Tetrahedron Vol. 45, No. 9, pp. 2665 to 2676, 1989
Printed in Great Britain.

0040-4020/89 $3.00+.00

ORGANIC SYNTHESIS PLANNING: AN ALGORITHM FOR SELECTING STRATEGIC BOND FORMING SEQUENCES

Luca BAUMER(#), Giordano SALA, Guido SELLO *.

Dipartimento di Chimica Organica e Industriale,
Centro di Studio per le Sostanze Organiche Naturali del C.N.R.,
Universita' degli Studi, via Venezian 21, 20133 Milano, Italy.
(#) Present address: Farmitalia Carlo Erba, Erbamont Group,
R & D Analytical Chemistry, Via dei Gracchi 35, Milano, Italy.

(*Received in UK* 31 *January* 1989)

Abstract - An algorithm part of a computer program (Lilith) for convergent retrosynthesis planning is described. It chooses the optimal forming sequence for a given set of strategic bonds. Its heuristic rules consider complexity distances, steric congestions and ease of cyclization. The procedure used for fast evaluation of approximated congestion at reaction center is outlined. Linkings of the algorithm with the previous parts and with the general approach of Lilith are emphasized. Applications to natural substance synthesis projects are exemplified.

Introduction.

This paper is part of a series (1-4) in which we describe our program Lilith, a system for OSP presently under development at our department. Previous papers were about topological analysis of a target molecule for ring systems perception, and the logical principles on which the selection of the strategic bonds is based. Of course, all of the following procedures must work on the set of strategic bonds which constitutes the retrosynthetical solution to the synthesis problem; in the next section we recall some of the main concepts and terminology.

Previous works in the field (5) often use heuristic rules to select a precursor of the target as a necessary (or privileged) intermediate on the way from the simple starting materials to the target, in order to reduce the otherwise enormous space of the solutions; deciding that a particular bond is to be made 'before' or 'after' a certain point in the synthesis is sometimes an implicit consequence. On the contrary, the strategic section of Lilith already outputs a very limited set of possible solutions for each fragmentation (and the convergence principle grants a low number of fragmentations); so we consider the sorting of bond formation order 'inside' each synthetic passage more as an optimization step devoted to the identification of the most suited precursors than as a method to prune the synthetic tree.

Hendrickson, whose SYNGEN (7) is currently the paramount example of a program for convergent retrosynthesis, has already demonstrated (8) that the optimized synthesis path is the sequence:

1) starting materials / affixation / intermediate 1;
2) intermediate 1 / cyclization / precursor 1
3) precursor 1 / affixation / intermediate 2
4) intermediate 2 / cyclization / target
neglecting the refunctionalization steps and iterating steps 2-3 as many times as necessary, depending on target's size.

However, the two or more bonds whose formation constitutes a cycle affixation cyclization are interchangeable, and no theory or heuristics is presently given to select which bond must be the affixation one, and which is the ideal order of cyclizations. Hendrickson was aware of the problem: in fact he stated (9) that "bondsets may be further defined by the order in which (the bonds) are constructed". But how is this "ordered bondset" chosen? The answer "convergency defines the order of bond making in a bondset" is sufficient only for the subsequent fragmentations of an acyclic compound.

At this level only skeleton construction is taken care of; refunctionalizations, activations, protections, etc. will be evoked at a later stage.

Furthermore, we always refer only to the first fragmentation; if a requirement of reaching small / simple / available starting materials is not fulfilled, either one of the two synthons obtained, or both, can become the target for a new, analogous processing.

Lilith premises and terminology

Depending on the context, we speak of 'bond forming (or making) sequence' or of 'bond breaking (or cleaving) sequence', which are the reversal of each other. However, no ambiguity should arise, because Lilith does not allow skeleton disconnections on the way from precursors to target; so, 'bond breaking' always implies a retrosynthetical (antithetical) walk on the reaction path, and conversely 'bond forming' implies a synthetical direction.

The basis of Lilith's strategy is the well-known principle of convergence. It means that the best solution is the one which makes the target out of two fragments of similar complexity, and the smaller the complexity difference, the better the solution. In our approach, "complexity" is defined (1) as a simple function that considers the number of atoms in each fragment as well as their species, substitution levels, stereochemistries and mutual positions.

