Tetrahedron. Vol. 34, pp. 841 to 852. Pergamon Press 1978. Printed in Great Britain

COMPUTER-ASSISTED STRUCTURE MANIPULATION

STUDIES IN THE BIOSYNTHESIS OF NATURAL PRODUCTS[1,2]

TOMAS H. VARKONY*, DENNIS H. SMITH and CARL DJERASSI

Contribution from the Departments of Chemistry and Genetics, Stanford University, Stanford, CA 94305, U.S.A.

(*Received in U.S.A.* 1 *August* 1977; *Received in U.K. for publication* 18 *October* 1977)

Abstract—A computer program ("REACT") that simulates chemical reactions by manipulating representations of structures has been utilized to study biochemical reactions. Biosynthetically plausible sterol side chains were generated and the number of plausible isomeric sterols were calculated. Possible label distributions in structures resulting from the biochemical conversion of a humulene precursor to some known fungal metabolites were followed by studying different biochemical pathways.

Manual elucidation of structures arising from chemical reactions which may yield a large number of products via a number of complex, interrelated pathways is a difficult problem. Such reactions are, however, natural candidates for computer-assisted studies because the computer can easily record all intermediates and products as well as interrelationships among them.[3-6] Examples of these reactions include carbonium ion rearrangements, reactions of free radicals and biochemical processes.

We developed the REACT program[3,4] initially as an extension to the CONGEN program[9] for computer-assisted structure elucidation. Although a detailed description of REACT[7] is beyond the scope of this presentation, a brief summary of the program is necessary for comprehension of the present work. REACT is designed to carry out representations of chemical reactions on representations of chemical structures. Reactions, defined by the chemist using the program, are carried out in the synthetic direction as opposed to the retro-synthetic direction of programs for computer-aided synthesis.[8] In structure elucidation problems, the set of structures undergoing reaction is the current set of candidate structures for an unknown. It was soon clear, however, that the program could also be used effectively in following reactions of a single, known compound participating in a complex sequence of reactions. In the examples given in this work, a single precursor was subjected to repetitive application of a set of reactions.

In this paper we demonstrate the utility of REACT in two examples where a given precursor of known structure is subjected to an extended sequence of reactions. At each step in the sequence one or more reactions may apply to the products from the previous step. As will be shown in the sequel such an approach is especially well suited to problems involving the biosynthesis of natural products.

Generation of biosynthetically plausible sterol side chains

Method. Sterols are naturally occurring steroidal alcohols (usually 3β-ols) which differ in the number and the position of Me groups and the degree of unsaturation (present as a double bond or cyclopropyl ring). New sterols are frequently isolated in minute quantities from natural sources. Because of their structural similarities and the large number of different sterols present as a mixture in the same source (a recent paper[10] documents the isolation of *ca.* 50 sterols from one marine source) it is often difficult to separate them and to obtain pure compounds in quantities large enough for structure determination by conventional methods. Some structural assignments are based on biogenetic considerations, assuming that compounds from the same origin are related to each other through formation along the same biochemical pathway. This pathway can be a series of complicated chemical reactions which yield a large number of intermediates and products. It is difficult to follow manually such a series of reactions in order to explore all possible structural alternatives. To date, over 100 different 3-hydroxysterols have been isolated, the majority of them based on the seven nuclear skeletons[11]

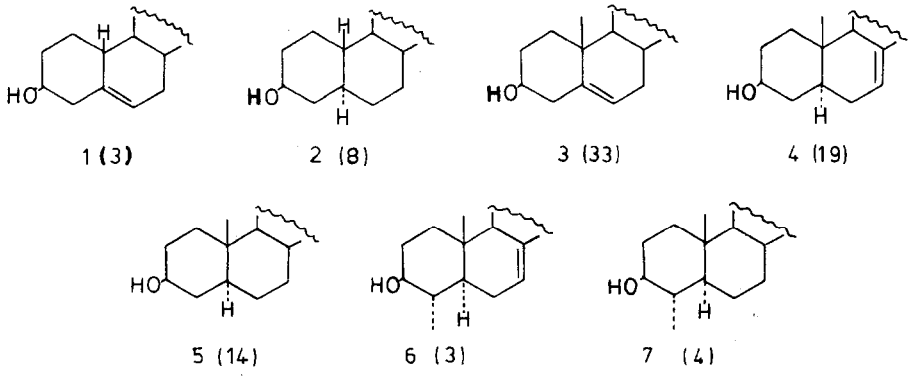


Fig. 1. Sterol nuclei of some naturally occurring sterols. The numbers in parentheses refer to the number of naturally occurring sterols, known to contain that skeleton.

