ACKNOWLEDGMENT

Financial support for this work from the Center for Scientific Databases, Academia Sinica, is gratefully acknowledged.

REFERENCES AND NOTES

(1) Gschneider, K. A., Jr.; Eyring, L. *Handbook on Physics and Chemistry of the Rare Earths*; North-Holland Publishing Co.: Amsterdam, 1979; Vol. 3. *Ibid.* 1982; Vol. 5.
(2) Marcus, Y.; Kertes, A. S.; Yanir, E. *Equilibrium Constants of Liquid-Liquid Distribution Reactions*; Pages Bros., (Norwich) Ltd.: Norwich, 1974.
(3) *Handbook of Extraction.* Rosen, A. M., Ed.; translated by Yuan Chengye; Atomic Energy Press: Beijing, 1981; Vol. 1. *Ibid.* 1983; Vol. 2. *Ibid.* 1988; Vol. 3.
(4) Zhong Shan University. *Chemicophysical Constants of Rare Earths*; Metallurgical Industry Press: Beijing, 1978.
(5) Mayer, I. P.; et al. *J. Phys. Chem.* **1962**, *66*, 693.
(6) Yang, Jiaan; Jiang, Yuansheng. The Graphic Character and Thermodynamic Properties of Aliphatic Alkanes. *Acta Chim. Sin.* **1983**, *41*, 884.
(7) Wang, Huayun; Lü, Tianxung; Xu, Lu; Wang, Erkang; Su, Qang. General a_N Index and its Applications. I. Extraction Reactivity of Y, Ce, U and Th. *Acta Chim. Sin.* **1990**, *48*, 1159.
(8) Wang, Huayun; Xu, Lu; Su, Qang. General a_N Index and its Applications. II. Properties of Phosphorus Compounds. *Acta Chim. Sin.* **1991**, in press.
(9) Kubaschewski, O.; Alook, A. B. *Metallurgical Thermochemistry.* Butterworth-Spring, Pergamon: Oxford, 1979.
(10) Chen, Nianyi; Xu, Zhihong; Liu, Hongling; Hu, Hua; Wang, Leshan. *Computational Chemistry and Application.* Shanghai Sciences Press: Shangchai, 1987.
(11) Jurs, P. C.; Kowalski, B. R.; Isenhour, T. L. Computerized Learning Machine Applications to Chemical Problems. *Anal. Chem.* **1969**, *41*, 21.
(12) Wold, S.; Sjostrom. SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy. *ACS Symp. Ser.* **1977**, *No. 52*, 243.
(13) Xiao, Yunde; Xu, Lu. Search for the Basic Regularities of the Transition Emission of Eu(II) Ion Complex Fluorides with Pattern Recognition. *Chemom. Intell. Lab. Syst.* **1991**, in press.

MAPOS: A Computer Program for Organic Synthesis Design Based on Synthon Model of Organic Chemistry

LUDĚK MATYSKA* and JAROSLAV KOČA†

Institute of Computer Science and Department of Organic Chemistry, Faculty of Science, Masaryk University, 611 37 Brno, Czechoslovakia

Received October 29, 1990

The program MAPOS is a logically oriented computer program for computer-aided organic synthesis design suitable both for forward and retrosynthetic synthesis planning. It is based on the synthon model of organic chemistry, introduced by the authors. The fundamentals of the model as well as the basic algorithms are described. Examples of the use of the program are given.

1. INTRODUCTION

Computer-aided organic synthesis design (CAOS) is currently an active field of computer chemistry. Programs for CAOS may be classified in two basic directions—the information-oriented programs and those logically oriented.

The former are based on a database of chemical reactions. They may be specialized for some area of organic synthesis (e.g., synthesis of heterocycles), but there exist attempts to cover organic chemistry as a whole.[1-7]

The logically oriented programs need a formal model of organic chemistry.[8,15] This formal (mathematical) model is usually based on discrete mathematics, and it is not directly related to quantum chemistry models. However, energy computational methods, even those based on quantum mechanistics, often serve as bases of the strong heuristics used. Several such programs have been noted in the literature;[9-13,15] many of them are based on the Dugundji-Ugi model of constitutional chemistry.[8]

Obviously, some CAOS computer programs do not fall strictly into any of the two mentioned groups. The empirical chemical knowledge on different levels together with a mechanistic approach is usually used in them.[14,16,17]

The program MAPOS, presented in this paper, belongs to the logically oriented programs. It is based on the synthon model of realistic constitutional chemistry.[18,19] In this model, the central role is played by the so-called valence states of atoms, the notation almost identical with Pauling's[20] and Van Vleck's[21] idea of "atom in molecule". The model incorporates the notion of the reaction center, mostly represented by one (addition and elimination reactions) or two (substitution reactions) atoms where a primary attack occurs. Only changes initiated by the primary attack are expanded to other atoms of the molecule.

The combinatorial nature of this mathematical model is further restricted by the use of heuristics based on the reaction distance,[18,19,22,23] defined as the minimal number of elementary steps of valence electrons reorganizations (ESRE).[8,24-28] Reaction distance is a notion very similar to chemical distance.[29] However, no one of these metrics may be directly transformed to the other; they represent different aspects of the chemical reality.[22]

A similar approach, based directly on the Dugundji-Ugi work, is used in the IGOR system.[30]

2. MATHEMATICAL MODEL

The synthon is the basic structural unit used in the model. The notion of synthon has been, in the frame of organic chemistry, initially introduced by Corey.[31] Based on Corey's approach, the formal synthon model of organic chemistry has been introduced.[18,19,22] It may be understood as a generalized concept of valence states of atoms and their combinations. From all used meanings of the notion "synthon", the meaning of "substructure reduced in the reaction" is the most closest to our formal definition.

The synthon $S(A)$ over an (arbitrary) atomic set A is defined as one or several molecules and/or their parts, composed of atoms from the set A (all atoms must be used). The most similar notion, introduced so-far in the scope of CAOS, is the definition of Ensemble of Molecules EM(A).[8] The notion of synthon is more general because free valences, i.e., bonds that do not connect two atoms from A but only start from one atom, may be also specified.

