J. Chem. Inf. Comput. Sci. **1990**, *30*, 426–430

Multistep Reactions: The RABBIT Approach

ALEXANDER J. LAWSON* and HARTMUT KALLIES

Beilstein Institute, Varrentrappstrasse 40-42, D-6000 Frankfurt/Main 90, FRG

Received August 10, 1990

A simple concept (RABBIT) for the coding of multistep reactions is presented. A RABBIT (**R**andom **A**ccess **B**lack **B**ox **I**ndexing **T**erm) code contains purely structural information and may be represented as a vector in two-dimensional space. Applications of the concept on a PC basis allow automatic structural checking to maintain chemical relevance in the tree search of the multistep path at query time. An example is discussed from an experimental implementation of the concept.

INTRODUCTION

The present authors are involved in the planning of a Beilstein Reaction Database.[1,2] This paper will accordingly concentrate on two specific aspects of computerized reaction management which are essential to the success of this project, over and above the current state-of-the-art achieved by commercially available systems.[3] These two aspects are

(A) Multistep reactions
(B) Analogy of reactions

The reasons why these two aspects are of central importance to the Beilstein project are briefly summarized as follows.

(A) Multistep Reactions. The printed *Beilstein Handbook* and its online counterpart both contain the seeds of a potentially outstanding database on chemical reactions, with a strong bias to synthetic usefulness. This is an automatic result of the Beilstein data structure of the printed work, in which individual chemical compounds are linked in a direct educt–product relationship after due consideration of the relevance and correctness of the individual publication in the original literature.[4] There is a further constraint which is operative, although this is not widely recognized: it has always been Beilstein policy to formulate an individual description of an organic preparation in terms of organic educts (starting materials) which are themselves present (or due to be present) in the *Handbook* with a corresponding entry, including details of preparation.

It is therefore immediately obvious that Beilstein data on preparations form an explicit *network* of recorded one-step transformations; this being equivalent to a map of practically all known synthetic pathways from almost any starting material to almost any product. This is a supportable statement within the boundaries of the reported chemical literature *as a whole* (as currently covered by the Beilstein products[5]) and is not limited by the constraints of any single publication in the original literature. Any reaction management system which wishes to do full justice to the Beilstein data must therefore make full use of this powerful multistep knowledge base.

(B) Analogy of Reactions. Reactions are traditionally regarded by chemists from a strongly classified[6] standpoint: reaction classes and methods mostly have names (Beckmann rearrangement, Diels–Alder addition, etc.) or potential names insofar as they are often associated with the chemist who propagates the method [Pd(0)-catalyzed allylation according to B. Trost, etc.]. One problem with this instinctive classification is that it is often not clear whether one is referring to a mechanistic assessment (detailed description of the reaction path, including transition states) or whether the key aspect is "what goes in and what comes out" ("Here-To-There" approach, HTT). While both aspects are important to users of a reaction database, in this paper we will concentrate on the HTT query, which is very firmly based in practice. The chemist is assumed to have a particular structural transformation in mind and wants to know how this can be accomplished. Here the principle of analogy can play a very strong role in the relevance of the information sought: a reaction is a description of *changes* taking place in a specific environment, and the case where the changes are locally *identical* although the starting points are only *similar* is in many cases a satisfactory answer for the chemist. One simple example is the transformation of benzophenone oxime, (**1**) to benzophenone imine (**2**) (see Figure 1).

Assuming for the moment that this specific reaction is being searched, and further that no such entry exists in the database, then the searcher might be interested to know that the analogous process in the fluorenone series (**3** to **4**) has been described, or that (even more relevant) the transformation **5** to **6** is known.

Naturally, this type of information can be obtained today from existing reaction database systems:[7] an appropriate query based on substructure searching is capable of finding these answers. The difficult question of *how relevant* these answers may be to any particular query will be deferred for the moment. It suffices to note that a high degree of subjectivity is involved and that most chemists would ascribe some degree of relevance to the answer set.

It should be noted, however, that the situation becomes much more complicated when the transformation of **1** to **2** is a logical part of the multistep synthetic pathway. Suppose the actual desired HTT query was from benzophenone (**7**) to 3,3-diphenyloxaziridine (**8**). Let us further suppose that the single step **7** to **1** is not explicitly described in the database (since this is an absolutely trivial derivatization), but that **2** to **8** is known. Remember that we have stipulated that **1** to **2** is not known, although the analogues **3** to **4** and **5** to **6** are well described.

