DATABASE COVERING THE SYNTHESES OF PYRIDINES FROM ACYCLIC REAGENTS: THE ROLE OF NONSTANDARD RETRIEVAL REQUESTS IN DEVELOPING THE STRATEGY OF COMPUTER SYNTHESIS

E. V. Babaev and S. V. Tsitovskii

The methodology of constructing a database covering the syntheses of pyridine ring in a standard ChemBase shell is discussed. An algorithm is proposed for introducing into the database nonstandard information on the electrophilic and nucleophilic nature of the reagents and on the structural image of the reaction (reaction graph). Subsequent retrieval requests for the introduced data proved to be heuristic for planning the synthesis of pyridines from the reagents of a prescribed polarity.

This paper is a continuation of our investigation in the methodology of computer synthesis and design of novel heterocyclic structures and reactions, carried out at the Chemical Department of Moscow State University [1–3]. Our earlier publications treated the development of formal models for describing heterocyclic rearrangements [1, 2]. The development of this approach led to the creation of a computer program for predicting new transformations of the cycles (or recyclizations) of heterocyclic systems [3].

Adequate utilization of the already available information on rearrangements offered a clue to the development of a theoretical model and to the prediction of new reactions. To this end, a possibility was provided in the computer program for creating an elementary database related to the already existing reactions, calling which enabled one, in particular, to make conclusions concerning the degree of novelty of the structural transformation being predicted.

In our present paper we consider further generalization of this methodology to heterocyclization reactions. In the first place, we discuss the question, which formal models should be used for describing heterocyclizations. In the second place, we discuss the criteria used by us in creating the database covering the syntheses of pyridines from acyclic structures. In the third place, we would like to demonstrate in what way a small database covering the known methods of the synthesis of a particular heterocycle can help in the development of a general methodology of computer synthesis of unprecedented examples of chemical transformations.

We consider separately the general criteria to be met by such databases: in particular, we illustrate how one can learn to derive nonstandard heuristic hints and rules for a more general problem of planning the synthesis of heterocycles, by using the most standard database shells, such as the ChemBase program, very popular with the chemists.

ABSTRACT SCHEMES OF THE SYNTHESIS OF HETEROCYCLES

Let us consider in brief the principles of formalization and classification of chemical reactions [4–6]. The main idea of the majority of approaches consists in performing mentally a superposition (identification) of the atoms and bonds of the starting reagents and final products. The "reaction image" thus obtained contains information on the location of the bonds being broken and formed. Consequently, one can neglect all the atoms whose environment has not changed and all the bonds whose multiplicity has remained unchanged.

The resulting object can be specified in the form of a reaction graph, if the left- and right-hand sides were represented by conventional structural formulas (molecular graphs). Another method is to describe the chemical transformation in the form of a reaction matrix (BE-matrix [Dugunji]), if the structure of the reagents and products is specified, for example, in the form of a bond matrix.

The resulting mathematical image of the reaction, viz., a graph or a matrix, shows the changes that occurred in the nearest environment of the atoms and in the bond orders in the course of the reaction (Fujita [7] proposed a very appropriate term for denoting such objects: "imaginary transient reaction state"). Such objects can be introduced into the computer, e.g., for searching analogies between different reactions or for predicting the reagents and final products.

This idea was adapted in previous publications [1, 2] to the needs of heterocyclic chemistry, namely, for describing recyclizations of heterocycles. The specific feature of the approach was that the most simplified image of the reaction (recyclization graph) contained, nevertheless, skeletons of the starting and final cycles. Such "incomplete" abstraction had simplified radically a very complicated problem of describing recyclizations and led to the creation of an hierarchical (and periodic) classification of heterocyclic rearrangements, which was used later [3] for computer prediction of new transformations of the heterocycles.

A natural generalization of this approach could be an analogous simulation of heterocyclizations, i.e., of the methods of assembling the heterocyclic ring from acyclic precursors. Fortunately, chemists developed an adequate formal image for heterocyclizations very long ago. Any review in which the syntheses of an arbitrary heterocycle are generalized (see, for example, Ref. [8]) demonstrates how this is done. Usually the structure of the target heterocycle is represented, in which the bonds forming in the course of the reaction are indicated with dotted lines.

Such objects (referred to in the literature as disconnection schemes) are neither the starting reagents nor the final cyclic product. This is a mapping between them, i.e., a reaction graph in the above-mentioned sense. The natural combinatorial character of this mapping is apparent: the number of ways to arrange the dotted lines in the cycle according to such scheme fully corresponds to the number of conceivable ways in which the target heterocycle can be assembled. (In our subsequent exposition we call such mappings synthesis schemes.)