Being a generalization of the concept of valence states of atoms, the synthon carries also a strong mechanistic content—conversions of valence states of atoms allow clear

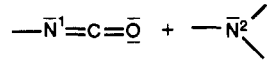
0095-2338/91/1631-0380$02.50/0 © 1991 American Chemical Society

and concise modeling of reaction mechanisms.[27,32]

From the formal point of view, a synthon may be modeled by the synthon matrix (S-matrix) or by the synthon graph (S-graph).[18,19,22] Throughout this paper, the S-matrix approach will be used.

2.1. S-Matrix. The S-matrix is the topological model of the synthon because it does not explicitly model stereochemistry. It is a special symmetric square matrix with nonnegative integer off-diagonal entries and four-component vectors on the main diagonal. Individual entries of these vectors are nonnegative integers. Each column/row of the matrix corresponds exactly to one atom of the synthon. The multiplicity of chemical bonds between individual atoms of the synthon is represented by off-diagonal entries of the matrix. The off-diagonal part of the S-matrix is an analogy of the so-called BE matrix introduced by Dugundji and Ugi.[9] The main diagonal vectors describe the valence states[25,32] of individual atoms. The first number is the number of lone electrons on the atom; the second, third, and fourth entries represent the number of single, double, and triple bonds, respectively. Free valences of the atom are understood as bonds to the so-called *virtual atoms*.

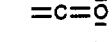
Example 2.1. Let us consider a synthon



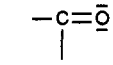
constructed on the set $A = \{C, O, N^1, N^2\}$ of atoms. Its S-matrix is

$$\begin{array}{c} \\ C \\ O \\ N^1 \\ N^2 \end{array} \begin{array}{c} \begin{array}{cccc} C & O & N^1 & N^1 \end{array} \\ \begin{pmatrix} (0,0,2,0) & 2 & 2 & 0 \\ 2 & (4,0,1,0) & 0 & 0 \\ 2 & 0 & (2,1,1,0) & 0 \\ 0 & 0 & 2 & (2,3,0,0) \end{pmatrix} \end{array}$$

A *subsynthon*[18,19] of a given synthon $S(A)$ is any synthon $S(X)$ formed on the atomic set X such that X is a subset of the set A, and the S-matrix of $S(X)$ is a submatrix of the S-matrix of $S(A)$. It means that all atoms from the set X have to retain their valence states in $S(X)$ with respect to $S(A)$. Thus, the synthon



is a subsynthon of the synthon $S(A)$ from example 1, while the synthon



is not, as the valence state of carbon is not preserved. The sybsynthon is a generalization of the well-known and widely used term "substructure" and has the same intuitive meaning.

2.2. SR-Matrix. The S-matrix is a static description of a chemical system. In order to model chemical reaction, the so-called *SR-matrix*,[18,19,22] R is introduced by the equation

$$R = P - E \qquad (1)$$

where E and P are the S-matrix of educt and product, respectively.

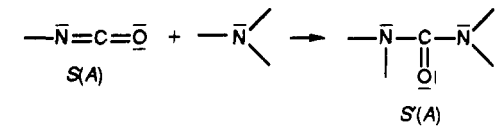
The matrix R defined by eq 1 is of the same kind as the matrices P and E. The operation "−" in eq 1 is understood as the standard matrix subtraction for off-diagonal elements, and vector subtraction for the diagonal ones.

In analogy to the Dugundji and Ugi approach,[8] eq 1 may be expressed as

$$E + R = P \qquad (2)$$

modeling thus chemical reaction.

Example 2.2. Let us consider the reaction of the synthon from the example 2.1:



which can be seen as the addition of ammonia to isocyanate. Its SR-matrix may be written as

$$R = \begin{array}{c} \\ C \\ O \\ N^1 \\ N^2 \end{array} \begin{array}{c} \begin{array}{cccc} C & O & N^1 & N^2 \end{array} \\ \begin{pmatrix} (0,2,1,0) & 2 & 1 & 1 \\ 2 & (4,0,1,0) & 0 & 0 \\ 1 & 0 & (2,3,0,0) & 0 \\ 1 & 0 & 0 & (2,3,0,0) \end{pmatrix} \end{array} -$$

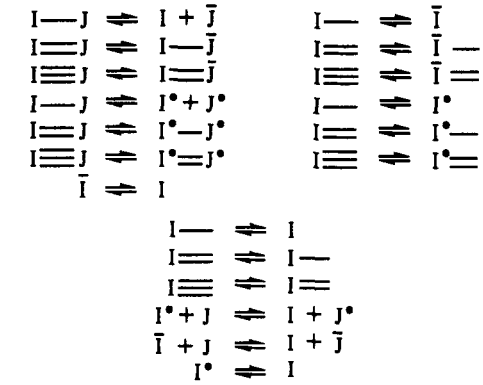
$$\begin{array}{c} \\ C \\ O \\ N^1 \\ N^2 \end{array} \begin{array}{c} \begin{array}{cccc} C & O & N^1 & N^2 \end{array} \\ \begin{pmatrix} (0,0,2,0) & 2 & 2 & 0 \\ 2 & (4,0,1,0) & 0 & 0 \\ 2 & 0 & (2,1,1,0) & 0 \\ 0 & 0 & 0 & (2,3,0,0) \end{pmatrix} \end{array} =$$

$$\begin{array}{c} \\ C \\ O \\ N^1 \\ N^2 \end{array} \begin{array}{c} \begin{array}{cccc} C & O & N^1 & N^2 \end{array} \\ \begin{pmatrix} (0,2,-1,0) & 0 & -1 & 1 \\ 0 & (0,0,0,0) & 0 & 0 \\ -1 & 0 & (0,-2,1,0) & 0 \\ 1 & 0 & 0 & (0,0,0,0) \end{pmatrix} \end{array}$$

The off-diagonal entries of the SR-matrix describe changes in the multiplicity of chemical bonds. Changes of valence states of atoms during the reaction are expressed by the diagonal entries of the SR-matrix. The diagonal entries of the SR-matrix may be used for the distinguishing between different kinds of chemical reactions modeled, i.e., additions, substitutions, eliminations, or their combinations may be recognized, reflecting thus the mechanistic contents of the synthon model.

2.3. Isomeric Synthons. Any two synthons $S(A)$ or $S'(A)$ constructed over the atomic set A are isomeric. The synthon isomerism is the generalization of the chemical isomerism. All synthons, constructable from the set A, belong to the *Family of Isomeric Synthons*, FIS(A),[18,19,22] an analogue to the *Family of Isomeric Ensembles of Molecules*, FIEM(A).[8] The synthons $S(A)$ and $S'(A)$ from example 2.2 are isomeric.