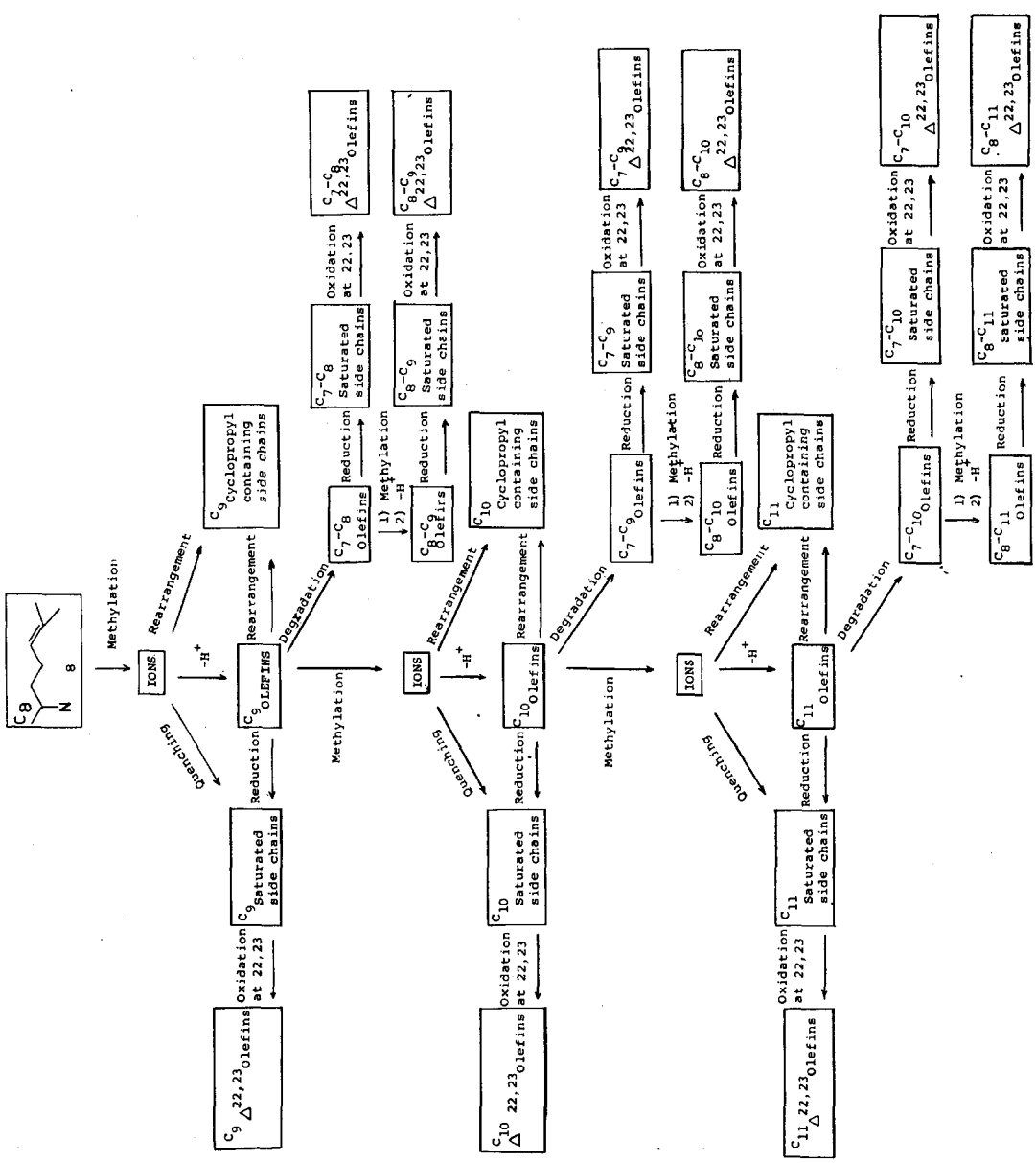


Fig. 2. Schematic representation of the computer assisted "Biosynthesis" of C7–C11 sterol side chains.

1–7 (Fig. 1) and a variety of C-17 alkyl side chains (some of the known side chains are marked in Figs. 3–8).

