Tetrahedron Vol 44, No. 4, pp. 1195 to 1206, 1988
Printed in Great Britain.

0040-4020/88 $3.00+.00

ORGANIC SYNTHESIS PLANNING: A NEW ALGORITHM FOR STRATEGIC BOND PERCEPTION.

LUCA BAUMER, GIORDANO SALA, GUIDO SELLO*.

Dipartimento di Chimica Organica e Industriale,

Universita' degli Studi, via Venezian 21, 20133 Milano, Italy.

(*Received in UK* 1 *December* 1987)

Abstract - A new algorithm which generates breaking suggestion for Organic Synthesis Planning (OSP) is described. A new strategical approach to organic synthesis design has been used. New definition for structural complexity weight, atom complexi_ ty distance, molecular centre, Minimal Sets of Strategic Bonds are presented. The need for separate program sections for the strategic part and the transformation model is addressed. Some results are given and comments made to illustrate the algo_ rithm performance.

Introduction.

The last two decades have seen new trends in OSP. Approaches based on analogy and authoritative synthetic metodologies have been substituted by a more logical and rational approach to synthesis design.

In the late sixties, E.J. Corey and his group elaborated a novel and revolutio_ nary approach to OSP, changing perspectives, methods and even the language used in the field (1).

Not longer after this, other groups presented their results and ideas, rapidly increasing the level of the knowledge in the field (2). Since then, the scienti_ fic community has learned to think of OSP as an autonomous intellectual challenge.

Particular effort has been put into the search for basic principles operating in the development of organic synthesis and for well-suited work methodologies. In OSP important attempts have been made to simulate human intellectual activity and to search for experimentally based theoretical modelling.

Computer applications, which calls for a straightforward rationalization of the matter, provides the best opportunity to correctly analyze OSP. Automation has allowed the expansion of the space for analysis and the unusual but important chance to put into effect large numbers of theoretical trials in

short time intervals; for the very first time chemists have had at their disposal a fast a powerful means of simulating thousands of reactions under different chemi_ cal situations.

An important aspect of OSP is the dualism between the logical evaluation of the best synthesis and the model for the transformation of one molecule into another.

Each approach has its own logical control, explicit or not, and its own trans_ formation model. In all but one case, the greatest drawback is the indistinct separation of the two parts, thus causing a partial misunderstanding of OSP prin_ ciples.

In fact, sometimes the logic used in the planning and at times the real struc_ ture of the transformation model are unclear. Hendrickson's approach (3) is the only one that exactly defines the inference areas of the two sections.
His approach has an explicit strategic section, totally devoted to screening the results for subsequent examination. This strategic section represents a novel method in OSP because the principles used are only logical and are unrelated to analogy or past experience.

It is in the model section that the program decides the feasibility of a solu_ tion, and performs the actual transformation.

Many of Hendrickson's ideas have, in fact, been incorporated in our own approach, the main being his suggestion to separate the two different working sections.

It is well known that, at the actual state of the art, OSP is characterized by the impossibility to find an exact solution. Its size (4), the principles it uses, the inadequate definition of rules and limits, are responsible for the difficult standardization of the work in the field.

For these reasons, we decided to confront the strategic part first.

Beginning from scratch, we chose a theoretical approach to OSP and thus were in a good position to elaborate new ideas. Some guidelines used in the proposed solution were heuristic, others logical, but all represents an attempt to obtain a correct behaviour of the automatic analysis of organic synthesis.

This article aims to be the first building block, the height of the building being unknown at the moment.

A logic for OSP.

Following Hendrickson's work, we are developing a system (named Lilith (5)) where the control part (strategy) is a self-standing section whose principal ac_ tivities are based only on logical suggestions.

When working on large problems, the early reduction of the alternatives is im_ portant so that subsequent work is easier and faster.

It is also important to search for the "best" solutions, discarding results with poor chances of success.

In some past programs this kind of solution pruning is performed here and there but not in a defined work activity (6). In some other programs, a similar reduc_

tion of the solution space could not be found (7).