The attention of chemists was attracted repeatedly to the combinatorial nature of the synthesis schemes. In some studies (see, for instance, review [6]) a detailed analysis of the "thought-up" synthesis scheme led to the experimental finding of a heretofore unknown approach to the construction of a heterocyclic ring. The heretofore unknown syntheses of the ring of indazole and thiazole were predicted and discovered exactly in such a manner "at the tip of the pen" [6]. Meanwhile, the very course of reasoning and the logics of passing over (from a very abstract scheme to quite concrete reagents for the heterocyclic synthesis) virtually never were the subject of an independent analysis on the part of specialists in computer chemistry.

Although in the methodology of computer synthesis there exist several approaches to the design of particular reagents for the synthesis of target molecules [6], their application often involves difficulties in connection with the effect of "combinatorial explosion" and with the appearance of a great number of precursors having no prospects, viz., noise structures. Solution of problems of such kind could be found in developing effective criteria for selection of acceptable reagents. This, in turn, requires a careful selection of the initial information as to which syntheses of a given heterocycle can or cannot be effected readily and selectively.

We would like to note that this particular aspect of the problem, an analysis of "good" and "poor" syntheses, was developed in most detail by experimental chemists, for instance, in reviews dealing with the syntheses of one or other heterocycle. Usually, a circumstantial analysis is given for the same synthesis scheme, as to which particular reagents are suitable or unsuitable for a given scheme (see, for example, Ref. [8]). Thus, there is a discrepancy between the real knowledge the experimenters have concerning the regularities of one or another of the synthesis schemes and the inability of theorists in the computer synthesis to put this knowledge to advantage.

THREE CLASSES OF REAGENTS FOR HETEROCYCLIZATIONS

The overwhelming majority of heterocyclizations are heterolytic processes, in which C—C, C—X, or X—Y bonds (X and Y being heteroatoms) are formed in heterophilic or nucleophilic addition or substitution. One of the most important ideas, which takes into account the dominance of heterolytic processes in the syntheses, as well as simplifies and systematizes the types of acyclic structures for the synthesis of heterocycles, is the division of the reagents into three groups [8, 9]: bioelectrophiles, bionucleophiles, and bifunctional compounds (i.e., those containing an electrophilic center and a nucleophilic center at the ends of the chain).

The only fundamental difference of the reagents within each of the three classes is the size of the chain N between polar functions. (The chain introduced by the reagent into the cycle is implied here.) For example, among the reagents with two electrophilic functions at the ends of the chain we can distinguish bioelectrophiles of the following subclasses:

(1,1) — for example, geminal dichlorides, aldehydes, ketones, and derivatives of carboxylic acids;

(1,2) — for example, dichloroethane, phenacylbromide, and glyoxal;

(1,3) — 1,3-dicarbonyl compounds, etc.;

(1,N) — bioelectrophiles with a large length of the chain.

Binucleophiles or bifunctional reagents can be classified in a similar manner [8, 9].

It was of reasonable interest to find out which particular polar types of reagents dominate in the synthesis of one or another heterocycle. (It is common knowledge, for example, that α-haloketones are very often used in the synthesis of azoles, but they are seldom used in the preparation of azines [8].) An answer to this question could provide a clue to the computer retrosynthesis of heterocycles, namely, to the transition from abstract synthesis schemes to particular classes of reagents. Such an answer should be sought, for instance, in a comparison of the polar nature of the reagents used in the synthesis of one or the other of heterocyclic structures.

COMPUTER DATABASE FOR THE SYNTHESES OF PYRIDINES

We believed that any "good" rules of the synthesis of heterocyclic structures can be developed only on the basis of analysis of a concrete overview of the syntheses of heterocycles. Such rules can be further refined by comparing their results for different heterocycles.

To this end, we deemed it reasonable to introduce the initial information on the syntheses of particular heterocycles into the computer database (hereafter referred to as DB) for subsequent statistical analysis.

Choice of objects. Pyridine ring was chosen as the simplest heterocyclic system; it was contemplated that the methods of synthesizing this ring should then be compared with those of synthesizing the closest allied structures (for instance, benzo- aza-, and hetero analogs). Maximum-unsaturated structures were primarily selected (pyridines proper, pyridones, etc.).

Sources of DB formation. To attain the highest representativity and to cover the literature of different periods, the main reviews and monographs dealing with the pyridine chemistry [8, 9–15] were chosen as the primary sources. In the course of gradual data collection recent abstract journals were analysed.