We use a method of combined gas chromatography/mass spectrometry (GC/MS) to analyze complex mixtures of sterols in a search for new compounds which may represent important biosynthetic intermediates. Part of this method involves research in interpretation and prediction of mass spectra.[12] We have used the REACT program as an additional tool to predict plausible structural candidates to guide both our manual and computer-based interpretations.

The set of reactions used in REACT to carry out possible transformations of sterol side chains have been suggested previously.[13] The precursor **8**, (a 24,25 unsaturated side chain) the order of application of the various reactions and the classes of products which result are shown in Fig. 2. The sequence of reactions consists of repetitive application of the following steps:

(1) *Methylation*

C-methylation of a double bond. In nature this reaction occurs via the ylide of S-adenosylmethionine. This reaction is constrained for general application later in the sequence to forbid the sterically unfavourable methylation of tetra-substituted double bonds.

(2) *The carbonium ion obtained by the alkylation can undergo several reactions*

(a) Proton elimination and formation of a double bond.

(b) Cyclization to form a cyclopropyl system with subsequent elimination of a proton.

(c) Quenching to form saturated side chains.

(3) *The olefin is allowed to undergo several additional reactions*

(a) Reduction to form a saturated side chain.

(b) Rearrangement to a cyclopropyl system.

(c) Degradation to shorter side chains via loss of allylic Me groups.

(d) Methylation to produce longer side chains.

Constraints on reactions of the olefin included

(a) Subsequent migration of the double bond is not allowed.

(b) Olefins obtained by degradation are allowed to undergo only one step of methylation.

(4) Subsequent oxidation of saturated side chains proceeds to form a new double bond at C-22, 23, a mechanism proposed by Knapp *et al.*[14]

This set of reactions was applied sequentially as shown in Fig. 2 a total of three times. Thus, side chains possessing from seven to eleven carbon atoms are accessible by this sequence.

Results

The different side chains generated by REACT are shown in Figs. 3–8. The letters M and P refer to side chains which have already been encountered in nature. M standing for marine and P for plant sources. Encircled structures are examples of interesting "missing links". If such structures are encountered in nature an important step will have been taken in demonstrating the existence of such alternate biosynthetic pathways.

A numerical summary of these results is presented in Table 1. The table is organized by summarizing the side chains produced by the different biochemical pathways. The only known, naturally occurring C7 saturated side

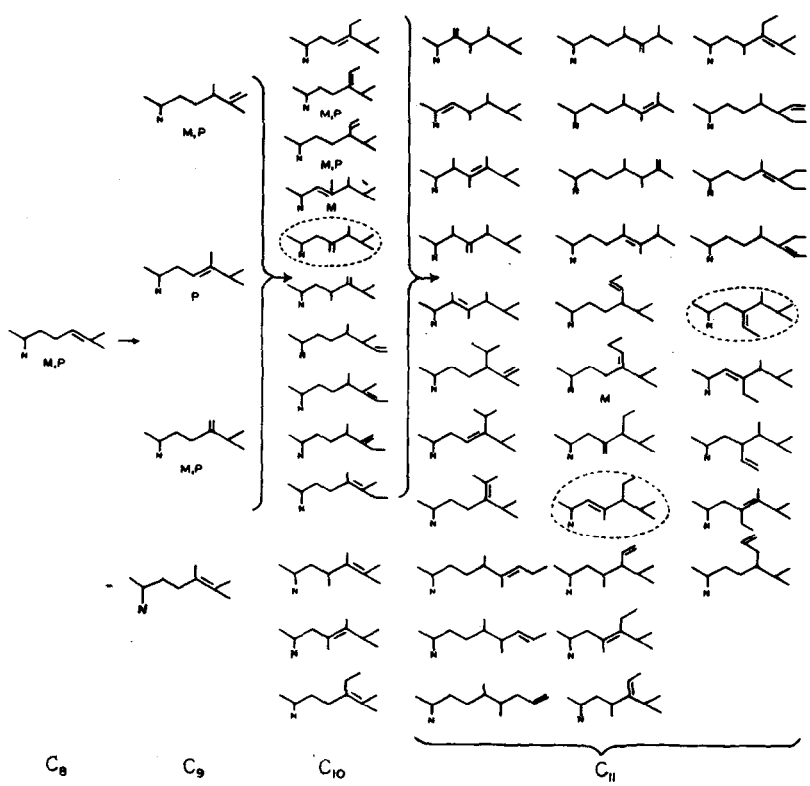


Fig. 3. Computer generated mono olefinic side chains. Methylation only.