Here we have a classic case of analogous thought, albeit with a simple reaction scheme. The chemist, once presented with the information in the database, would interpret a viable three-step reaction from **7** to **8**. However, a computer-based management system would be hard pressed to provide the searcher with this information merely on the basis of definition of the starting point **7** and end point **8**. This is one type of skill which would be invaluable when used in conjunction with the Beilstein Reaction Database.

THE RABBIT CONCEPT

To test these ideas we developed the RABBIT concept. RABBIT is not a program, but rather an indexing concept for certain aspects of storage and retrieval of information on reactions, with particular emphasis on reaction similarity and multistep processes. In the confines of this paper the term

0095-2338/90/1630-0426$02.50/0

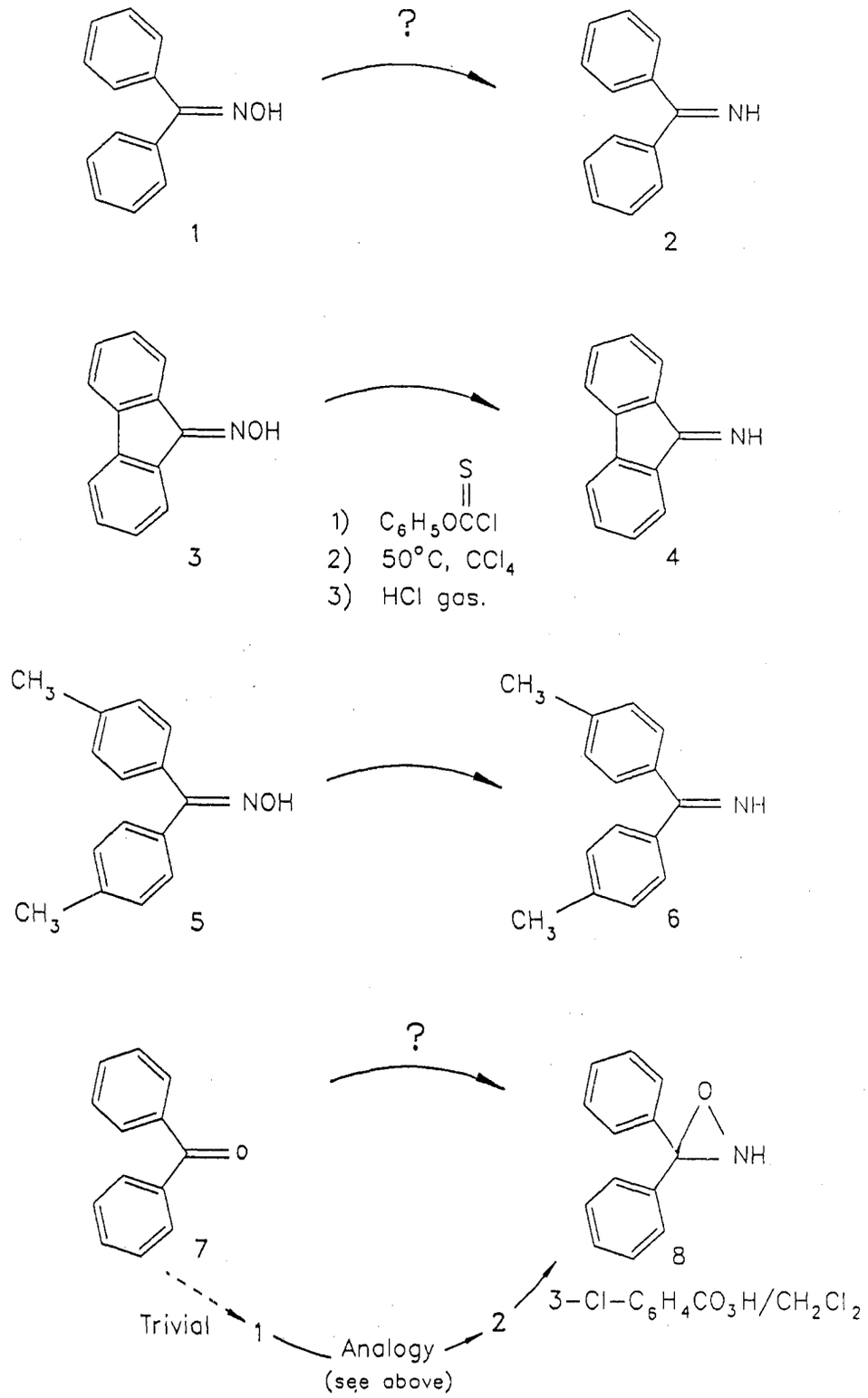

Figure 1. Example of a reaction scheme with analogous processes (see text).