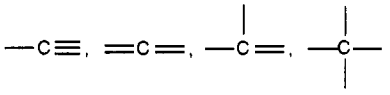
For each couple of synthons S(A) and S'(A) and FIS(A), there exists one SR-matrix describing the isomerization $S(A) \rightarrow S'(A)$. Such a matrix is defined by eq 2, and it can be decomposed into a finite sum of the so-called elementary SR-matrices.[18] These matrices express the following elementary electronic processes:



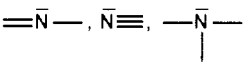
The smallest number of elementary SR-matrices from which the final SR-matrix can be composed is called reaction distance[18,19,22,23] RD[$S(A)$,$S'(A)$] between the synthons $S(A)$ and $S'(A)$. Chemically interpreted, the reaction distance between educt and product is the smallest number of elementary electronic processes on the path *educt* → *product* (or in the opposite direction).

2.4. Stable Synthons. The set of "*stable*" valence states, corresponding to states in observable products or stable intermediates, is joined with each chemical element. A synthon is called stable if the valence state of each atom of the synthon belongs to the appropriate set of stable valence states. Although the stable synthon does not need to be a stable chemical structure, this notion helps further reduce the combinatorial nature of the model. The valence states without charge are usually called stable.

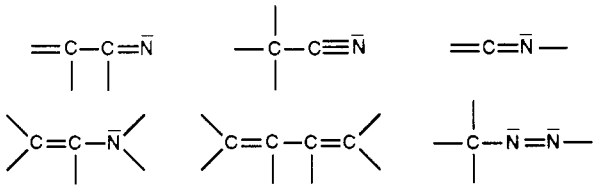
Example 2.3. Let the valence states



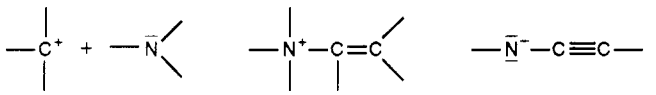
of carbon, and the valence states



of nitrogen are considered as stable. Then the following synthons



will be declared as stable. At the same time, the synthons



will be classified as unstable.

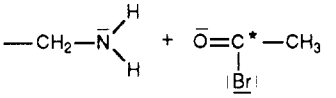
2.5. Synthon Precursor and Successor. Generation of the Synthon Precursors/Successors,[18,19] (SPS) is the fundamental task for most of the CAOS programs. Intuitively, SPS of a synthon $S(A)$ should be any stable synthon $S'(A)$ that is relatively close (from some distance point of view) to the $S(A)$. However, the above definition is too wide to be of direct use. Therefore, the notion of SPS is further specified, and the idea of reaction center is used. The reaction center is a subsynthon $S(X)$ of $S(A)$ on which the reaction starts or which undergoes the most fundamental change during the reaction (the reaction center may be understood as the "synthon" in its commonly used meaning[31]). The reaction center is usually chosen to be the most reactive part of the molecule for synthesis in the forward direction, and the skeletal part or functional group which may be synthesized in the last step for the retrosynthesis. First, let us turn our attention to the reaction center changes.

Let $S(A)$ be the starting synthon and $S'(A)$ be the SPS of $S(A)$. The subsynthon $S(X)$ of $S(A)$ is its reaction center for the isomerization $S(A) \rightarrow S'(A)$ if the following conditions are fulfilled by the change $S(X) \rightarrow S'(X)$:

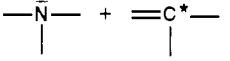
1. The reaction distance of $S(X)$ and $S'(X)$ is non-zero, i.e., at least one electronic process has to take place on the reaction center.
2. The electronic change has to be "contiguous", i.e., two or more mutually independent reactions in one planed step are forbidden.
3. The upper bound of the number of elementary electronic processes taking place on one atom is two. At the same time, the number of new σ-bonds, created to virtual atoms, is limited by two.

The conditions for reaction center change have been introduced in a more detailed manner recently.[18,19,22] Below, all these conditions will be considered as one condition, marked with an asterisk (*).[36] The meaning of the (*) condition will be illustrated by the following example.

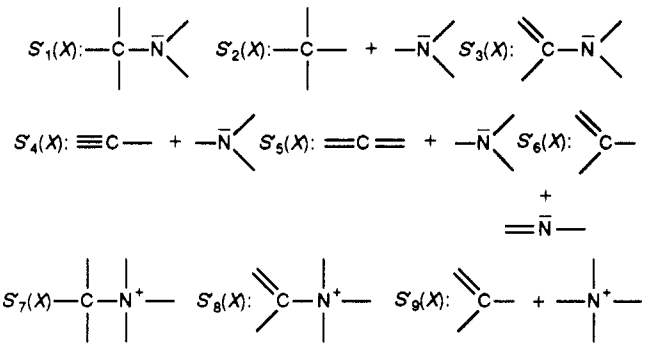
Example 2.4. Let us consider the synthon



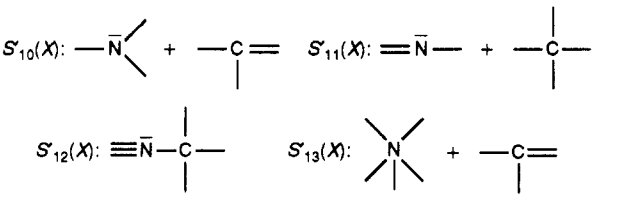
Let us suppose the reaction center set $X = \{C^*,N\}$. It implies that $S(X)$ is



(a) The subsynthon $S(X)$ may be changed to the following synthons:



(b) The changes that lead to the following synthons are not accepted:



Condition 1 is conflicted by the change $S(X) \rightarrow S'_{10}(X)$, condition 2 by the change $S(X) \rightarrow S'_{11}(X)$, and condition 3 by the change $S(X) \rightarrow S'_{12}(X)$ (two new bonds to virtual atoms have been created on the nitrogen atom).

The following condition 4 together with the above three conditions have to be fulfilled by any SPS of $S(A)$.

4. The creation of a new σ-bond between an atom a which is not a part of the reaction center $S(X)$ ($a \notin X$) and any atom from the set A is forbidden in one step.