From the definition of atomic complexity, the concept of "complexity distance" (between two atoms) follows: it is the shortest complexity path joining them.

Distance calculation identifies the "centre" of the molecule as a set of atoms halfway between the farthest points of the structure. All the strategic bonds chosen for convergent retrosynthesis span from these central atoms.

The output of the strategic section, which we call MSSB, is a set of solutions; each solution, in turn, is a set of strategic bonds (a subset of the MSSB). Each solution contains only the bonds necessary to split the target into two parts for a particular fragmentation; the synthetic subgraph representing a single solution contains only the intermediates originating from different bond forming sequences, and they all converge into the double leaf of precursors.

The object of this paper, is the sorting of the N bonds in every proposed solution, i.e. choosing the 'best' chronological sequence in which they must be formed among the N! possible sequences. When exiting this procedure, the synthetic graph is reduced to a tree made by a single unbranched line for each solution. Of course, some intermediate points may coincide, i.e. the first breakage(s) can be common to some solutions; this allows an alternative representation of the final synthetic graph, as shown in FIG. 1.

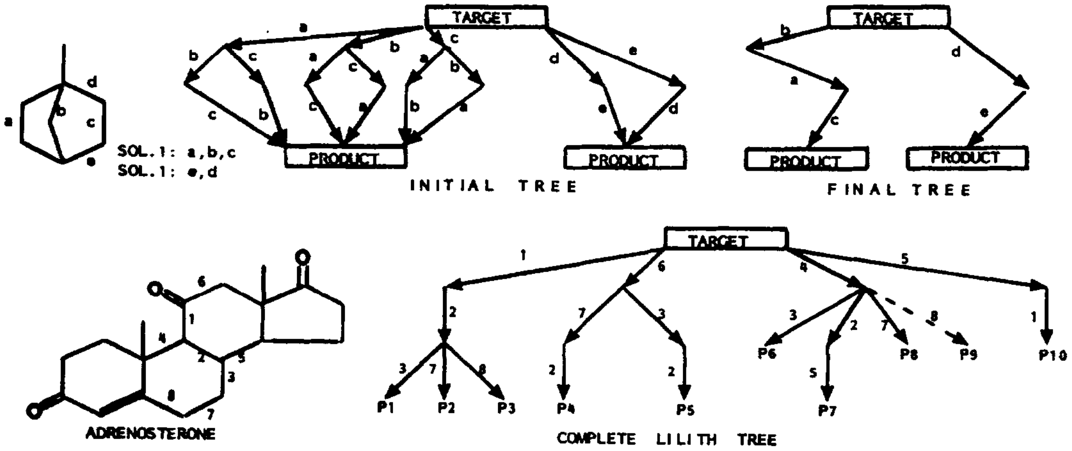


FIG. 1

The procedure applies only to solutions which require the breaking of more than one bond; this is equivalent to say that the procedure applies only to substances that contain one or more rings in their centre. This follows from the forcing requirement which divides the target into two (and no more than two) separate parts.

Principles

No absolute rule applies to this subject; like many other points in a not-computerized process of synthesis planning, it generally relies upon the chemist's experience and sensitivity. A computer needs rules, though. We chose these rules applying logic and coherence with the strategic section to some basic, general principles. We thus examine bond centrality, ease of reaction, steric hindrance and ring closures.

The main criterion is strategic efficiency, that is, adherence to strategic rules. We stated that the most important atoms are the central ones and that only bonds spanning from them can be part of MSSB; consequently, bond centrality is also the first topic to consider when deciding the bond forming sequence.

Another important point to consider is the different ease with which an intramolecular reaction can take place compared to an intermolecular one: this is related to steric hindrance evaluation. For this reason the routine is divided into two parts: the first part selects the first bond to be made, and the second sorts all the remaining bonds. This follows from the consideration that the first bond is the only one made through an intermolecular reaction, while all the following reactions are intramolecular. In general, intramolecular reactions are easier; in particular, they have less strict requirements from the point of view of steric accessibility, if they have sufficient conformational freedom. It is therefore sensible to choose to prepare first a bond whose reaction sites are not heavily hindered.