"reaction" will be limited to consideration of the purely structural changes taking place in organic molecules; i.e., aspects of descriptions of conditions (yields, solvents, catalysts, temperature, etc.) will not be dealt with, important as these are.

Multistep Transformations: Some of the Problems. From the point of view of computer-based systems, the HTT approach is more easily treated than mechanistic classifications, since it is easier to define: connection tables (CT) for educts (starting materials) and products then lead to reaction-center matrixes which can then be manipulated in a number of ways, in combination with the usual tools (substructure search, etc.) which are available for the CTs of educts and products.[2] This approach is widely implemented, in one form or another, in the major reaction-information systems available today. The instincts of the chemist are accommodated by the ability to concentrate his search on the structural change at the reaction center, qualified by further boundary conditions imposed by supplementary structural features of the educt and the product.

However, this position is not so satisfactory if one extends the HTT query to include multistep transformations. The conventional approach defines each "reaction" as a single

structural transformation. However, chemists actually often require information of the following type: starting from compound A, how can I *best* get to compound B. The word "best" can carry a number of implications (yield, cost, purity, etc.), but it is not a priori confined to single-step transformations. Most present systems tackle this problem with an implicit requirement that the user concentrates his actual chemical query into a series of key one-step queries.

The importance of true multistep queries increases with the size (coverage) of the database. Unfortunately, the problems involved in satisfying such queries increase also. There are problems at registration time (database generation) and at query time (user).

The task could be solved in principle by establishing (at registration time) all possible routes from all educts to all products: this involves an astronomical degree of computer power (recursive registration) and is clearly out of the question. However, it should be mentioned that a related process is viable in practice: registration of all routes from all educts to all products *in the confines of a single publication.* This is the solution being developed by certain systems today,[8] although the approach has obvious drawbacks.

At first sight, a second solution appears to be attractive: at registration time each educt points to the corresponding product of one-step reactions, and inverted lists based on Registry Numbers (RN) are created. At query time, these lists are successively manipulated on the basis of branching trees. In practice, it is obvious that this approach has one fundamental flaw: the RN contains no structural information per se. This fact leads in turn to two key drawbacks in practice:

(i) The degree of branching at query time is not contained to relevant paths. This leads to an immediate explosion of the data with a consequent breakdown of the process.

(ii) A relevant path through a database will often contain logical "jumps" (see Figure 1) which are chemically valid but not explicitly registered. An obvious example is the case of the branch leading to a compound with an unmodified group (acid, carbonyl, amine, etc.) and the further path continuing from a trivially modified derivative (ester, oxime, salt, etc.), although the explicit pointer for the process of modification is not set. This causes a valid path to be missed.

The way out of these dilemmas requires the interpretation of structural information at each step in the tree search. The interpretation algorithm itself must be complex, since it demands a measure of the relevance of the alternative steps open to the process. One possible approach is the use of "chemical distance"[9] but irrespective of the algorithm, the manipulation of connection tables at each step would involve considerable computation and thus affect processing times adversely.

An alternative possibility is to combine certain structural information with the registration information in such a way that the interim manipulation of CTs is avoided. This boils down to the use of a concise, efficient, and relevant coding system for any single-step transformation, which automatically relates (by its value) to the next relevant single-step transformation in a multistep process. This is the basis of the approach used in the RABBIT concept.