The conditions for the whole synthon $S(A)$ change to any SPS $S'(A)$ of $S(A)$ have been introduced in a more detailed way recently.[18,19,22] Below, they together will be marked with two asterisks (**).

The notion of SPS has been introduced in a close connection with a reaction center $S(X)$. Therefore, we define a SPS of $S(A)$ with respect to a reaction center $S(X)$, and the set of all the SPS of $S(A)$ with respect to $S(X)$ is denoted[18,19] $\mathcal{F}[S(A/X)]$.

The model presented makes it possible to elaborate a part of the starting structure (e.g., a part of skeleton, functional group, etc.) as "rigid", i.e., without any change. This "rigid" part is denoted as $S(\bar{X})$ ($X \cup \bar{X} = \emptyset$). The set of all the SPS of $S(A)$ with respect to the reaction center $S(X)$ and reduced by the subsynthon $S(\bar{X})$ is denoted[18,19] $\mathcal{F}[S(A/X/\bar{X})]$.

3. BASIC ALGORITHMS

The fundamental problem for the program MAPOS, **MA**thematical **P**rogram for **O**rganic **S**ynthesis Design, is the generation of the set $\mathcal{F}[S(A/X/\bar{X})]$. The simplest way is to generate all the isomeric synthons of $S(A)$ and to choose synthons fulfilling the conditions (*) and (**). However, it

is easy to see that this way is, in any but a trivial case, unfeasible because of the possible combinatorial explosion. Therefore, this fundamental problem is divided into two parts. In the first one, only the SPS of the reaction center $S(X)$ are generated by the usual combinatorial algorithm, while all the SPS of $S(A)$ are generated by a different, more efficient algorithm.

3.1. Generation of the SPS of the Reaction Center. The algorithm corresponding to problem of the generation of the SPS of a reaction center $S(X)$ is formulated in three basic steps.

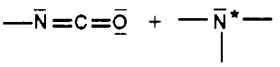
In the first level, all the possible synthons on the set X are generated such that each atom x from the set X is only in one a priori valence state. Accordingly, a subset of FIS(X) will be generated. This level is performed by the algorithm GENFIS1 (see appendix). Note that the problem formulated in such a way is very close to the problem of structure generation described in the literature, cf., for examples, refs 33–35. Since the set X is small in any practical application (usually 1–3 atoms), the combinatorial explosion is negligible in this case.

In the second level, such a subset of FIS(X) is generated where an atom x from set X is in more a priori specified valence states. This level is performed by the algorithm GENFIS (see appendix).

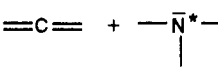
In the last level, synthons $S'(X)$ which fulfill the condition (*) are generated by the algorithm GENFIS2. One from the basic heuristics used by the algorithm GENFIS2 is the condition $RD(S(\{x\}),S'(\{x\})) \leq 2$, which has to be fulfilled for any atom x from set X and the change $S(X) \rightarrow S'(X)$. The algorithm GENFIS2 may be implemented as a straightforward extension of the algorithm GENFIS1, where each synthon generated is filtered out through the condition (*).

The performance of the algorithm GENFIS2 is illustrated by the example in section 2.4, part (a) and by the following example 3.1.

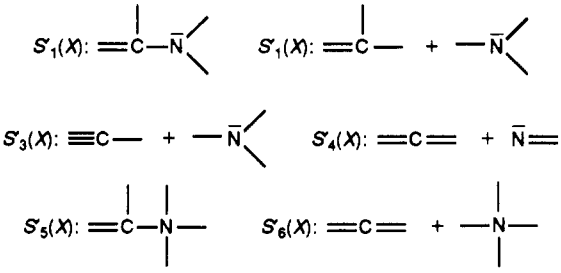
Example 3.1. Let us consider the synthon



from the example 2.1. Let us suppose that the set X of reaction center is $X = \{C,N^*\}$. The corresponding subsynthon $S(X)$ is



The following synthons $S'(X)$ will be obtained by the application of the algorithm GENFIS2:



3.2. Generation of All the SPS of the Starting Synthon. Now, the main problem may be formulated as follows. Let $S(A)$ be a starting synthon on the set A of atoms, let $S(X)$ be the reaction center, $S(X) \subset S(A)$. From the above constructions, we have got a set of all the synthons $S'(X)$ isomeric to $S(X)$ and fulfilling the condition (*). For each synthon $S'(X)$, we should find all the synthons $S'(A)$ isomeric to $S(A)$ such that $S'(X) \subset S'(A)$ and the condition (**) is fulfilled by $S'(A)$. The approach introduced is called stabilization,[18] and it is realized by the algorithm STAB (see appendix).

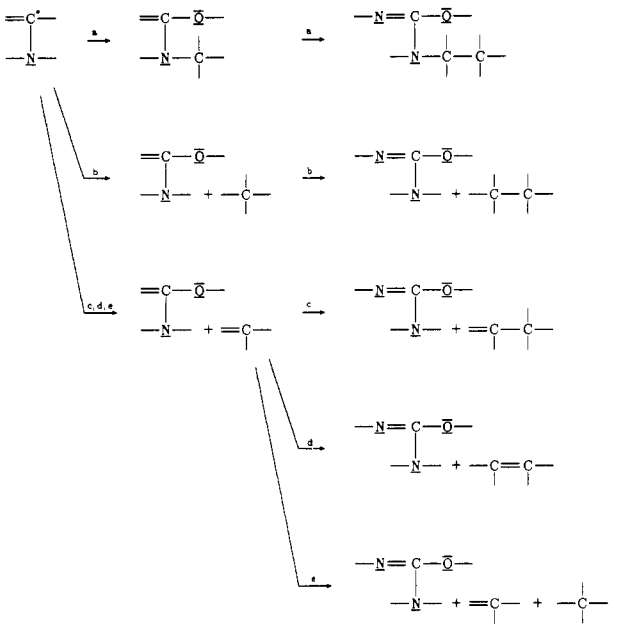
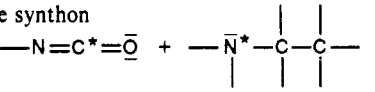
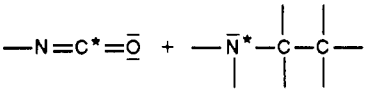