Choice of DB shell. For many reasons the ChemBase system was chosen as the standard shell. On the one hand, this program has become almost the most widespread among the chemists. On the other hand, it has extensive potentialities for graphic input and processing of information not only on the structures, but on the reactions as well. Finally, we were interested to find out whether a standard database can be "compelled" to answer nonstandard requests, for example, concerning the polar nature of the reagents used in the synthesis. In addition to the standard (internal) format for storing data on the reactions, the program allows the data to be transformed into formats readable by other computer programs. This feature was supposed to be employed later on.

Choice of reactions. The reactions were chosen in accordance with the following criteria:

Limitation to heterocyclizations. The methods of synthesis comprising rearrangements or recyclizations were left out of consideration.

Limitations imposed on the heterolytic mechanism. The requirement of unambiguous description of heterocyclizations in terms of the interaction of electrophilic and nucleophilic centers resulted in leaving some pericyclic reactions out of account.

Selectivity. Reactions must yield predominantly one product.

Laboratory reaction conditions. Some vapor-phase industrial syntheses of pyridines were not analyzed.

Arbitrary number of components for the synthesis. In this case, the database covered most examples of named reactions [16].

The way information was given. For each synthesis one entry with the following fields was allocated in the database:

Graphic part contained structural formulas of the starting heterocyclic systems.

Text part consisted of 11 fields, including, in particular, identification number, authors of publication, review source, imprint of original paper, name of synthesis (if reaction was personal), and comments field (conditions of synthesis, reagents, and yield).

STANDARD REQUESTS ON THE SYNTHESES OF PYRIDINES STIPULATED IN DB

The created database makes it possible to search for the starting reagent or product at User's option, by entering the results into a separate list. A similar search in text fields is possible (by the names of authors, years of publications, or, e.g., by named reactions). The attributes according to which the search is effected can be combined into logical expressions. For example, there is no difficulty in obtaining in a few seconds the information on the use of a given compound in the synthesis of a given pyridine derivative during the period of the last 20 years.

An additional possibility offered by the DB is the search by the substructure. (For this to be done, it is necessary to specify an arbitrary fragment of the molecule, for instance, a functional group or an ensemble of carbon atoms joined by the bonds of a specified multiplicity.) Furthermore, several such fragments can be combined into logical expressions. Note that already in such standard form the number of examples entered into the DB generates a new property: if the structural scheme of the synthesis of a desired structure is unknown, it is possible to request information on analogous syntheses. As a result, the created DB is of a definite methodological value: it facilitates the initial acquaintance (say, of students) with the diversity of synthetic approaches in the chemistry of heterocycles.

METHODS OF DB ADAPTATION FOR RESPONSES TO NONSTANDARD REQUESTS CONCERNING THE POLAR STRUCTURE OF REAGENTS

We pursued the objective of elucidating the polar nature of the reagents most often used for the syntheses of the pyridine ring. It should be stressed that this type of information is not contained explicitly in the structural formulas of the reagents: a certain mental effort is required for assigning a definite polar type to a given reagent in a given reaction. Naturally, the ChemBase program at the very start seemingly could not be adapted for solving such problems. Thus, we had the choice: either to create a new shell of graphic input of structures, or to adapt the already existing program with the vast molecular graphic potential for our purposes.

In this connection, we resorted to the option provided in the ChemBase program, of introducing the structure of intermediates into the DB. These intermediates can serve later for carrying out search, analogous to that by the reagents or products. We believed that the polar type of the reagent can be represented just as an imaginary intermediate. In other words, we thought it reasonable to "mask" an abstract synthesis scheme as an intermediate, having indicated in the scheme in some way the electrophilic-nucleophilic nature of the centers forming the bonds of the cycle.

It turned out to be far from trivial to solve such a problem. On the one hand, the program assigns automatically the symbol of an absolutely concrete element to any atom. On the other hand, an abstract synthesis scheme is, generally speaking, a connected signed graph, having two "colors" of the edges: conventional (skeleton bonds of the reagent, that have passed into the cycle) and dotted (cyclic skeleton bonds that have originated in the course of the reaction). Consequently, a distinction had to be made between the dotted and the conventional bonds, which in the ChemBase system are regarded as identical. Finally, the program was not sensitive to distinguish any marks on the atoms (even if they were envisaged in the graphic input of the symbols, e.g., charges, isotope labels, etc.) and treated them as identical during search.

The only feasible solution to the problem how to "deceive" the program proved to be the following "know-how" algorithm:

Standard synthesis schemes are represented as imaginary intermediates, which the program can compare with one another, irrespective of the reagents or products.