The rules that an approach uses in pruning must obviously be consistent with the spirit of the theory operating in the system. These rules must be well defined and must always operate correctly, regardless of the molecule.

Strategy.

Many different logics are commonly used when choosing a strategical approach to organic synthesis; some are experimentally based, some analogically based and some are not based at all.

In searching for our strategy, we decided to use general rules only, forgetting physical properties.

Strategy is, in our opinion, something related to efficiency of the planning and not to experimental reliability.

After an accurate search for strategic principles present in the literature, we focused on convergence and structure simplification as the guidelines for our strategy. (As stated by Hendrickson, convergent strategy are much more efficient than linear ones.)

The usual meaning of convergence is to get, retrosynthetically, pieces of similar "weight". The cornerstone of the strategy is the identification of the weight; it is not sufficient to have precursors that are similar just because they contain a similar number of carbon atoms or because they have a similar number of rings. A way to determine the "complexity" weight of a structure must be found. (8)

We realized an algorithm able to evaluate complexity and to generate precursors of similar weight.

First of all, we had to specify what aspects participate in determining the complexity weight of a structure. The choice is not easy because, sooner or later, different structural features can divert chemists' attention from the main road.

Let us first consider a molecule as a graph. Features characterizing this graph are points, lines and their mutual spatial relationship. We can find different kind of points (reducible or irreducible (9)) and of lines (open or closed); we can also find particularly "crowded" parts of the graph and lines differently space-oriented. The problem is the identification of what is really important for OSP.

We chose two aspects as weight determining: 1) point substitution; 2) point stereochemistry.

We created a function evaluating the complexity of a structure; it is:

$$(1)\quad CHI = SUM(i)(SIGMA(i) + RO\ (i))$$

where SIGMA(i) is the substitution level of atom (i) and RO(i) is the effective stereochemical contribution of atom (i). The additivity function is a fast way to evaluate structure weight.

SIGMA evaluation.

SIGMAs represent the substitution level of an atom, excluding those atoms that

are zero-weighting for one of the following reasons:

1) they are monovalent atoms (e.g. F, Cl, Br, I);
2) they are n-valent atoms but have n-& bonds with hydrogens, C and Si atoms excluded (e.g. OH, NH2, but not CH3);
3) they are n-valent atoms but have one n-i valent bond and (i) bonds with hydrogens, C atoms excluded (e.g. =O, =NH, but not C=CH2);
4) they are particular superatoms like NO2 or monosubstituted phenyls.

RO evaluation.

In considering stereochemical weight three different situations can be found:

1) isolated stereocenters;
2) stereocenters alfa or beta one to another, but not in the same ring;
3) stereocenters on a ring containing other stereocenters.

Each situation has its own weight, decreasing in the order cited.

Having defined a function for structure complexity evaluation, let us consider other principles useful in the search for the "best" convergent synthesis.

We defined the complexity distance between atom I and atom J as the minimum possible sum of atom weights onpath; i.e. the shortest complexity path between I and J.

We then evaluated the maxima of the minimum distances.

After this we had sufficient information to search for the complexity "centre" of a structure. This is the area of the structure in which we must operate to obtain precursors of similar weight and to enhance synthesis convergence.

Molecular centre definition and identification.

We defined the complexity centre of a molecule as the set of atoms situated half-way between the atoms at maximum distance. In this way, we identify exactly that part of the structure (here represented by a set of atoms) which is the centre as regards substitution, number of atoms, atom stereochemistries.

If we can elaborate a method to break the target structure at its centre, we should be certain to obtain precursors which are similar in complexity.

The atom set identified may be composed of different numbers of atoms, this identification depends on: the starting atom, the number of starting atoms, and the crowding of the structure. We must find the number of bond breaks that really separate the molecule and the method to operate their breaking.

Quantity and quality of breaks.

We defined the minimum number of breaks necessary to separate a molecule as the number of central atoms found starting from each peripherical point at maximum distance and moving towards the centre. This determines different numbers for different centre crowdings.

Furthermore, a correct solution can be defined that which breaks one bond spanning from each central atom, allowing only one double affixation (10).