Figure 1. Stabilization of the synthon

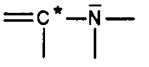


The stabilization process is started at the first neighborhood of the reaction center. It means that new valence states of atoms of the first neighborhood are examined, and those corresponding to the changes realized on the reaction center are found. If all the atoms of the first neighborhood remain without any change, i.e., it is possible to create the same bond, which existed before the change, between each atom from the reaction center to the atoms from the first neighborhood, we say that it is possible to *immerse* $S'(X)$ into $S(A)$. Realizing this immersion, we get the synthon $S'(A)$. If at least one atom of the first neighborhood has to change its valence state as a consequence of the isomerization $S(X) \rightarrow S'(X)$, we join the atoms of the first neighborhood to $S'(X)$, and we get a new synthon denoted as $S'(Y)$ $(Y \supset X)$. Now, the process of stabilization is repeated starting with $S'(Y)$. The whole process is finite because at each step at least one atom is added and the number of atoms in the set A is finite. The stabilization is illustrated by the following example 3.2.

Example 3.2. Let us consider the synthon $S(A)$:

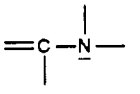


Let the set X of the reaction center be $X = \{N^*,C^*\}$. Let us take the synthon



The synthon is taken from the set of all the synthons fulfilling the condition (*), and the example 3.1 may serve as a source. The process of stabilization is illustrated in Figure 1.

It is seen from the example that the starting synthon for stabilization, i.e., the synthon



[denoted $S'(X)$] cannot be immersed into $S(A)$. Therefore, stabilizations have been performed. One stabilization has been necessary in the processes (a,b,c). Two stabilizations have been

necessary in the processes, (d) and (e). However, the examples (a)–(e) are not an exhaustive description of all the stabilizations of $S'(X)$ with respect to $S(A)$.

4. COMPUTER IMPLEMENTATION OF THE PROGRAM MAPOS AND ITS USE

4.1. Input. The basic input of the program is the starting synthon as well as the reaction center together with the part of the structure to be "rigid". The permitted order[18] of SPS may be input as well. The order of SPS is dependent on the changes of valence states, and it measures skeletal fragmentation. It is illustrated by the examples below. The program "knows" allowed and stable valence states of many atoms, but as the number and quality of generated SPS may be very dramatically affected by the choice of stable valence states, they may be changed by the user. For example, marking the valence state —$\overline{C}$— as stable, the program will produce carben structures.

4.2. Output. The program produces one level of the synthetic/retrosynthetic tree. However, each of the one-level tree synthons may be used as input for the building of further levels of the tree. Each of the synthon produced is numbered and may be drawn on the screen, as well as the whole tree.

4.3. Computer Implementation. The program MAPOS is implemented on an IBM PC-compatible personal microcomputer, running MS DOS 3.30. Its core part, which is based on the algorithms GENFIS2 and STAB, is written in C language, while the user interface and the drawing routines are written in Pascal. The program uses the whole memory (640 Kbytes) addressable by the PC under MS DOS, and it is capable to manipulate, for example, several hundreds of 20 atoms synthons in one (retro) step. There are no explicit limits on the number of atoms in a synthon, on the number of synthons, etc.; the only requirement is that all the structures generated have to be stored in the main memory of the computer. The total number of synthons in the tree is restricted only by the capacity of external storage (hard disc), and the program is well suited for real synthetic problems. The implementation for an UNIX-based computer is in progress (the virtual memory of the UNIX operating system will allow processing transformations of large structures with many atomic reaction centers).

5. EXAMPLES

In order to provide examples of the program use, and in the same time to illustrate its deductive power, we used the program MAPOS on two examples discussed earlier in the literature.[8,13]

The first example is taken from the work[13] where, among others, the synthesis of 3,4-dihydro-1,2,3-oxathiazin-4-one-2,2-dioxine (I) was studied with the aid of computer program TOSCA. The MAPOS program, while analyzing this problem, found not only the halogensulfonyl isocyanate, substituted methylacetone, and acetoxypropene, discussed in ref 13, but it also generated two new interesting precursors (II and III) and the precursors (IV) of a rather exotic synthetic path (the symbols X represent different virtual atoms, and are generated by the program MAPOS itself) (see Figure 2).

The second example is taken from the work[8] where the guanine (V) synthesis was studied with the aid of computer program EROS. The MAPOS program found the same precursors as the EROS program, but it was able to find other precursors of the first or the second cycle of guanine as well. Cyanamides, ketenimines and isocyanates (and hydrocyanide), carbodiimides, and substituted acetylenes were found as potential precursors for the guanine synthesis. Several examples of the found use of cyanamides are given in Figure 3 (VI–VIII).

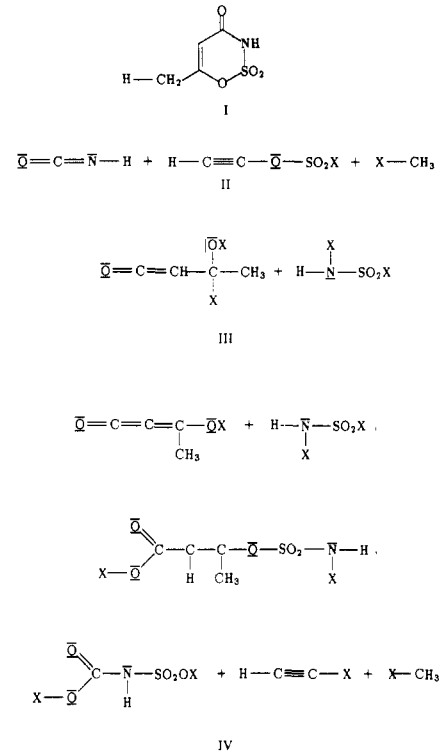


Figure 2.

