with multistep reactions, and incorporates the present model of reaction similarity. Clearly, in a multistep process the reaction chain must be allowed to pass through either of the educts (chloroethane or benzene) to the product (ethylbenzene) and beyond.]

## CHOICE OF DEGREE OF "FUZZINESS" IN CLASSIFICATION

In general, the full resolution of the coder is too detailed for an effective clustering of the set. For instance, the desired class concept of the above-mentioned Friedel–Crafts reaction is more likely to be "C-alkylation of cyclic hydrocarbons" than the more precise terms given above. In terms of the number of descriptors used in the coding, this means that the use of the first five descriptors is often optimal. The remaining three (discrete) terms would usually separate the tabular reactions too finely ("methylation, ethylation, propylation ...of... cyclohexane, cycloheptane, cyclooctane") for use in the sense here desired.

It should be stressed at this point that the coder does not in general classify reactions in the normal mechanistic nomenclature of the chemist. To take a simple example, the Beckmann rearrangements of cyclohexanone oxime, (*E*)-acetophenone oxime, and (*Z*)-acetophenone oxime show different coding patterns, since the nature of the respective products (oxo heterocycle, cyclic amine + acyclic acid, acyclic amine + cyclic acid) are different. The coding for the two acetophenone stereoisomers is naturally more similar than that for the intramolecular rearrangement to caprolactam, but the main point is clear: there is no coding specific or typical for "a Beckmann rearrangement".

## SUMMARY OF RESULTS

Reactions can be classified in a chemical context on the basis of the resulting codes. For the purposes of the task in hand (as noted above), classification using the **first five descriptors** as a class identifier gives an excellent fit with chemical intuition. There is clearly no a priori reason for choosing the first five descriptors. The important point is that the codes (generated immediately in the context of the single publication) allow a fully automatic inspection of the assembly of all reactions described in the paper in question and establish the "best fit" for the number of descriptors used for the reaction context contained therein.

In the particular example mentioned in the Introduction, the algorithm used here separated the assembly into the three general reactions mentioned by the authors in their abstract. The tabular examples of each reaction fell neatly into the classes defined by the use of the first five descriptors.

Performance tests on a file (ca. 20 000 preparations and reactions from 4000 publications randomly selected from the current literature) showed that the coder is sufficiently fast (ca. 6 h batch time on a 386-Compaq) and sufficiently selective (reduction to ca. 10 000) for the task in hand.

## OTHER USES FOR THE CODER

The coding described here has its prime function in the cost reduction of the production of the printed *Beilstein Handbook*. However, other uses can be envisaged. Since the 8-byte code is designed as a bit screen, it is possible to implement the results into a retrieval tool for large amounts of reactions. Possibly more interesting is the management of hit-list display after the search has been carried out. Experience shows that users of computerized systems in general are currently faced in many cases by too much information, even after a carefully constructed search strategy. In the case of reactions, it should be possible to implement a tabular display in analogy to the intention of the original author.

The results of further research in this direction will be reported at a later date.

## REFERENCES AND NOTES

(1) (a) Vladutz, G. Do we still need a classification of reactions. In *Modern Approaches to Chemical Reaction Searching*; Willett, P., Ed.; Gower Publishing Co.: Brookfield, VT, 1986; p 202. (b) Willet, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchforth, Hertfortshire, England, 1987. (c) Grethe, G.; Moock, T. E. Similarity Searching in REACCS. A New Tool for the Synthetic Chemist. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 511–520 and references therein. (d) Bawden, D. Classification of Chemical Reactions: Potential, Possibilities, and Continuing Relevance. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 212–216 and references therein.

(2) Goebels, L.; Lawson, A. J.; Wisniewski, J. L. AUTONOM: System for Computer Translation of Structural Diagrams into IUPAC-Compatible Names. 2. Nomenclature of Chains and Rings. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 216–225.

(3) (a) Beilstein, F. K. *Handbuch der Organischen Chemie* (Erste Auflage); Bd. I, Springer-Verlag: Berlin, 1881. (b) Prager, P.; Jacobson, P. *Beilstein's Handbuch der Organischen Chemie* (Vierte Auflage); Bd. I, Julius Springer-Verlag: Berlin, 1918. See also ref 5.

(4) Jochum, C. J. In *The Beilstein Online Database*; Heller, S. R., Ed.; ACD Symposium Series No. 436; American Chemical Society: Washington, DC, 1990, pp 10–23 and references therein.

(5) Lawson, A. J. In *Software Entwicklung in der Chemie 2*; Gasteiger, J., Ed.; Springer-Verlag: Heidelberg, 1988; p 1.

(6) Lawson, A. J.; Kallies, H. Multistep Reactions: The RABBIT Approach. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 426–430.

(7) Lawson, A. J. In *The Beilstein Online Database*; Heller, S. R., Ed.; ACS Symposium Series No. 436; American Chemical Society: Washington, DC 1990, pp 143–145.

(8) Lawson, A. J. In *Graphics for Chemical Structures*; Warr, W., Ed.; ACS Symposium Series No. 341; American Chemical Society: Washington, DC 1987; p 80.