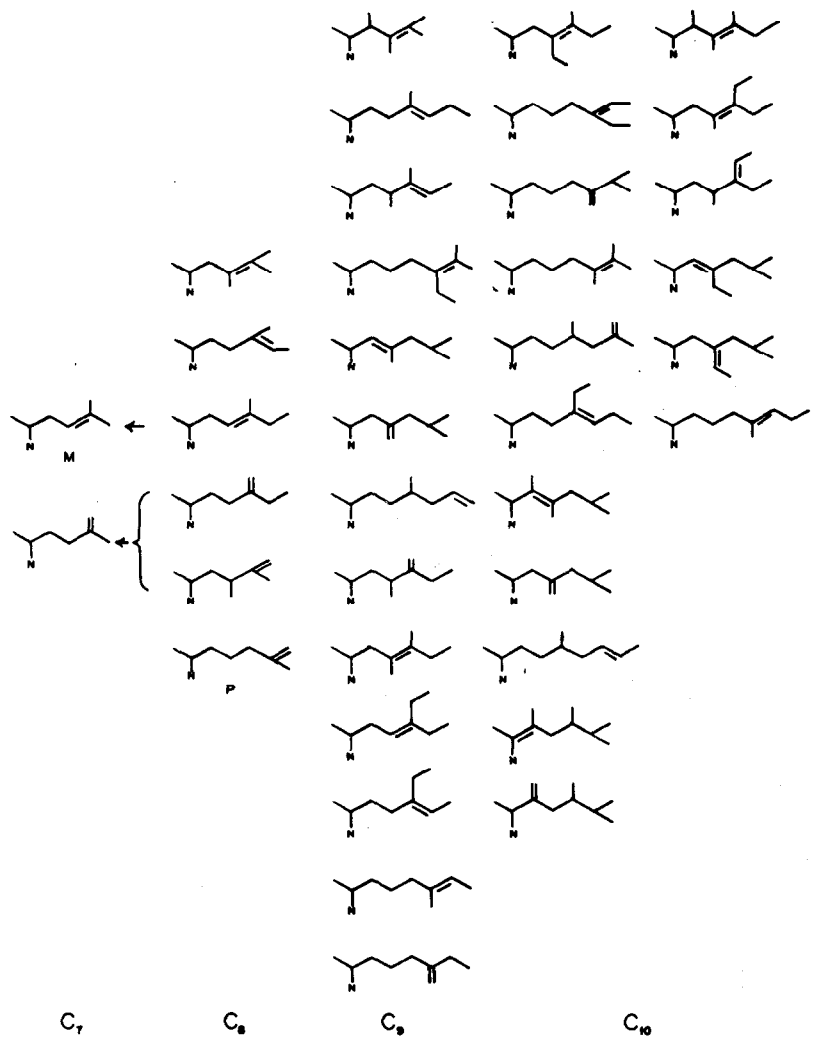


Fig. 4. Computer generated mono olefinic side chains. Degradation only.

Table 1. Number of isomeric Hydrocarbon side chains of sterols obtained by biosynthetic pathways

Number of C in side chains	SATURATED				OLEFINS					CYCLOPROPANES		
	A	B	E	Nature	A	B	E	F	Nature	C	D	Nature
7	-	1	-	1	-	2	-	1	2	-	-	-
8	1	2	-	1	1	6	3	2	5	-	-	-
9	1	7	-	1	4	13	6	4	4	2	4	-
10	3	12	-	4	13	17	19	8	6	8	11	1
11	8	-	8	1	31	-	37	8	1	17	21	1

A methylation only.
B methylation followed by degradation only.
C rearrangement of carbonium ion.
D rearrangement of olefin.
E degradation followed by methylation only.
F oxidation of saturated side chains at 22,23 position.

chain was correctly predicted by REACT. Three C7 unsaturated side chains were predicted. Two of these three exist in nature. In the C8 series five unsaturated side chains out of 12 predicted are observed in nature. For the longer side chains, more are possible but fewer are observed. For example, only one out of the 76 predicted C11 side chains has so far been found in nature.

By combining the results from Table 1 with the seven most common steroidal skeletons 1–7 (Fig. 1) we obtain the minimum number of possible structural isomers, which are ordered by molecular weight in Table 2.