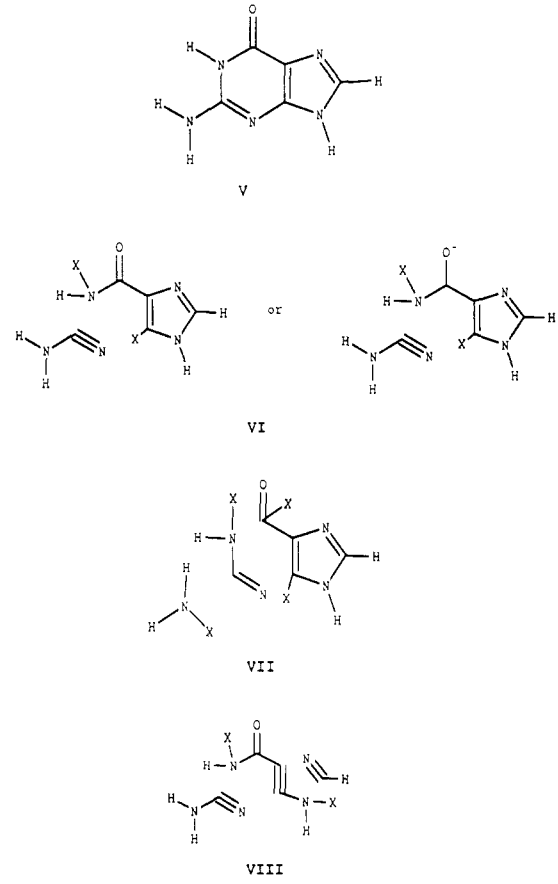


Figure 3.

6. CONCLUSIONS

The program MAPOS has been developed for the planning of the synthetic precursors or successors. It can be used for the computer organic synthesis design in the forward as well as retrosynthetic direction. It is an interactive program which serves as an "formulator" of synthetic ideas. Because of the

possibility to choose valence states, reaction center, and "rigid" parts, it is very flexible. The model on which the program is based is a guarantee that, under the correct use, no structure will be omitted. The program would like to be easy usable for the organic chemist, as the model itself is close to the "synthetic" chemist's thinking.

ACKNOWLEDGMENT

We are grateful to Prof. M. Kratochvíl for many helpful suggestions and comments on the ideas, program, and earlier drafts of this paper. Comments and suggestions of both the referees, which helped to clarify this paper, are also gratefully acknowledged.

7. APPENDIX: OUTLINE OF KEY ALGORITHMS

STAB Algorithm:

Purpose: to find all stabilizations of the synthon S′(X) with respect to S(A) (X ⊂ A).

1. If $|X| = |A|$ then $S'(A) = S(A)$ is output.
2. Find the set $Y \subset A$ of all atoms from the first neighborhood of set X.
3. Find the set $C \subset Y$ of atoms, whose valence state must be changed in order to connect them to (or disconnect from) atoms in $S'(X)$. There may be at most two such atoms.
4. If $|C| = 0$ (there is no such atom), immerse the $S'(X)$ to $S(A)$, and the resulting synthon $S'(A)$ is output.
5. For each $c_i \in C$ perform all permitted changes of its valence state and connect all atoms from Y to $S'(X)$ obtaining $S'(X \cup Y)$ [all atoms from set Y except atoms from set C may be connected to $S'(X)$ without change of their valence state].
6. For each synthon $S'(X \cup Y)$ call recursively the algorithm STAB. Joining of all these outputs (without duplicates and isomorphic synthons) is output of the algorithm.

GENFIS1 Algorithm:

Purpose: to generate all synthons $S(X) \subset \mathrm{FIS}(X)$, where each atom $x_i \in X$ is in the valence state $v_i \in V$.

1. If $|X| = 1$, then output the synthon $S(X)$ with $x \in X$ in valence state $v_i \in V$.
2. If $|X| > 2$, then go to step 7.
3. Generate all possible combinations of atoms x_1 and x_2 in valence states v_1 and v_2, respectively. This is performed by the following steps:
   4. If the valence states of both atoms permit at least one single-bond bonding, connect both atoms with this bond and add the synthon to the output set.
   5. Repeat step 4 for the double and triple bond, if possible.
   6. Add both atoms disconnected to the output set and end the algorithm.
7. Remove arbitrary atom α from the set X, and call recursively the algorithm GENFIS1 on the resulting synthon Y. Let the output of this call be denoted $\mathcal{L}$.
8. For each synthon $S(Y) \subset \mathcal{L}$ generate all possible combinations with the atom α in analogy to steps 4–6 above.
9. The set of all obtained synthons (without isomorphic duplicates) is the output of the algorithm.

GENFIS2 Algorithm:

Purpose: to generate all synthons $S(X) \subset \mathrm{FIS}(X)$ such that each atom $x_i \in X$ is in some valence state from the set $\mathcal{D}_i \subset \mathcal{D}$.

1. Generate some permutation V_j of valence states v_{ij} from $\mathcal{D}$.
2. Call GENFIS1 $(S(X), V_j, \mathcal{N})$.
3. Join $\mathcal{N}$ with results obtained thus far.
4. Generate new permutation V_j of valence states v_{ij} from $\mathcal{D}_i$; end the algorithm when such new permutation does not exist, otherwise continue on step 2.