Mutual topological relationship between bonds is the third factor necessary to the task.

The algorithm

In order to select the first bond to be formed, the target is nearly completely fragmented cutting all but one of the strategic bonds of the current solution. (This means that all the necessary bond and connectivity matrices that constitute the computer's representation of the molecular structure are modified). No atom or group is added or substituted to saturate the valence of the separated atoms. This new structure is processed again by the procedure (3) devoted to the calculation of complexity distances and to the identification of the couples of atoms whose shortest joining path is the longest existing in the molecule. We called (1) this distance between molecular extremes MAXD; we now distinguish between MAXD0 (the target's MAXD) and MAXD1, MAXD2..., respectively in the structure where only the strategic bonds 1, 2..., are not cleaved. Two points must be stressed: firstly, as a general rule, the extremes of these structures will not coincide with each other's extremes nor with the target's extremes. Secondly, no MAXDi can be less than MAXD0, because breaking bonds decreases the number of possible paths between extremes, and, of course, no new shorter path can appear.

From our definition of central atoms and strategic bonds, it follows that the bond which minimizes MAXDi is the most central (with respect to the target); it means that this structure is the most 'similar' to the target. This should also be graphically evident from a comparison of the structures in FIG. 2, which also shows that the partially cleaved structure with minimal MAXDi is also the one which maximizes the coincidence of the extremes with those of the target (10).

So, if this bond is constructed first, the first synthetic step will arrive at a product which is as similar as possible to the target itself, and it should be easier to perform the subsequent ring closures in the desired direction. This is the principle of maximum strategic efficiency, which follows directly from the joining of simplification and convergence (11).

FIG. 2 exemplifies the results for a few structures. (Dots highlight atoms that are members of extreme couples, whose MAXD is reported.)

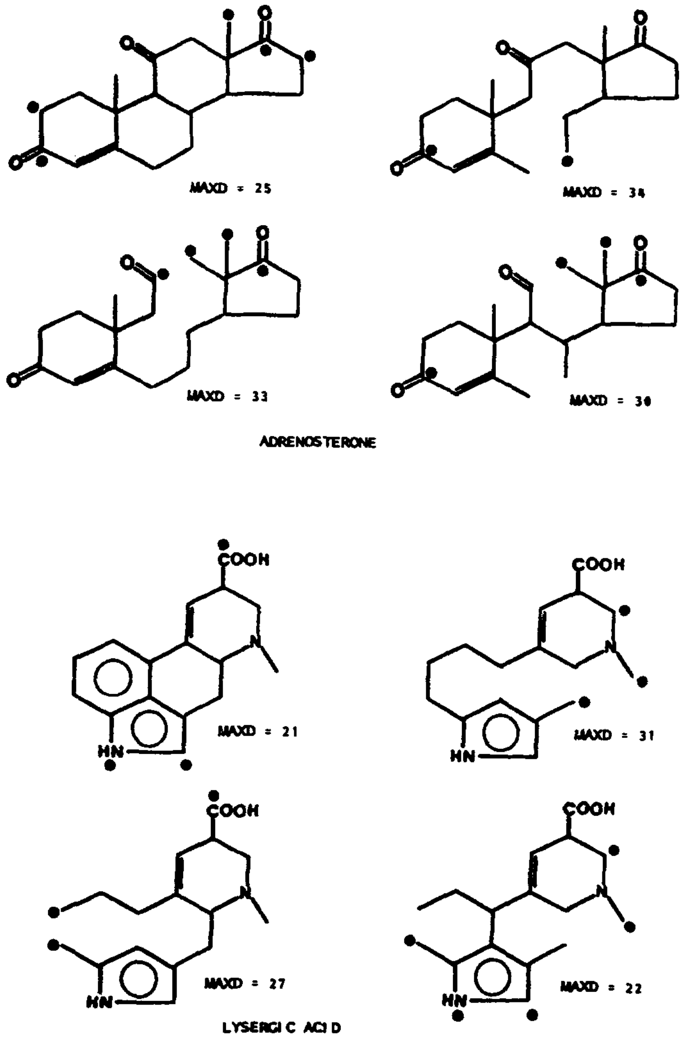


FIG. 2

The simplification made on the complete synthetic tree is evident in the figure. In both structures there are three strategic bonds whose breakage, without an ordering, would produce 6 possible solutions, that the algorithm cuts down to 1. (The situation is similar to the one shown in the first half of FIG. 1, in the left part of the tree.)