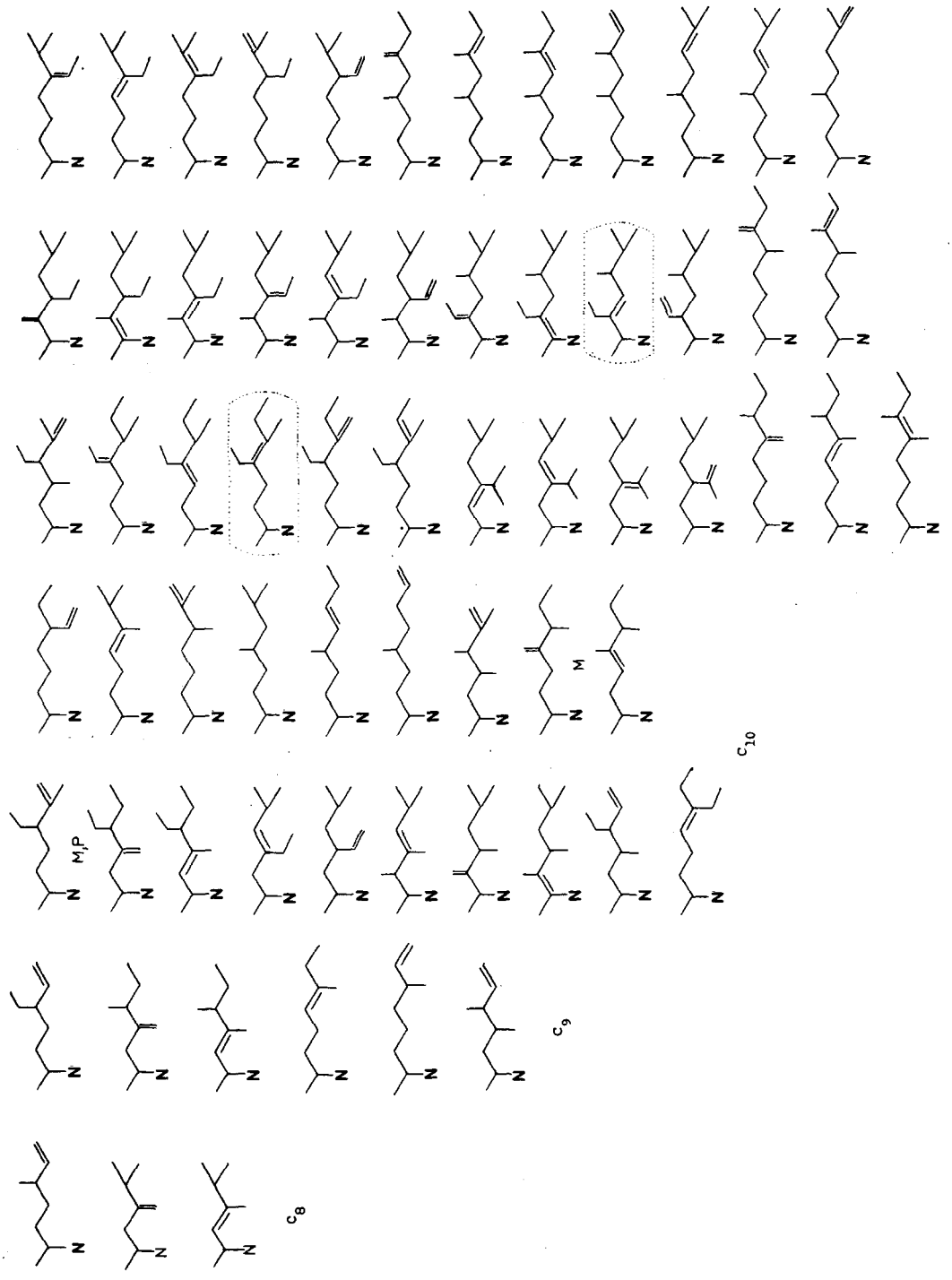


Fig. 5. Computer generated mono olefinic side chains. Methylation following degradation only.