REFERENCES AND NOTES

(1) Corey, E. J.; Long, A. K. J. *Org. Chem.* **1978**, *43*, 2208. Corey, E. J.; Long, A. K.; Mulzer, J.; Orf, H. W.; Johnsonand, A. P.; Hewett, A. P. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 221. Long, A. K.; Rubenstein, S. D.; Joncas, L. *Chem. Eng. News* **1983**, *61*, 22. Corey, E. J.; Longand, A. K.; Rubenstein, S. D. *Science* **1985**, *228*, 408.
(2) Wipke, W. T.; Dyott, T. M. *J. Am. Chem. Soc.* **1974**, *96*, 4825. Gund, P.; Grabowski, E. J. J.; Hoff, D. R.; Smith, G. M.; Andose, J. D.; Rhodes, J. B.; Wipke, W. T. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 88. Carter, R. E.; Wipke, W. T. *Kem. Tidskr.* **1981**, *93*, 20.
(3) Gelernter, H.; Sridharan, N. S.; Hart, A. J.; Yen, S. C.; Fowler, F. W.; Shue, H. J. *Top. Curr. Chem.* **1973**, *41*, 113. Gelernter, H. L.; Bhagwat, S. S.; Larsen, D. L.; Miller, G. A. *Anal. Chem. Symp. Ser.* **1983**, *15*, 35.
(4) Choplin, F.; Laurenco, C.; Marc, R.; Kaufmann, G.; Wipke, W. T. *Nouv. J. Chim.* **1978**, *2*, 285. Laurenco, C.; Villien, L.; Kaufmann, G. *Tetrahedron* **1984**, *40*, 2731.
(5) Bersohn, M. *J. Chem. Soc. Jpn.* **1972**, *45*, 1897. Bersohn, M.; Esack, A.; Luchini, J. *Comput. Chem.* **1978**, *2*, 105. Bersohn, M. *ACS Symp. Ser.* **1979**, *No. 112*, 341.
(6) Moreau, G. *Nouv. J. Chim.* **1978**, *2*, 187.
(7) Barone, R.; Chanon, M.; Metzger, J. *Tetrahedron Lett.* **1974**, *32*, 2761. Barone, R.; Chanon, P.; Metzger, J. *Chimia* **1978**, *32*, 216. Barone, R.; Chanon, P.; Cadiot, P.; Cense, J. M. *Bull. Soc. Chim. Belg.* **1982**, *91*, 333. Barone, R.; Chanon, M.; Contreras, M. L. *Nouv. J. Chim.* **1984**, *8*, 311.
(8) Dugundji, J.; Ugi, I. *Top. Curr. Chem.* **1973**, *39*, 19. Ugi, I.; Bauer, J.; Brandt, J.; Friedrich, J.; Gasteiger, J.; Jochum, C.; Schubert, W. *Angew. Chem. Int. Ed. Engl.* **1979**, *18*, 111.
(9) Wochner, M.; Ugi, I. *Chem. Ind.* **1986**, 498.
(10) Zefirov, N. S.; Gordeeva, E. V. *Usp. Khim.* **1987**, *56*, 1753. Zefirov, N. S.; Tratch, S. S. *Zh. Org. Khim.* **1975**, *11*, 225, 1785. *Ibid.* **1976**, *12*, 697. *Ibid.* **1981**, *17*, 2465. *Ibid.* **1982**, *18*, 1561. *Ibid.* **1984**, *20*, 1121. Tratch, S. S.; Zefirov, N. S. *Ibid.* **1982**, *18*, 1561. Zefirov, N. S. *Acc. Chem. Res.* **1987**, *20*, 237. Tratch, S. S.; Zefirov, N. S. In *Principles of Symmetry and Systemology in Chemistry*; Stepanov, N. F., Ed.; Moscow University Publishers: Moscow, 1987; p 54. Zefirov, N. S.; Tratch, S. S. *Chem. Scripta* **1980**, *15*, 4.
(11) Gasteiger, J.; Jochum, C. *Top. Curr. Chem.* **1978**, *74*, 93. Gasteiger, J.; Hutchings, M. G.; Christoph, B.; Gann, L.; Hiller, C.; Low, P.; Marsili, M.; Saller, H.; Yuki, K. *Top. Curr. Chem.* **1987**, *137*, 19.
(12) Bauer, J.; Herges, R.; Fontain, E.; Ugi, I. *Chimia* **1985**, *39*, 43. Fontain, E.; Bauer, J.; Ugi, I. *Chem. Lett.* **1987**, 37.
(13) Doenges, R.; Groebel, B. T.; Nickelsen, H.; Sander, J. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 425.
(14) Weise, A. *Z. Chem.* **1973**, *13*, 155. *Ibid.* **1975**, *15*, 333. Westphal, G.; Klebsh, A.; Weise, A.; Sternberg, U.; Otto, A. *Z. Chem.* **1977**, *17*, 295. Weise, A.; Scharnow, H. G. *Z. Chem.* **1979**, *19*, 49. Weise, A. *J. Prakt. Chem.* **1980**, *322*, 761. Weise, A.; Westphal, G.; Rabe, H. *Z. Chem.* **1981**, *21*, 218.
(15) Hendrickson, J. B. *Top. Curr. Chem.* **1976**, *62*, 49. Hendrickson, J. B. *J. Am. Chem. Soc.* **1977**, *99*, 5439. Hendrickson, J. B. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 129. Hendrickson, J. B.; Grier, D. L.; Toczko, A. G. *J. Am. Chem. Soc.* **1985**, *107*, 5228.
(16) Peishoff, C. E.; Jorgensen, W. L. *J. Org. Chem.* **1985**, *50*, 1056. *Ibid.* 3174. Bures, M. G.; Roos-Kozel, B. L.; Jorgensen, W. L. *J. Org. Chem.* **1985**, *50*, 4490. Metivier, P.; Gushurst, A. J.; Jorgensen, W. L. *J. Org. Chem.* **1987**, *52*, 3724.
(17) Cf., e.g., Hippe, Z. *Anal. Chim. Acta* **1981**, *133*, 677. Hippe, Z. *Stud. Phys. Theor. Chem.* **1981**, *1*, 249. Hippe, Z. *Przem. Chem.* **1985**, *64*, 285.
(18) Koča, J. *J. Math. Chem.* **1989**, *3*, 73. *Ibid.* 91.
(19) Koča, J.; Kratochvíl, M.; Kvasnička, V.; Matyska, L.; Pospíchal, J. *A Synthon Model of Organic Chemistry and Synthesis Design*; Lecture Notes in Chemistry 51; Springer-Verlag: Heidelberg, 1989.
(20) Pauling, L. *J. Am. Chem. Soc.* **1931**, *53*, 1367.
(21) Van Vleck, J. H. *J. Chem. Phys.* **1934**, *2*, 20.
(22) Koča, J. *Coll. Czech. Chem. Commun.* **1988**, *53*, 1007. *Ibid.* 3108. *Ibid.* 3119.
(23) Kvasnička, V.; Pospíchal, J. *J. Math. Chem.* **1989**, *3*, 111.
(24) Kratochvíl, M. *Chem. Listy* **1983**, *77*, 225.
(25) Koča, J.; Kratochvíl, M.; Kunz, M.; Kvasnička, V. *Coll. Czech. Chem. Commun.* **1984**, *49*, 1247.
(26) Kratochvíl, M.; Koča, J.; Kvasnička, V. *Chem. Listy* **1985**, *79*, 807.
(27) Koča, J.; Kratochvíl, M.; Matyska, L.; Kvasnička, V. *Coll. Czech. Chem. Commun.* **1986**, *51*, 2637.
(28) Schubert, W. *MATCH* **1979**, *6*, 213.
(29) Jochum, C.; Gasteiger, J.; Ugi, I. *Angew. Chem. Int. Ed. Engl.* **1980**, *19*, 495. Jochum, C.; Gasteiger, J.; Ugi, I. *Z. Naturforsch.* **1982**, *37b*, 1205.
(30) Bauer, J.; Ugi, I. *J. Chem. Res., Synop.* **1982**, *11*, 298. *J. Chem. Res., Miniprint 11*, 3101, 3201. Ugi, I.; Fontain, E.; Bauer, J. *Anal. Chim. Acta* **1990**, *235*, 155.
(31) Corey, E. J. *Pure Appl. Chem.* **1967**, *14*, 19. Corey, E. J.; Cheng, X.