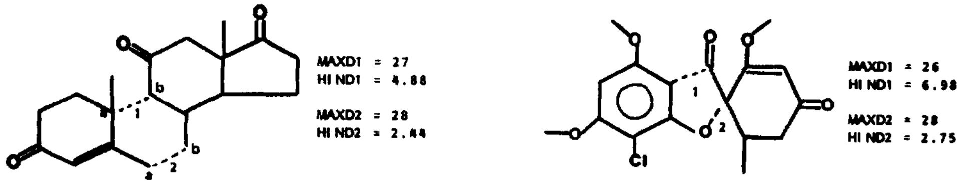
Let us now consider the molecule of adrenosterone to explain the work done by the algorithm.

The target has 3 couples (formed by 5 atoms) with MAXD equal to 25, the first solution has 1 couple with MAXD equal to 34 and only 1 atom coinciding with one of the target, the second solution has 3 couples with MAXD equal to 33 and 2 atoms coinciding with 2 of the target, the third solution has 3 couples with MAXD equal to 30 and 3 atoms coinciding with 3 of the target. This last is evidently the solution giving the starting structure most similar to the target.

Strategic efficiency is thus the main criterion we use to order the bond forming sequence: the bond with the smallest MAXDi must be formed first, the bond with the largest MAXDi will be the last, the others (if any) in order of increasing MAXDi.

This basic criterion could be overridden if the first resulting reaction (the intermolecular one) is difficult because of heavy steric congestion at the reaction site. Congestions of atoms A and B, which must form the i-th strategic bond A-B, are classified as G.A and G.B by the subroutine described below. The total congestion G.T(i) of bond A-B is quantified through a heuristic function, which is essentially an additive parameterization of G.A and G.B (12). If the i-th bond has minimum MAXDi, but there is a j-th bond such that (MAXDj - MAXDi) < (G.T(i) - G.T(j)),then bond j replaces bond i in the priority sequence.

FIG. 3 shows two examples of this kind of swap.



The best swapped for its hindrance.

FIG. 3

Another possibility which must be checked is the double-affixation case (breakage of two bonds spanning the same atom (1)). Lilith contains no explicit reference to any particular reaction (no reaction database), but from experience it is known that cyclizations involving a double-affixation can often be performed in a single synthetic step. Forming two bonds at the same time has a higher strategic efficiency (closer approach to the target) than forming a single, though central, bond. So, if there are three strategic bonds a, b, c such that MAXDa < MAXDb and MAXDa < MAXDc, but b and c are in double-affixation with each other, while a is not, then a is replaced by the lower-MAXD bond between b and c.

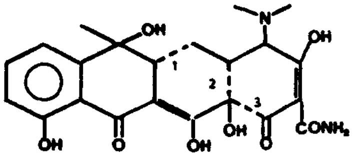
At this point, the last breakage is definitively chosen, and the procedure which sorts the others is entered.

Its purpose is to obtain a breakage sequence which maximizes the ease of the cyclizations. The first step is to place the bonds in double-affixation with the best bond, if any, in the position adjacent to it.

The adjacency of bonds which are part of the same ring is at a lower priority level. We have already stated that an MSSB containing more than two bonds implies a polycyclic structure, either fused or bridged, at the center of the molecule.

So, for every pair of strategic bonds, we could find a ring or envelope of rings containing them both. But we always refer only to the Smallest Set of Smallest Rings (SSSR), as defined and described in a previous paper (3). Using this SSSR for sorting bonds, the algorithm gives precedence to the creation of the smallest and 'relevant' rings.