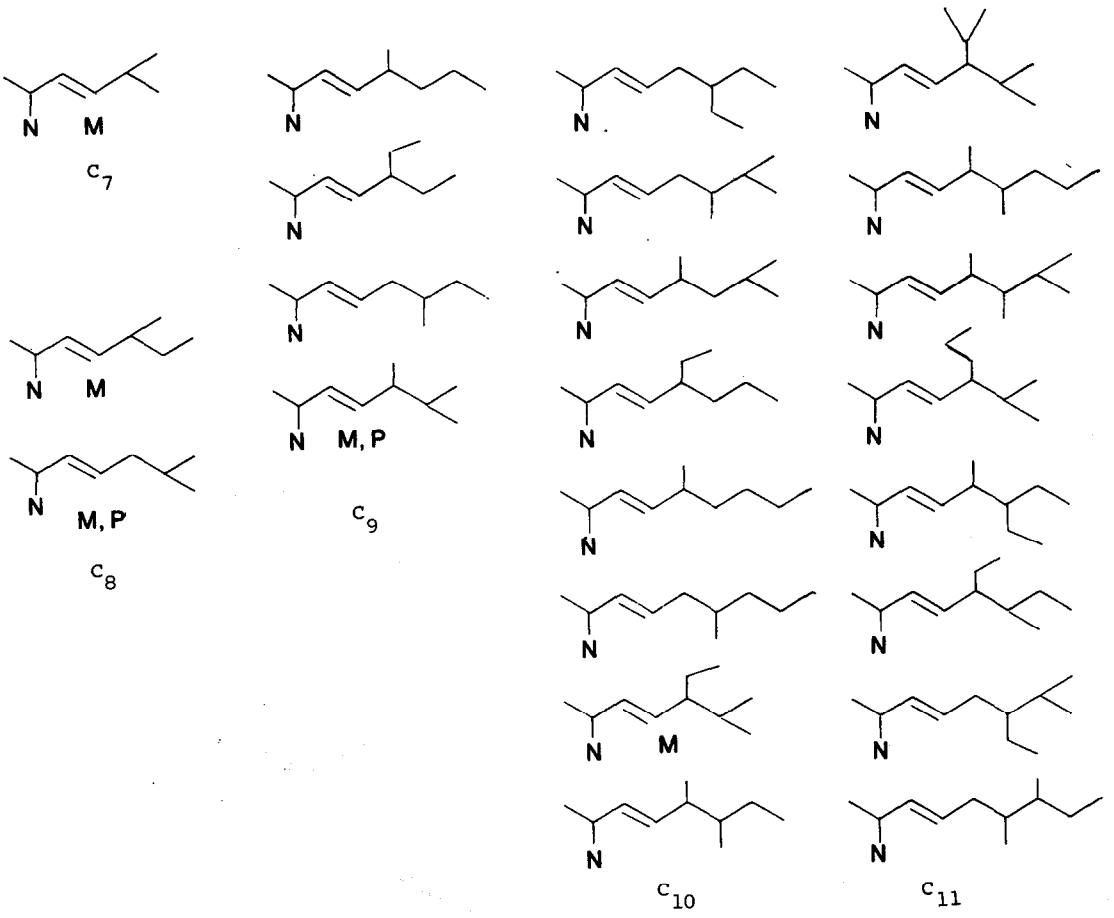


Fig. 6. Computer generated mono olefinic Δ^{22} side chains.

Table 2. Isomeric distribution of computer generated sterols

Empirical Formula	Molecular Weight	Saturated #	Olefins #	Cyclopropanes #	Total #
C31H560	444	16	-	-	16
C31H540	442	-	92	21	113
C31H520	440	-	76	21	97
C30H540	430	31	-	-	31
C30H520	428	47	133	32	212
C30H500	426	-	209	53	262
C29H520	416	39	-	-	39
C29H500	414	54	160	36	250
C29H480	412	-	217	47	264
C28H500	402	26	-	-	26
C28H480	400	34	96	15	145
C28H460	398	-	123	19	142
C27H480	388	12	-	-	12
C27H460	386	15	42	4	61
C27H440	384	-	54	4	58
C26H460	374	4	-	-	4
C26H440	372	5	15	-	20
C26H420	370	-	18	-	18
C25H440	360	1	-	-	1
C25H420	358	1	3	-	4
C25H400	356	-	3	-	3
TOTAL		285	1241	252	1778