M. *The Logic of Chemical Synthesis*; John Wiley and Sons, Inc.: New York, 1989.
(32) Kvasnička, V.; Kratochvíl, M.; Koča, J. *Mathematical Chemistry and Computer Aided Synthesis Planing*; Academia: Prague, 1987 (in Czech).
(33) Lindsay, R.; Buchanan, B. G.; Feigenbaum, E. A.; Lederberg, J. *Applications of Artificial Intelligence for Organic Chemistry*; McGraw-Hill: New York, 1980.
(34) Carhart, R. E.; Smith, D. H.; Gray, N. A. B.; Nourse, J. G.; Djerassi, C. *J. Org. Chem.* **1981**, *46*, 1708.
(35) Munk, M. E.; Lind, R. J.; Clay, M. E. *Anal. Chim. Acta* **1986**, *184*, 1.
(36) The same notation is used in all papers on the synthon model in order to facilitate the reader's task.

Fast Drug-Receptor Mapping by Site-Directed Distances: A Novel Method of Predicting New Pharmacological Leads

ANDREW S. SMELLIE,† G. M. CRIPPEN,*,‡ and W. G. RICHARDS§

BioCAD Corporation, 1091 North Shoreline Boulevard, Mountain View, California 94043, College of Pharmacy, University of Michigan, Ann Arbor, Michigan 48109-1065, and Physical Chemistry Laboratory, South Parks Road, University of Oxford, Oxford, OX1 4AR England

Received December 3, 1990

The searching and characterization of large chemical databases has recently provoked much interest, particularly with respect to the question of whether any of the compounds in the database could serve as new leads to a compound of pharmacological interest. This paper introduces a fast and novel method of determining whether any of a given series of compounds are able, on geometrical grounds, to interact with an active site of interest. The C program written to implement the method is able to make a qualitative prediction for a given compound in about 1 s per structure (for drug-sized molecules), while still permitting the compound complete conformational freedom. However, the algorithm is sufficiently flexible to permit distance constraints to be placed on the molecules while docking. The test system studied was a family of Baker's triazines docking into the active site of dihydrofolate reductase (DHFR), as defined by a methotrexate/NADPH complex.

INTRODUCTION

(a) Overview. There exist in the literature numerous algorithms for performing docking of small molecules (usually drugs) into a larger binding site (usually a protein). This is the drug-receptor mapping problem, which breaks down into various classes. However for all classes of problem, the ultimate goal is always the same—can the binding conformer(s) of molecules be predicted for a given receptor site? The binding conformers are referred to as *binding modes*.

The following sections describe each class of problem. "Known" is taken to mean that the structure or conformer of the receptor and/or drug is known. "Unknown" means that the topology of the molecule is known, but the conformation is not.

Known Receptor (Site), Known Drug (Molecule). Here a conformation is assumed for the molecule, and predictions are made for the binding of this conformer only. Hopfinger proposed molecular shape analysis as a good criterion for mapping site and molecule[1] but this is not suitable for introducing conformational flexibility in the molecule.

Known Receptor (Site), Unknown Drug (Molecule). The scope of this paper falls into this category. Given a known receptor site and an unknown molecule, an algorithm is presented that will predict possible binding modes for the molecule. Complete conformational flexibility of the molecule is permitted in the site.

The drug-receptor mapping problem is represented as a bipartite graph in which the atoms of the site form one set of nodes (x) and the atoms of the molecule form a second set of nodes (o). By definition the edges in this bipartite graph link nodes from set x to nodes from set o. It can be seen from Figure 1 that the bipartite graph representation is analogous to the docking problem, where graph nodes are site or molecule atoms and graph edges are site–molecule interactions.

Binding modes can be extracted very rapidly from the bipartite graph by the docking graph approach,[2] which is described later in this paper.

Unknown Receptor (Site), Known Drug (Molecule). An example of this problem would be where binding data of several molecules are available, and it is known that they are binding at the same site. However, the structure of the site is not known.

The Voronoi binding site model of Crippen et al.[3,4] proposes a mathematical model of the binding site that is able to reproduce and predict binding data (in the form of equilibrium constants K_i, and hence free energy of binding from $\Delta G_{obs} = RT \ln K_i$). However, the abstract model site can bear little resemblance to the actual protein.

The clique graph algorithms of Kuhl[2] first introduced the docking graph concept used in this paper but without the fast filtering used here to greatly increase computational efficiency.

(b) Definitions. A few graph theory definitions will serve to clarify some of the points to follow, but for a more complete description of graph theory see ref 5.

An undirected *graph*, G, is a set, $u_g = \{u_i\}$, of *nodes* or *vertices* and a set $e_g = \{(u_i, v_i) | u_i, v_i \in u_g\}$ of unordered pairs of nodes interpreted as *edges*. A molecule can be thought of as a special type of graph where the atoms are the nodes and the bonds are the edges of the graph.

A *subgraph*, S, of G is composed of a subset of vertices and edges of G. Thus $u_s \subset u_g$ and $e_s \subset \{(u, v) | (u, v) \in e_g, u, v \in u_s\}$. A graph is *completely connected* if there is an edge between all pairs of vertices. For a completely connected subgraph of G, S_T, with N_T vertices, there are $N_T(N_T - 1)/2$ edges in set e_T where $e_T = \{(u, v) \; \forall \; u, v \in S_T\}$. The graph G* is a *clique* of G if it is a *maximal* completely connected subgraph of G.

A *bipartite* graph, B, is a graph where the nodes have been partitioned into two sets; $u = \{u_1, ..., u_N\}$, $v = \{v_1, ..., v_M\}$ and each edge involves exactly one vertex from set u and one vertex

† BioCAD Corporation.
‡ University of Michigan.
§ University of Oxford.

0095-2338/91/1631-0386$02.50/0 © 1991 American Chemical Society