The RABBIT Code Expressed As a Straight Line. The fundamental idea behind RABBIT coding is simple and can best be illustrated by the use of an analogy; namely, that any structural change can be represented by the equivalent of a straight line, as drawn on two-dimensional graph paper. In this representation, a line is defined by:

- the coordinates of a starting point
- the coordinates of a finishing point
- further attributes of the line (e.g., color of line)

This model is convenient for the discussion of RABBIT coding, since it is easily accommodated by a graphical representation on computer screens and can be fitted to chemical logic. Thus, the coordinates of the starting and end points of the line represent the reacting environments of the educt and product structures, respectively, while the attribute "color" (for instance) represents the classification of the *changes* in structure involved in moving from educt to product. Assuming for the moment that suitable algorithms can be found for the generation (directly from CTs) of coordinates and color in the above sense, it can be readily appreciated that the manipulation of RABBIT "lines":

- is computationally simple and fast
- allows quick comparisons of similarity (on the basis of gradient and color)
- allows the automatic assessment by the system itself of whether a "path-jump" (see above) is reasonable
- allows the system to assess the missing "lines" in an otherwise complete path, and go searching for a suitable analogue "line" in the database

There is nothing sophisticated about this model, as the informed reader will appreciate. But the above advantages arise from the speed of computation made possible by the simplicity. RABBIT stands for **R**andom **A**ccess **B**lack **B**ox **I**ndexing **T**erms, but the **R** could just as easily be used for **R**apid.

EXAMPLE FROM A TEST DATABASE

The test database was constructed on the basis of independently entered, completely uncoupled single-step transformations. The material used was not in any sense a complete compilation of all organic chemistry, but contained many interesting reactions. Then a particular multistep query was posed (for the purpose of illustration, this example is chemically simple, which in no way influences the principles involved here). The query: What is the best path from benzene (**9**) to aniline (**11**) (Figure 2).

The response of the system was threefold (in a matter of seconds):

Route A:

(I) nitration to nitrobenzene (**10**)
(II) reduction of nitrobenzene to aniline (**11**)

This is a standard method and needs no further comment.

Route B:

(I) alkylation of **9** to ethylbenzene (**12**)
(II) selective oxidation of **12** to acetophenone (**13**)
(III) Beckmann rearrangement of acetophenone oxime (**14**) to acetanilide (**15**)

This path requires some comment. With respect to the chemistry of the suggestion, the RABBIT system cannot be judged on the value of the route: each of these single-step processes was present in the database, and as such the answer is valid. However, there are two points to be noted here. First, the system automatically performed a "path-jump", by allowing the equivalence of the ketone with the oxime (this fact was of course drawn to the user's attention). Secondly, the system allowed the product actanilide to be set equivalent to aniline itself for the purposes of an HTT query. The trivial hydrolysis and oximation reactions of these particular compounds were not explicitly present in the database.

Route C:

(I) bromination of **9** to bromobenzene (**16**)
(II) conversion to the Grignard **17**
(III) carboxylation to benzoic acid (**18**)
(IV) conversion to the sulfoxide **19**
(V) reduction of **19** to acetophenone (**13**)