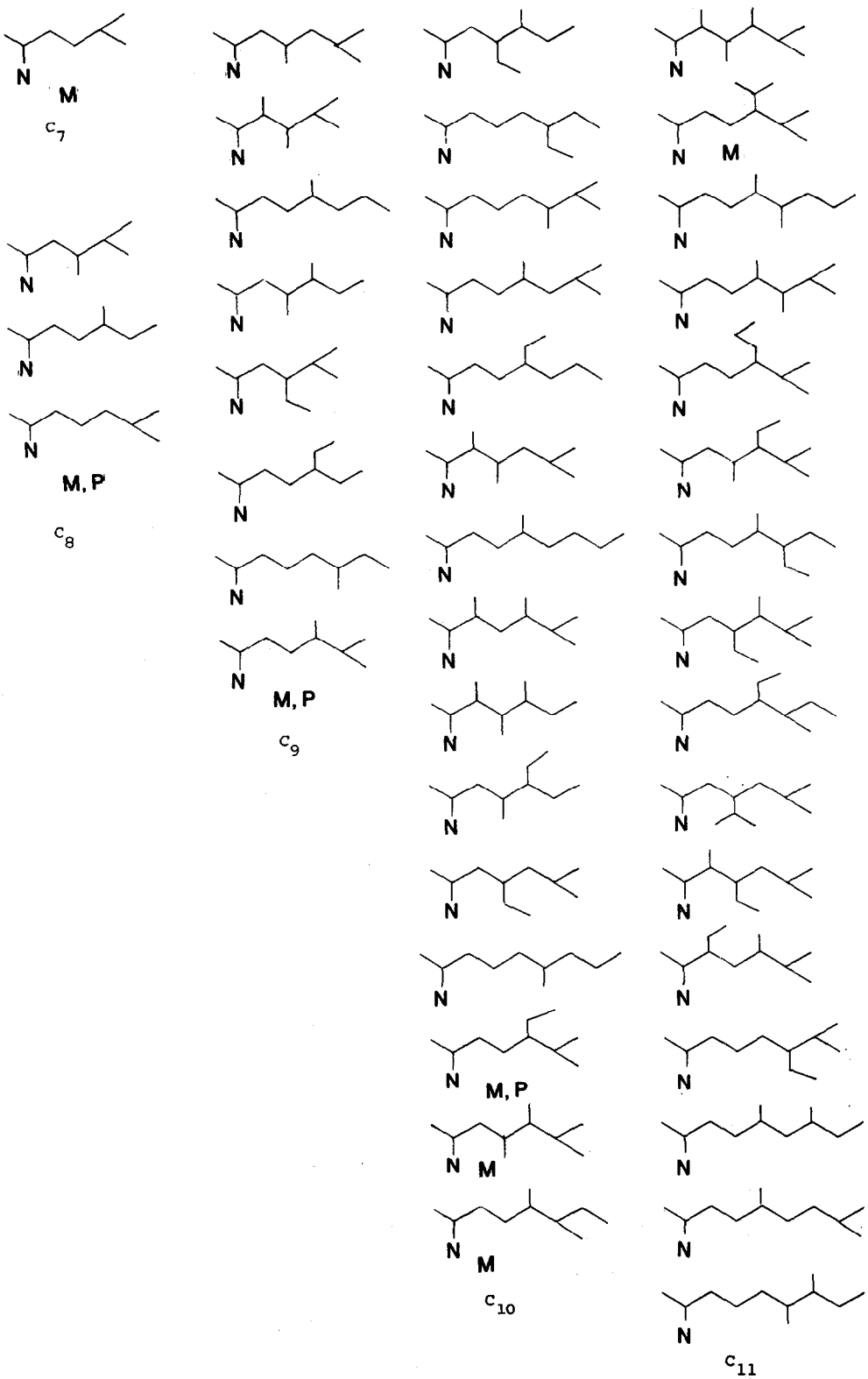


Fig. 7. Computer generated saturated side chains.