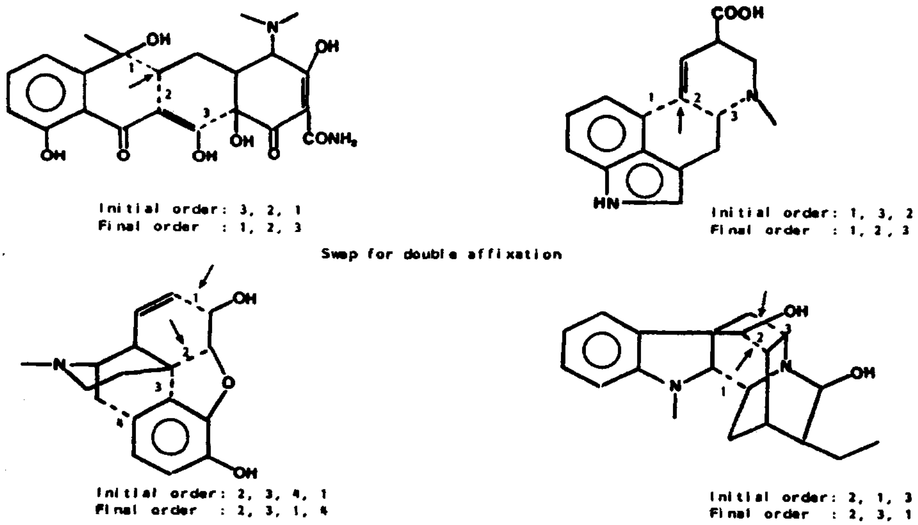
An example of a double swap is shown in FIG. 4 and few examples of single swaps related to double affixation or to the participation of two bonds to the same ring are shown in FIG. 5.



Initial order: 1, 2, 3
Final order : 2, 3, 1

Double swap for double affixation

FIG. 4



Initial order: 3, 2, 1
Final order : 1, 2, 3

Initial order: 1, 3, 2
Final order : 1, 2, 3

Swap for double affixation

Initial order: 2, 3, 4, 1
Final order : 2, 3, 1, 4

Initial order: 2, 1, 3
Final order : 2, 3, 1

Swap for same ring participation

FIG. 5

Steric congestion

A correct evaluation of 3-dimensional hindrance at a particular atom in a molecule is possible only after a complete conformational analysis, but such a calculation requires far too much computation time than is acceptable for use inside Lilith that needs a fast procedure to evaluate gross hindrances. In fact, since this calculation must be performed many times for every strategic solution proposed for the target, its principal feature must be speed. This is also coherent with the principles of simplification and Initial-Approximation / Increasing-Accuracy (IAIA) approach.

Thus we had to neglect every 3-dimensional approach, and focus on a way to extrapolate an approximation of real steric hindrance from the 2-D molecular graph.

Steric hindrance is a property of a particular reaction mechanism, (13, 14) affecting the activation entropy (frequency factor) for the transition state of that mechanism (15).

At this level of the target processing, the reaction to use to form the bond under examination is still unknown; no mechanism can thus be simulated. We have to work on the most general situation and evaluate a hindrance of general applicability. It is more correct to speak of steric 'congestion', defined by Wipke (16) as "a property of the substrate molecule in its ground state, and thus only one part of the total effect called steric hindrance".

However we can consider that this 'undefined' reaction must contain a weighted average of the most common mechanisms. In particular, there is a good confidence that the use of SN2 steric requirements as main factor to calculate congestions will bring the most useful approximation Lilith can reach with the small amount of established data. In fact, SN2 pentagonal activated state is more crowded than many others (e.g. addition to double bonds, tetragonal substitution at carbonyl, etc.) and thus it must have a higher weight when considering the role of steric congestion. Steric effects in SN2 depend much more (17) on the alkyl groups (and thus on the target structure) than on the entering and leaving groups and thus congestion is a closer approximation to steric hindrance than it would be for other reactions.

SN2 is also one of the reactions which received the largest attention since the first mechanistic studies (13 and related papers). As a consequence, there is a sound basis of experimental data and theoretical interpretations. (18, 19).