The skeleton bonds of the reagent (that pass without changes into the cycle) are represented in the synthesis scheme by double bonds, whereas the bonds being formed are represented by ordinary bonds. In this case the program, distinguishing between the ordinary and double bonds, carries out search by the

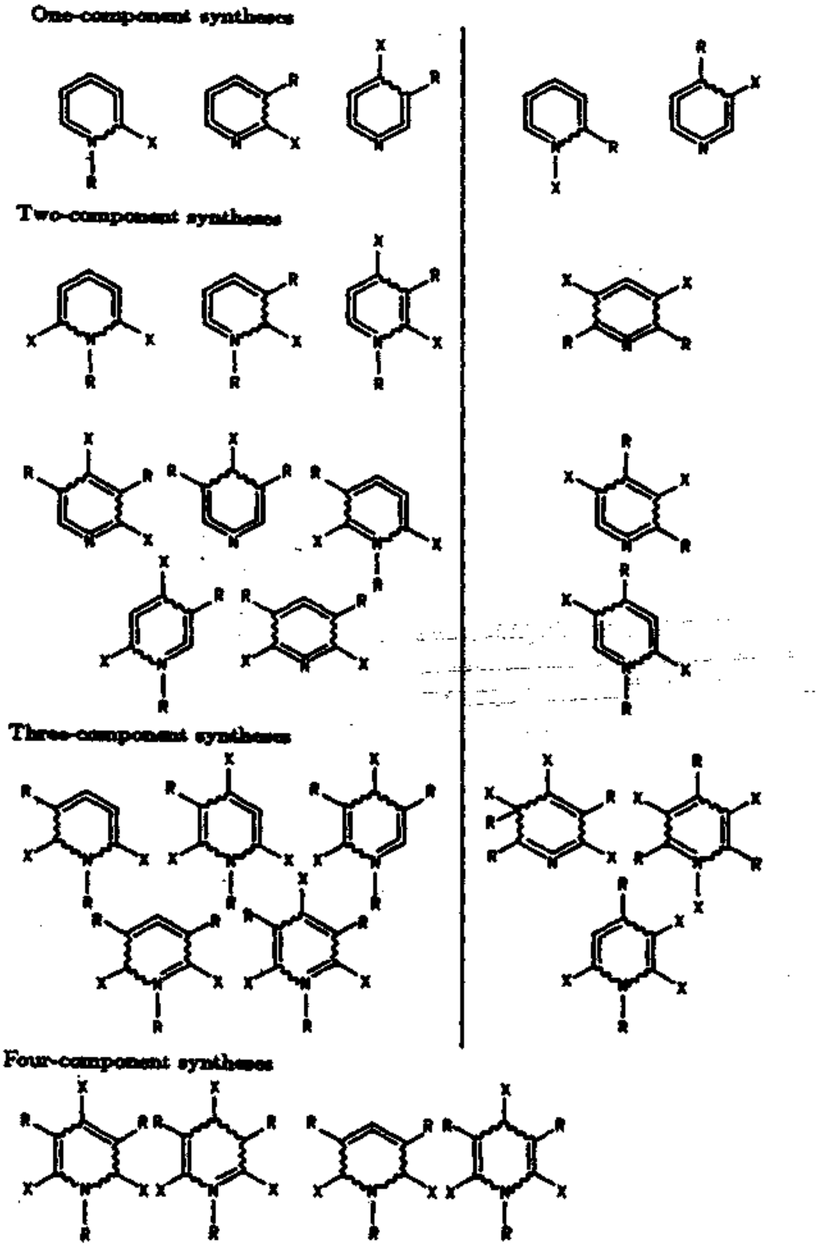

Fig. 1

The known structural types of pyridine ring syntheses contained in the database. Double lines in the diagrams correspond to the skeleton bonds of the reagents that pass into the skeleton of the pyridine cycle. The wavy line indicates the cycle bonds being formed. Symbols R and X correspond to the nature of reactive centers in the reagent (R denotes nucleophilicity of the neighboring atom; X denotes its electrophilicity). The line separates the most typical syntheses (left) from the extremely rare ones (right).

structure of the intermediate as a whole, i.e., actually by the synthesis schemes. For the User's convenience, ordinary bonds are masked by a wavy line recognized by the computer as a conventional ordinary bond.

The electrophilic and nucleophilic ends of the reagent are expressed by introducing "phantom-groups" of two types, X and R, respectively, into the synthesis scheme. (The symbols X and R are the only notations of the functional groups, reserved by the program, which are distinguishable during search.)