Depending on the connectivity of the molecular centre, the bond breaks may generate two or more pieces. Following our strategical model, we decided to accept only two-piece solutions. The objective was to have at our disposal a completely consistent methodology which, level by level, leads the analysis towards a binary tree.

Minimal Sets of Strategic Bonds (MSSB).

The next step is the identification of the MSSBs. These are sets of bonds whose breaking separates the target into two precursors of similar complexity (as defined earlier). These sets are unrelated to reactivity.

Computer power is the ability of the computer to work on large quantities of data; thus the analysis can be extended to the whole family of sets, but as some bond sets are better than others we must consider them first; we therefore evaluate the "goodness" of a solution.

In our approach, the "value" of a solution must be a measure of the similarity of their components; the value of a solution is defined as the complexity difference of its parts. The smaller the difference the better the solution; the sorting of the results is accomplished in this way.

The algorithm.

Schematically represented this part of our program may be divided in four main sections: maximum distance evaluation and peripherical atoms identification; centre identification; MSSB filling; solution evaluation and sorting.

It will suffice, in this paper, to draw attention to only the principal features of the various parts.

A new and efficient method to calculate the maximum complexity distnce and to identify the pairs of atoms positioned at that distance has been realized.
This routine package is also used for ring perception (11). It is based on an original method for problem size reduction, allowing work on a reduced structure representation.

Also the search for pairs of atoms at maximum distance (MAXD) is an example of the application of rules to obtain solutions without evaluating the entire problem. In fact, on the crunched (12) structure the pairs at MAXD are identified during the ring perception procedure (in the absence of rings the search is a lot easier); then partial expansion of the structure to arrive at the correct pairs are made. We then assume the information present in the crunched structure is sufficient to furnish the correct points for the successive elaboration; the growth of complete trees from each atom can be avoided; instead, a few trees, swapped on the structure, are sufficient to identify the pairs of atoms at MAXD.

The results obtained are correct for our test structure set. (A check has been

made by growing all the possible trees and evaluating all the possible two atom distances) (13).

With the pairs of atoms at MAXD available, a search is made for the centre of the structure and, at the same time, the set of the middle atoms is filled. Partial trees, starting from atom in the pairs, are grown and a search made for successive spheres until reaching a distance equal to MAXD/2. Every atom found in the sphere will be inserted in the set of middle atoms.

A different problem is calculating the number of bond breaks sufficient to separate the structure into two parts.

Our approach has been very pragmatic. We think that if N atoms are found in the centre of the molecule, coming from the periphery, it is highly probable that N would be the connectivity of the centre. We assume that the breaks of N bonds in the centre will separate the molecule in parts. For each atom in each pair the number of breaks is determined. During the execution of the breaking one bond break is performed for each middle atom, allowing for one double affixation only. (Obviously a double affixation eliminates another bond break, thus maintaining the total number constant).

The next step is represented by the filling of the MSSB.

This is done by looping on the bonds spanning each central atom, for all possible combinations. A certain number of check controls operate on the solutions obtained.

The first check is on the number of pieces created by a break combination; unfragmented and multifragmented results are discarded.

The second check is on identical solutions with the consequent elimination of any overlapping.

The third check is on the bond order of the broken bonds; triple bonds being considered unbreakable and the breaks being moved towards alfa positions.

The routine works searching for good solutions. This implies that we try, as soon as possible, to prune the solution set, maintaining however all the acceptable solutions.

Results and discussion.

We report some results obtained with our algorithm. Different kinds of structures taken from the literature are used.

The results are partially reported to save space but the best are ranked. (The total numbers of solutions are indicated) (see TAB. 1)

At this point, some comments are called for. With regard to the strategical value of the solutions, it can be observed that the number of solutions proposed is easily manageable, thus allowing for profitable subsequent work.

The consistency of the solutions with the principles proposed is evident. The centre of the molecule is always correctly identified, and the fragments obtai-

ned are always similar in complexity. Looking at each structure, it is possible to suggest different results but the formalism of our approach is fully satisfied and we disregarded proposals outside it.