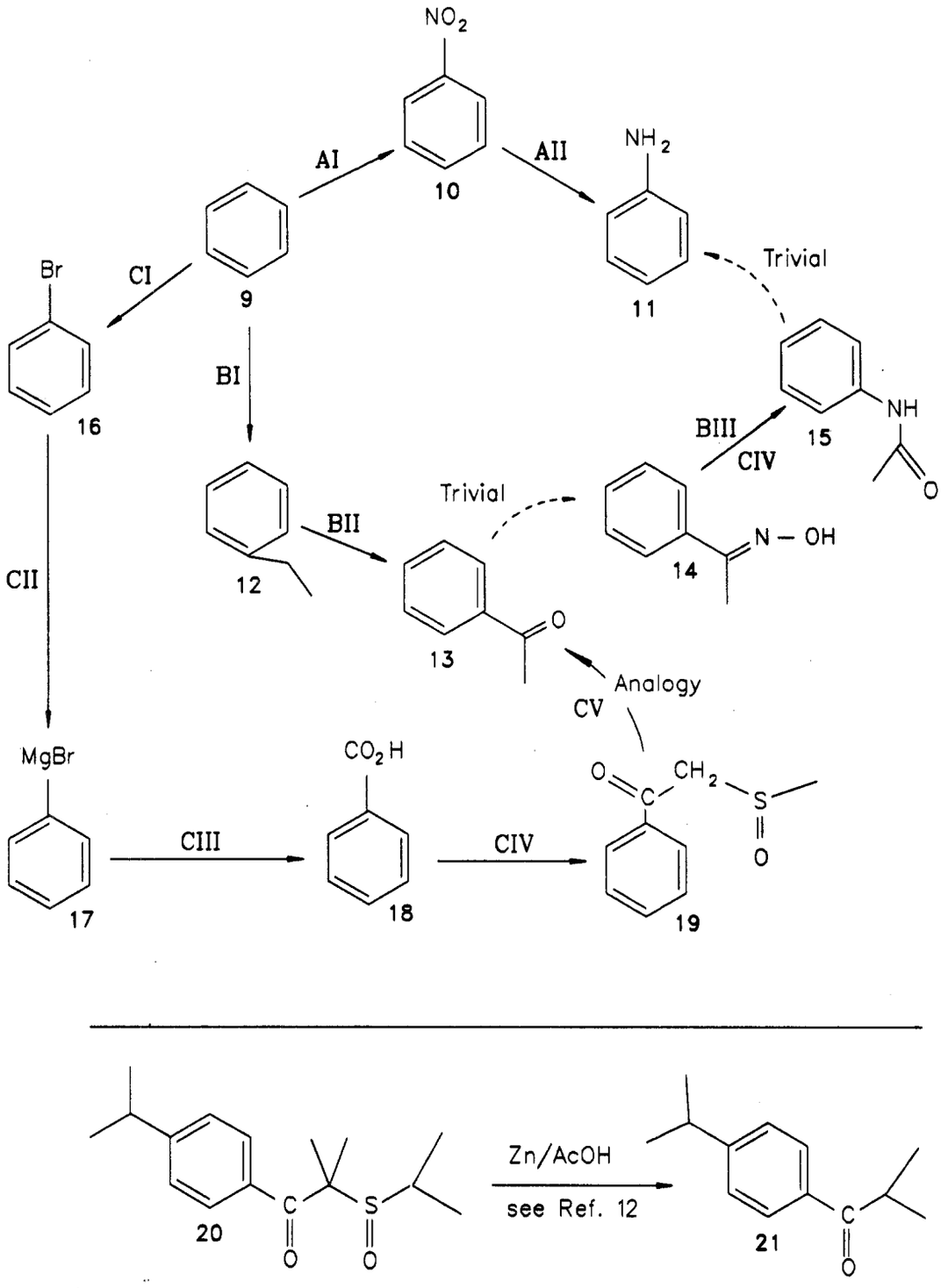

Figure 2. Example of a multistep query from the test database.

(VI) Beckmann rearrangement of acetophenone oxime (**14**) to acetanilide (**15**)

The point to note here is the reaction V in Route C. This reaction was not present in the database. It is a pure suggestion of the system, a "missing line" in an otherwise complete system of lines (as noted in the previous section). The reaction is the result of the system's constant automatic checking for analogous reactions, wherupon the model **20** to **21** (Figure 2) was found in the database, and found to be relevant (similarity of lines within the tolerances set by the user at run-time). Naturally, the source of the suggested reaction and the nature of the model are communicated to the user. The important aspect is the completely automatic nature of the definition of the relevant searches which the system itself then automatically performs.

Note to the Coding Used in This Test. The RABBIT concept was designed to operate with large databases (at least several hundred thousand reactions) with acceptable response times (and answer sets) using the computing power of personal computers alone. To meet these demands, it is clear that the decisive factor is a compact, well-distributed code for structural changes at an extended reaction site, as indicated above. We chose aspects of the LN algorithm[10,11] for initial testing in this report, since the code is known to be very evenly distributed, very compact (2 bytes), and of high resolution, allows a predictable degree of similarity, and is rapidly generated on PCs directly from the structure. In principle any similar (or more sophisticated) algorithm could produce analogous results.