After Ingold's fundamental work (19) attention focused more on different aspects (20) until the end of the 70's when especially DeTar and co-workers (22,23) re-examined SN2 reactions using molecular mechanics in theoretical calculations. They revised Ingold's experimental data and hypothesis, and eventually confirmed the essential (and almost quantitative) correctness of Ingold's conclusions. But the set of useful data is still limited to the same series of structures already used by Ingold. No extension to structures of intermediate congestion (e.g. 2-butyl, 3-pentyl, diisopropyl) has been made, and the lack of additivity of steric effects (24) prohibits any quantitative interpolation from known data. The extension to structures more congested than neo-pentyl, sometimes examined for other reactions (25) has no practical utility from the point of view of SN2 reactivity, but it would be necessary in order to have a homogeneous scale of general applicability.

We could however find a way of getting from such a limited set of data an estimation within the desired approximation level.

Let C(i) be the corrected connectivity of i-th atom, i.e. its substitution level neglecting zero-weighting (phenyl group excluded) substituents.

The congestion G.(i) at the i-th atom, which bears A alpha-substituents a1... aA, is expressed by a discrete classification of the connectivity of the molecular graph, around the i-th node, of the type:

$$G.(i) = f\ (\alpha, \beta M, \beta T)$$

where $\alpha = C(i)$, $\beta T = \sum(j=1,\ A)\ (C(aj))$, βM = MAX (C(aj)) and G. is an integer in the range [0 - 7].

The calculus of G. for atom (i) is done in the following way: 1) the strategic bond object of the examination is cut; 2) atoms in the alpha sphere of atom (i) are counted (α) (only skeleton atoms (1) are considered); 3) the number (βT) of alpha substituents, excluding atom (i), for each atom in the alpha sphere is evaluated (only skeleton atoms (1) are considered); 4) the maximum (βM) of beta atoms on each alpha atom is evaluated.

At this point a choice, based on α, βM and βT values, is done and atom (i) is assigned to its congestion class.

It means that we consider the number of alpha-substituents (α), the number (βT) and the distribution (βM) of beta-substituents and their atomic species (17, 26, 27).

An example can make the algorithm clearer. Let us consider FIG. 3. The first example is adrenosterone and the solution proposed considers the breakage of two bonds.

For bond 1 we have: atom (a), α = 3, βT = 3, βM = 2, the class G. is equal to 5 (cfr. FIG. 6); atom (b), α = 2, βT = 3, βM = 2, the class G. is equal to 3 (cfr. FIG. 6).

For bond 2 we have: atom (a), α = 1, βT = 2, βM = 2, the class G. is equal to 2 (cfr. FIG. 6); atom (b), α = 1, βT = 2, βM = 2, the class G. is equal to 2 (cfr. FIG. 6).

The bond 1 congestion is greater than the bond 2 congestion.

The returned value G. is a rough classification of the steric congestion at the atoms, generated by the first two neighboring spheres. Where required, this value is modified to take care also of the case of bridgeheads of ring systems.

There are 34 possible skeleton arrangements for a central tetravalent carbon atom, and they are grouped by their G. value into 7 major classes (28).

An example for each class is reported in FIG. 6.

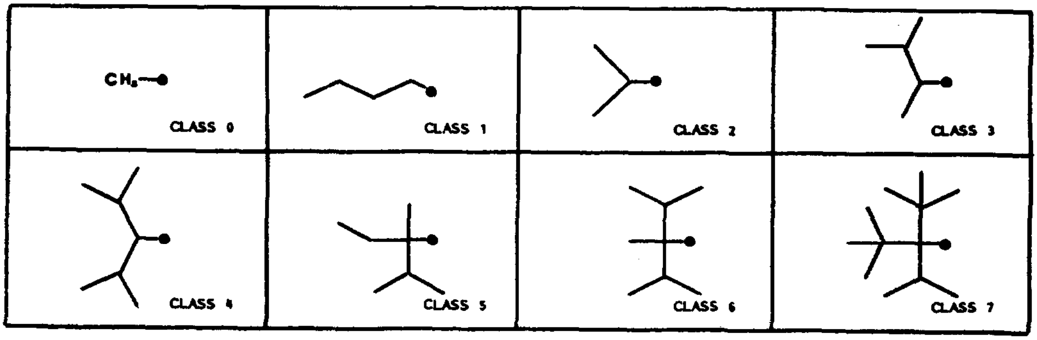


FIG. 6

If the nucleophile is a complex molecule, as it is in skeleton-forming reactions, chain branching around the attacking atom is also important (see refs. 114, 115 in 19) (even if not as much as chain branching around the attacked atom).