On the other hand, physical feasibility of the solutions is absent and we think that our objective regarding a rational strategy, neglecting physical soundness, has been obtained. Obviously a different section of the program should take care of the reactivity and other problems, so the chemical value of a solution will soon become part of the overall evaluation.

A further comment concerns the different weight that topology and stereochemistry have. Evaluation of a correct weight ratio of the two is a heuristic matter. If the ratio is good the solutions are good. Judgement, at this level, can be made only by synthesis experts (and their role of know-how transfer is basilar). Some of them consider the proposals of the algorithm "interesting". Is it good or not? It is probably too early to answer. Our feeling is positive because the proposed results are so clear and simple that they suggest a good heuristic pruning.

In conclusion let us comment on the feasibility of the solutions and on the relative importance of the strategy and the model.

We began our work with a postulate in mind: "If a chemist decides to operate a transformation, s/he can do it".

At present, we may discuss the postulate, but we are confident of its correctness. We do plan a transformation model, but think that the strategy must be as autonomous as possible. Only logical aspects of a synthesis must be considered; everything else will be taken care of by the model (and by the chemist, of course).

We wrote our algorithm in Fortran 77 and implemented it on an IBM 3083 machine, using the Fortvs compiler. Performance is quite good, considering the work done; debugging has reached the level of standard use.

Conclusion.

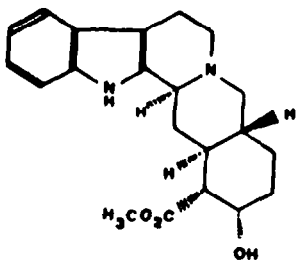
We have outlined a new algorithm, part of a program for Computer Assisted Organic Synthesis under development in our department.

The present level represents a first part of a full program. The algorithm is devoted to strategic bond selection in organic structures.
Results fulfil the basilar objectives stated at the beginning of the work.

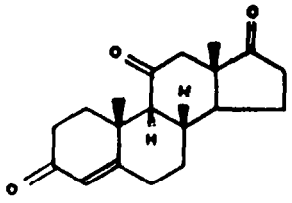
The attempt to develop an "intelligent" system to help chemists in OSP has been pursued as far as possible and, even if we are really far from an expert system, we feel we have paved part of the way.

The transformation model will be the subject of our future research developments and will be just as hard to realize as the present one. Eventually we hope to have at our disposal an instrument suggesting ideas ready for the laboratory.

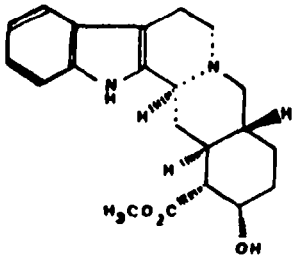
Acknowledgements. - Partial financial support by Tecnofarmaci S.p.A. is gratefully acknowledged. We appreciated helpful discussion with Profs. G. Jommi and F. Pellizzoni and F. Orsini. One of us (G. Sala) acknowledges a scholarship by Tecnofarmaci.