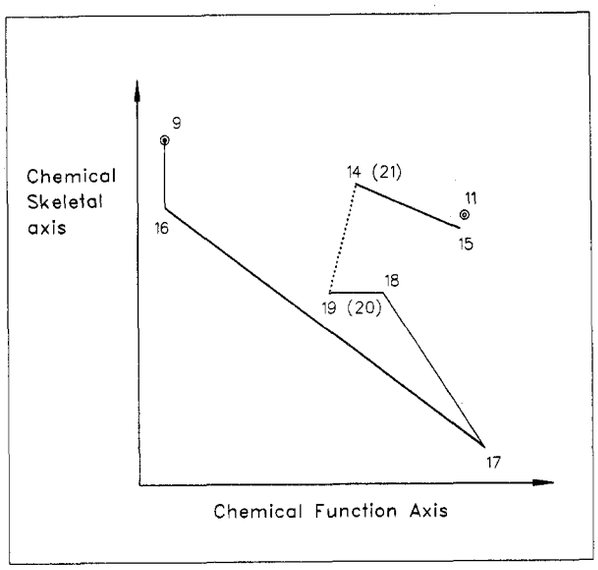

Figure 3. Schematic representation of the RABBIT analysis of route C (see Figure 2) from benzene (**9**) to aniline (**11**). (Line "color" not shown.) The axes are based on the elements of chemical function and chemical skeletal aspects of the Lawson Number (LN) hash code for educt- and product-reacting environments.

We also chose to separate and extend the aspects of chemical functionality and skeletal morphology which are analyzed in the generation of the LN hashcode. Each of these aspects can be then be described in 2 bytes, making a total description of 10 bytes (2 * 2 *X* coordinates, 2 * 2 *Y* coordinates, + 2 for the line attribute). Typical for the extension with respect to skeletal morphology (*Y* axis of Figure 3) was a quantification of the relative effects of chemical node features separated by an integral number of edges, diminishing with increasing distance. The values assigned to the standard parameters were in fact adjusted to give a useful spread in a significant sample set, but no true optimization has yet been carried out. The chemical function coding (*X* axis) followed the LN algorithm closely.

The specific example of Figure 3 shows the type of line tracing resulting from the RABBIT concept for Route C (above), including the missing line (**19** → **14**) suggested automatically by the system after searching for (and finding) the analogous process **20** → **21**. It can be readily seen that the successive introduction of the groups bromo, Grignard, carboxylic acid, keto sulfoxide, and amine all cause a marked distortion of the skeletal morphology of the benzene ring (as measured by the *Y* hash), the most marked effect being due to the organometallic. When one notes that the units of *X* axis, *Y* axis, and the line attribute each generally have a resolution better than ca. 1 part in 10 000 structures, it is clear that the RABBIT coding of extended reaction centers is an extremely selective hash; practically the only source of collisions is found in the (present) complete neglect of stereochemical aspects. Further testing will be carried out on this aspect in the near future.

REFERENCES AND NOTES

(1) Jochum, C.; Lawson, S. *Modern Approaches to Chemical Reaction Searching*; Willett, P., Ed.; Gower Press: Aldershot, 1986; p 165.
(2) Hicks, M. G. Reactions in the Beilstein Information System: Nonaporic Organic Synthesis. *J. Chem. Inf. Comput. Sci.* **1990**, 30, 352–359.
(3) See papers in this issue.
(4) Luckenbach, R.; Sunkel, J. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 271.
(5) The currently available Beilstein data covers the literature from 1790 to 1980. Plans for the decade 1980–1990 are well advanced.
(6) Vladutz, G. Do We Still Need a Classification of Reactions? In *Modern Approaches to Chemical Reaction Searching*; Willett, P., Ed.; Gower: Aldershot, 1986; pp 202–220.
(7) 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.
(8) Blake, J. E.; Dana, R. C. CASREACT: More than a Million Reactions. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 394–399.
(9) Jochum, C.; Gasteiger, J.; Ugi, I.; Dugundji, J. The Principle of Minimum Chemical Distance and the Principle of Minimum Structure Change. *Z. Naturforsch.* **1982**, *37B*, 1205–1215.
(10) Lawson, A. J. Chemical Structure Browsing. In *Chemical Structure Information Systems: Interfaces, Communication, and Standards*; Warr, W. A., Ed.; ACS Symposium Series No. 400; American Chemical Society: Washington, DC, 1989; pp 41–49.
(11) Lawson, A. J. In *Software-Entwicklung in der Chemie 2*; Gasteiger, J., Ed.; Springer Verlag: Heidelberg, 1988; p 1.
(12) Gassmann, P. G.; Richmond, E. D. *J. Org. Chem.* **1966**, *31*, 2355.