Its congestion can be evaluated with the same criteria, and thus the calculus is applied to both of them.

Discussion and conclusion

The creation of a breakage sequence is an important part of a CAOS program, since it actually selects a very limited part of the synthetic tree and is a guideline for the reactivity subprograms.

In Lilith it is executed at a very early stage of processing, and in spite of the insertion of some chemical heuristics, it is essentially topology-based, like the previous strategic section.

This coherence ensures the self-consistency of the system, but neglects some factors which could eventually prove very important; e.g., approximations in congestion evaluation.

Besides, Lilith arrives at this point without having any definite idea either on the type of reactions needed to create these bonds, or on the interferences that will arise. However, the answer to these problems was partially foreseen when Lilith's global structure was first planned.

The next step of Lilith process is the calculation of the outstanding electronic properties of the target (4) that are used to address reactivity.
So, we solve the approach to reactivity by reversing the more common procedure: instead of selecting strategic bonds after the identification of useful / known / feasible reactions, the reactions will be chosen, together with the necessary group transforms / additions, according to the bond breaking sequence.

The first interference block, which is three steps away from this subroutine and is currently being tested, will be able to modify the bond forming sequence and consequently the reaction plan.

A more exact 3-D analysis of hindrance is a future foreseen improvement, to be applied at a deeper reactivity and interference analysis level.

The general trend in Lilith is always from approximation towards a more and more exact knowledge (IAIA approach).

The strategy generates a small number of solutions, which are evaluated to process an even smaller number (the 'best' ones only) at any time.

As shown in FIG. 7, this will allow the use of a widely recursive flow-chart, with a higher reliability at every loop, but without unacceptable increasing of processing time.

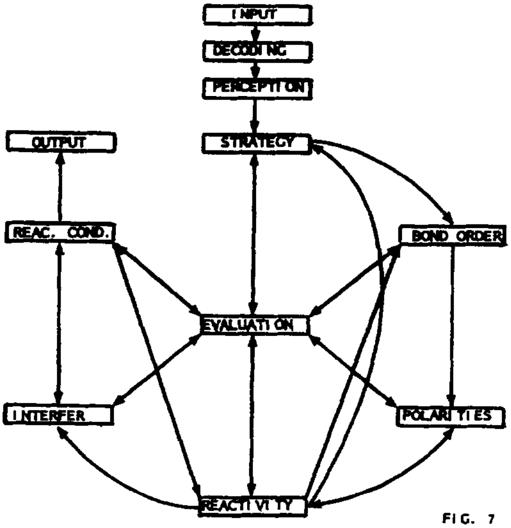


FIG. 7

In the figure the different activities of the program are reported.

1) An alphanumeric INPUT of the target is followed by its DECODING (creation of the adjacency matrix, identification of atomic species and of bond orders, etc.).

2) In the PERCEPTION phase some molecular features are identified: rings (SSSR), aromaticity, atomic strategic weights, etc.).

3) The STRATEGic part finds the MSSBs (1).

4) BOND ORDER determines the optimal bond forming sequence (the present algorithm).

5) POLARITIES identifies the natural polarities of bonds.

6) REACTIVITY is concerned with the absolute bond reactivity.

7) INTERFER evaluates possible interferences either with other bonds in the target or with reaction conditions.

8) REAC.COND. suggests optimal reaction conditions.

9) EVALUATION is a central block that assigns a value to each decision made by the program. As evidenced by the arrows most of the blocks exchange information and their activities can be influenced by the results of the other blocks. This allows a continuous evolution of the search towards better suggestions.