Typical examples of synthesis schemes introduced by the User in such a way are presented in Fig. 1.

SOME RESULTS OF ANALYSIS OF DB RESPONSES TO NONSTANDARD REQUESTS

At present the DB is at the stage of formation. Nevertheless, an analysis of several hundred reactions introduced into the database has led us to conclusions that surpass to some extent our initial expectations. First of all, we tried to introduce a maximum number of nonequivalent synthesis schemes into the DB. For this to be done, the structure of an "intermediate" being introduced into the DB is compared to other structures, already present in the computer memory. Only if such a scheme is absent in the DB, a decision is made of the priority of introducing the chemical equation as such. Thus, Fig. 1 illustrates all the nonequivalent schemes of the synthesis of pyridines (known to the authors) with indication of the polar nature of the reagents.

The vertical line in Fig. 1 corresponds to the preliminary statistics of the distribution of the syntheses of pyridines into classes: the syntheses for which hundreds or tens of examples are available are disposed in the left-hand part of Fig. 1; the syntheses whose examples are unique are disposed in the right-hand part of Fig. 1.

Analyzing the reagents in the left-hand part of Fig. 1, it is easy to see that a clear determinism of preferable disposition of polar functions in the reagents is displayed. For example, it is required that the nucleophilic centers should be located in those places that correspond either to the β-positions of the pyridine cycle being formed, or to the pyridine nitrogen atom as such. On the contrary, the electrophilic centers are fixed in those places of the chain of the reagents that will pass into the α- or γ-positions of the pyridine cycle. We would like to note that in the 99% of cases this rule is observed strictly for one- and two-component syntheses of maximum-unsaturated pyridines. Deviations are observed in individual examples of multicomponent syntheses, and also in preparing hydrogenated structures. Thus, a clear-cut heretofore unknown relation of the "structure–synthesis" type manifests itself. This relation can be used in the computer design of new heterocyclisations. A detailed analysis of this relation was the subject of a separate communication [17]. Here we have demonstrated the way leading to this relation.

The opportunity of using the ChemBase program (Version 1.50, Molecular Design Ltd., San Leandro) was kindly given to us by A. V. Cheprakov, Moscow State University (Disk Serial BAI 2150271). The present investigation was included into the curricula of the Chair of Organic Chemistry (Graduation Thesis of S. V. Tsitovskii; June, 1993).

REFERENCES

1. E. V. Babaev and N. S. Zefirov, *Khim. Geterotsikl. Soedin.*, no. 6, p. 808, 1992.
2. E. V. Babaev and N. S. Zefirov, *Bull. Soc. Chim. Belg.*, vol. 101, no. 1, p. 67, 1992.
3. E. V. Babaev, D. E. Lushnikov, and N. S. Zefirov, *J. Am. Chem. Soc.*, vol. 115, p. 2416, 1993.
4. *Chemical Information Systems: Beyond the Structure Diagram*, Chichester, 1990.
5. R. Herges, *Tetrahedron Comput. Methodol.*, no. 1, p. 15, 1988.
6. P. Barone and M. Chanon, *Computer Aids to Chemistry*, Chichester, 1986.
7. J. Fujita, *J. Chem. Inf. Comput. Sci.*, vol. 28, p. 128, 1988.
8. *Comprehensive Heterocyclic Chemistry*, vols. 1–8, Oxford, 1984.
9. A. R. Katritzky, *Handbook of Heterocyclic Chemistry*, p. 382, Oxford, 1985.
10. H. Maier-Bode and J. Altpeter, *Das Pyridine und seine Derivate in Wissenschaft und Technik*, Halle, 1934.
11. Yu. I. Chumakov, *Pyridine Bases* (in Russian), Kiev, 1965.
12. K. Schofield, *Heteroaromatic Nitrogen Compounds. Pyrroles and Pyridines*, London, 1967.
13. F. Brody and R. Ruby, *Pyridine and Its Derivatives*, Part 1, p. 99, New York, 1960.
14. N. S. Boodman, J. O. Hawthorne, P. X. Maskiantonio, and A. W. Simon, *Pyridine and Its Derivatives*, Suppl. 1, p. 185, New York, 1974.
15. *Pyridine and Its Derivatives*, Suppl. 5, New York, 1984.
16. K. V. Vatsuro and G. A. Mishchenko, *Named Reactions in Organic Chemistry* (in Russian), Moscow, 1976.
17. E. V. Babaev, *Khim. Geterotsikl. Soedin.*, no. 7, p. 937, 1993.

14 September 1993

Chair of Organic Chemistry