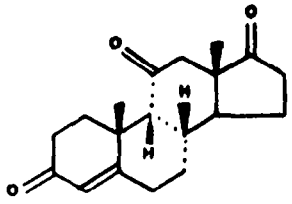


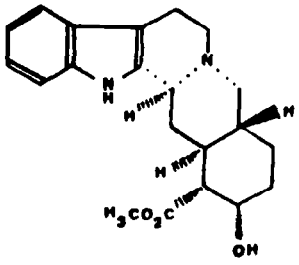
YOHIMBINE

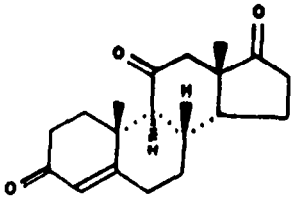


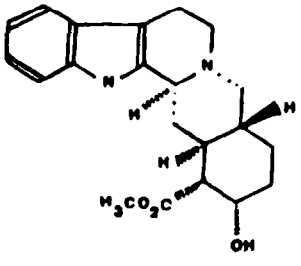
ADRENOSTERONE



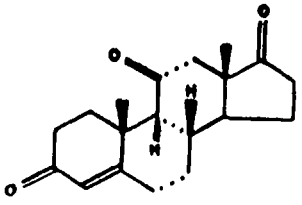






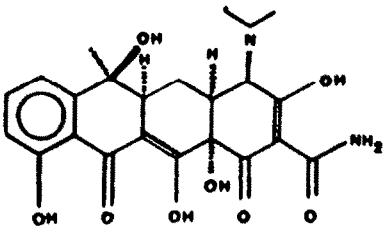


Total solutions = 6

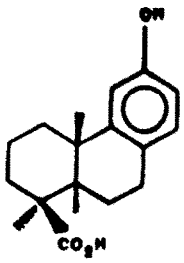


Total solutions = 10

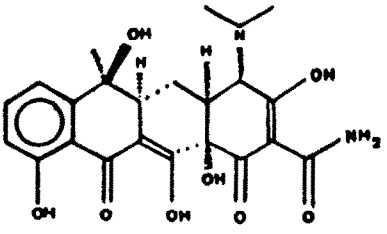
TAB. 1

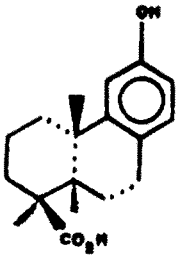


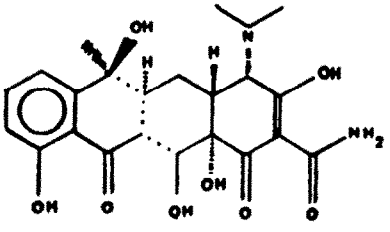
TETRACYCLINE

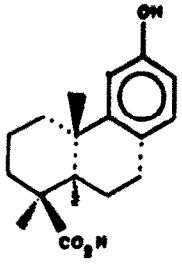


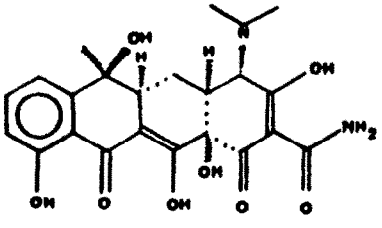
PODOCARPIC ACID



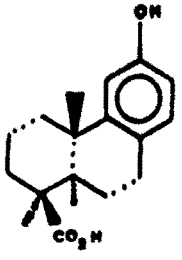






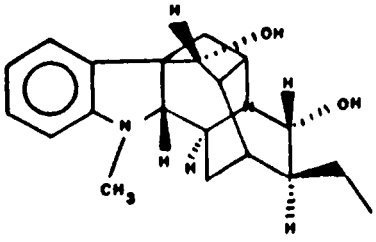


Total solutions = 4

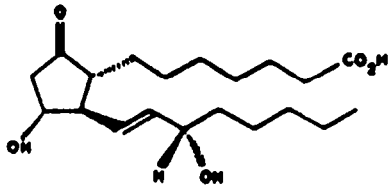


Total solutions = 7

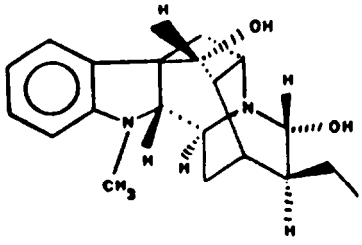
TAB. 1 (contd)