Blocks 1, 2, 3, 4 and part of 9, are already working, block 5 is also complete and will be the object of a future paper.

As far back as 1974, Bersohn, writing one of the first reviews about CAOS (6), stated repeatedly his expectation for "an effort by chemists to enunciate precise rules" and for "the chemist's imagination to be dissected, analyzed and made automatic". These expectations are still unfulfilled, and this is the root of CAOS expert systems' approximations. If our more effective method is a trial-and-error one, the computers we teach cannot do better.

We wrote our algorithm in Fortran 77 and implemented it on a Honeywell-Bull X-20 microcomputer, using the SVS Fortran compiler.

Acknowledgement - Partial financial support by Tecnofarmaci S.p.a. is gratefully acknowledged. Two of us (L. Baumer and G. Sala) acknowledge a postdoctoral fellowship by Tecnofarmaci.

References.

1) Baumer L., Sala G., Sello G., Tetrahedron 44, 1195 (1988)
2) Baumer L., Sala G., Sello G., Gazz. Chim. Ital., 118, 745 (1988)
3) Baumer L., Sala G., Sello G., Comp. Chem., submitted
4) Baumer L., Sala G., Sello G., in preparation
5) see (6) for a first review; and the other papers cited in the following.
6) Bersohn M., Esack A., Chem Rev., 76, 269 (1976)
7) Hendrickson J. B., Acc. Chem. Res., 19, 274 (1986) & refs therein
8) Hendrickson J. B., J. Am. Chem. Soc., 99, 5439 (1977)
9) Hendrickson J. B., Braun-Keller E., Toczko G. A., Tetrahedron. suppl. 1, 37, 359 (1981)
10) This is true only qualitatively: there is no evidence for a mathematical relation between the bond centrality and the number of kept (or lost) extremes and their distance from the target's.
11) Bertz S. H., "Studies in Physical and Theoretical Chemistry", 28, 206, ed. R.B.King, Elsevier, Amsterdam (1983)
12) Also the overall G. of all the bonds of the solution is calculated. Solutions with a very high global congestion are considered 'worse', and can eventually be rejected.
13) Dostrovsky I., Hughes E. D., Ingold C. K., J. Chem. Soc., 173 (1946)
14) Hughes E. D., Quart. Rev., 2, 107 (1948)
15) Ivanoff N., Magat M., J. Chim. Phys., 47, 914 (1950)
16) Wipke W. T., Gund P., J. Amer. Chem. Soc., 96, 299 (1974)
17) De La Mare P. B. D., et al., J. Chem. Soc., 3200 (1955)
18) Streitwieser A., Chem. Rev., 56, 571 (1956)
19) Ingold C. K., Quart. Rev., 11, 1 (1957)
20) e. g. solvolysis, or the steric requirements of the carbonyl group reactions (21). We do not mention here also the numerous free-energy linear relationships derived from experimental data since the steric parameters used in these equations are related to well defined combinations of reaction mechanisms and classes of substances. The variations required to generalize the use of these parameters would cover such a widespreaded range to loose any quantitative information content.
21) Wipke W. T., Gund P., J. Amer. Che. Soc., 98, 8107 (1976) and refs therein
22) De Tar D. F., McMullen D. F., Luthra N. P., J. Amer. Chem. Soc., 100, 2484 (1978)
23) De Tar. D. F., J. Org. Chem., 45, 5174 (1980)
24) Popov A. F., Matveev A. A., Piskunova Z. P., Org.React. (Tartu), 300 (1986) (C.A. 108, 5384m (1988))
25) De Tar D. F., Binzet S., Darben P., J. Org. Chem., 52, 2074 (1987)
26) Lowry T. H., Richardson K. S., "Mechanism and Theory in Organic Chemistry", Harper & Row, New York, (1976)
27) Metivier P., Gushurst A. J., Jorgensen W. L., J. Org. Chem., 52, 3724 (1987)
28) The 8th class, corresponding to b = 0, contains only isolated skeleton atoms (terminal methyl groups), a case forbidden by the strategic fragmentation rules, but provided for the following interference analysis section.