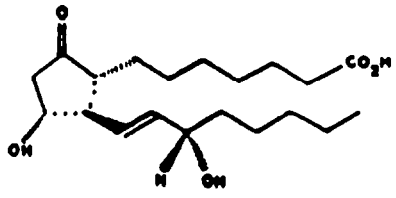


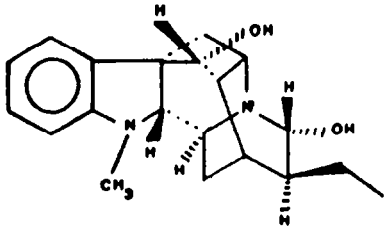
AJMALINE

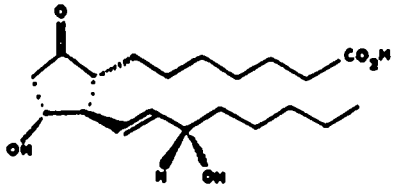


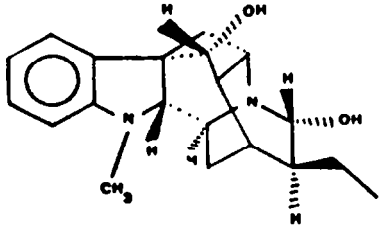
PROSTAGLANDIN E1



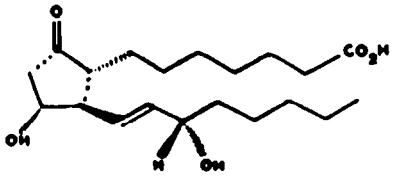






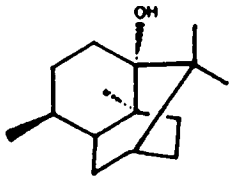


Total solutions = 4

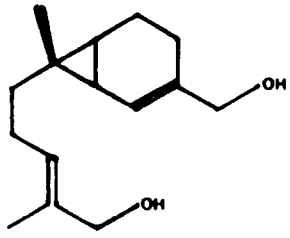


Total solutions = 5

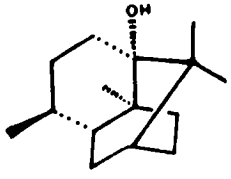
TAB.1(contd)

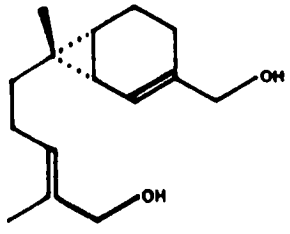


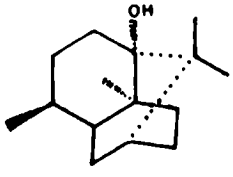
PATCHOULI ALCOHOL

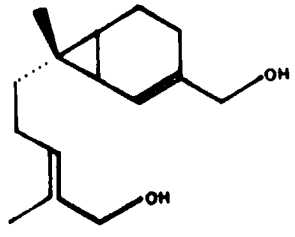


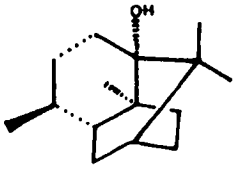
SIRENIN











Total solutions = 3

Total solutions = 2

TAB.1(contd)

References.

1) E.J. Corey, Pure Appl. Chem., 14, 19, (1967); E.J. Corey, W.T. Wipke, Science, 166, 78, (1969).
2) J. Gasteiger et al., Topics in Curr. Chem., 137, 19, (1987); J.B. Hendrickson et al., J. Chem. Am. Soc., 107, 5228, (1985); W.L. Gelertner et al., Anal. Chem. Symp. Ser., 15, 35, (1983); E.J. Corey et al., J. Org. Chem., 50, 1920, (1985); W.L. Jorgensen et al., J. Org. Chem., 50, 4490, (1985); W.T. Wipke et al., J. Chem. Info. Comput. Sci., 24, 71, (1984); and references cited.
3) J.B. Hendrickson, E. Braun-Keller, G.A. Toczko, Tetrahedron, 37, Supp. 1, 395, (1981).
4) See ref. 3, note 7.
5) Lilith is not an acronym, but the name of a feminine daemon.
6) This is true in Data-Base based programs. (LHASA, SECS and others).
7) This is not completely true, because the tree pruning occurs at the very end of the program execution, but this does not seem to be very efficient.
8) For a deeper discussion concerning convergence and complexity, see S.H. Bertz, J. Am. Chem. Soc., 104, 5801, (1982).
9) A.T. Balaban, P. Filip, T.S. Balaban, J. Comput. Chem., 6, 316, (1985).
10) A double affixation is the making of two bonds spanning the same atom. (See ref. 3).
11) Submitted for publishing.
12) The meaning of "crunched" is related to the action of reducing molecular size.
13) We usually determine a fixed maximum number of pairs. The use of more pairs seems to be unimportant in searching for the middle atoms.