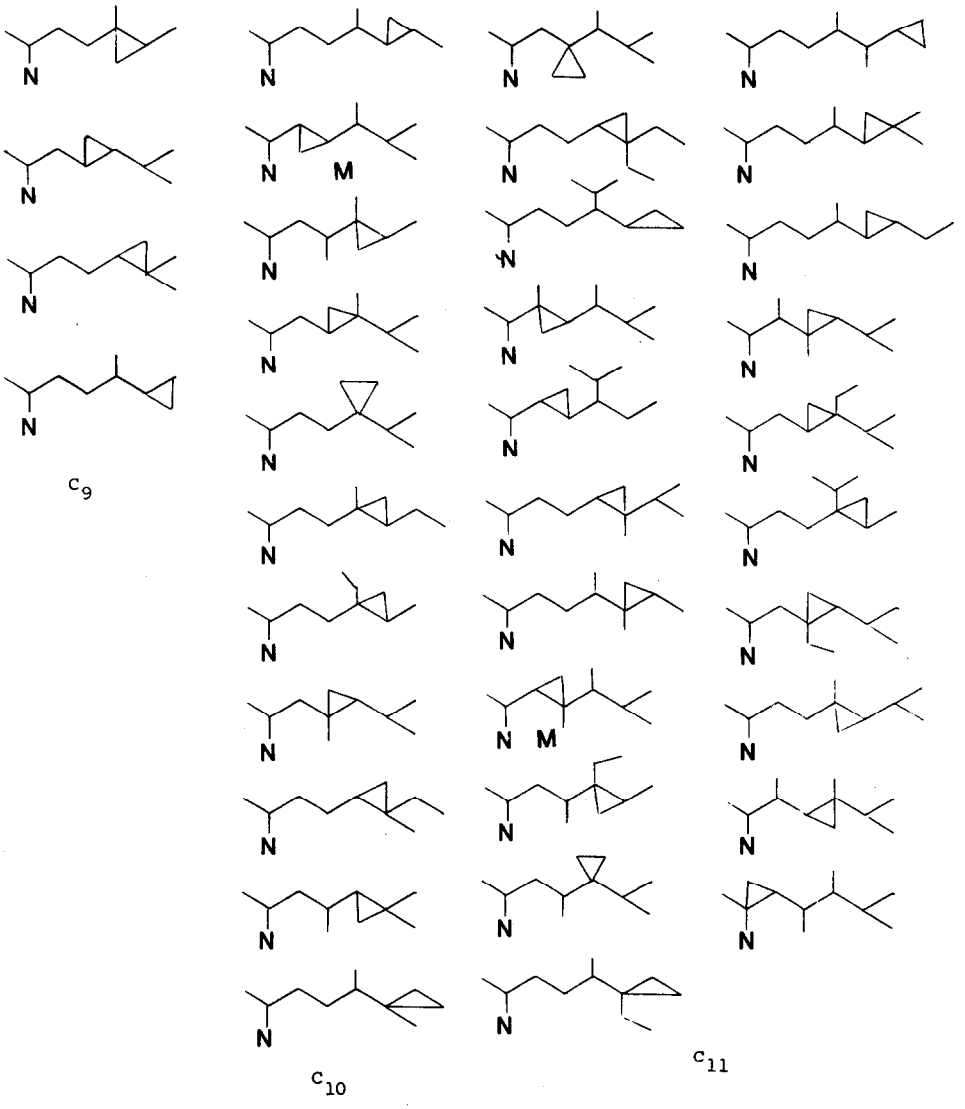


Fig. 8. Computer generated cyclopropyl containing side chains.

The total number of sterols which obey our biosynthetic constraints is 1778. This number is manageable by techniques of computer-assisted structure elucidation. Separating the structures by molecular weight reduces considerably the number of candidate structures which must be considered in a given problem. Thus, in a GC/MS experiment the maximum number of structures we have to consider is not larger than 264 (the number of isomers with empirical formula C29H480). Any additional spectroscopic or chemical data reduce this number still further. For other molecular weights the number of possibilities is considerably fewer. Structural information from the mass spectral fragmentation pattern of the molecule may leave only a small number of possibilities from which to choose.

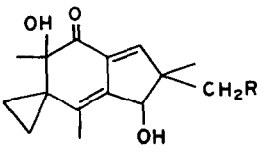
Elucidation of biosynthetic pathways via modelling of labelling experiments

Method. Elucidation of biosynthetic pathways can be accomplished in several ways, including for example co-occurrence of structurally related compounds or use of mutant organisms which accumulate intermediates.[15] These methods usually leave the structures of intermediates and/or the details of the biochemical pathways open to question. More detailed experiments are required to establish rigorously reaction pathways from precursor to product.

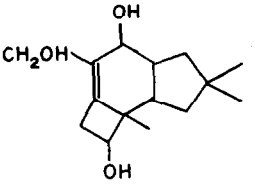
Isotopic labelling experiments are capable of providing additional detail through synthesis of labelled precursors followed by incorporation of labelled substrate and determination of the labelling pattern of the products of biochemical transformation. The incorporation of labelled precursors into desired products is generally low and elucidation of the labelling pattern in minute amounts of product is difficult. Thus, these experiments are generally time consuming and costly. They can be complicated by the existence of different biochemical pathways, some of which yield products with the same distribution of isotopic labels. Therefore, care must be used in designing such experiments. It is important to select a labelled precursor that will allow one to distinguish among most of the possible pathways, and that will lead to a product with labels distributed in easily detectable positions. Manual methods are often

insufficient to determine all the theoretically possible pathways when the number of possible pathways and the number of intermediate structures is very large. However, this type of problem is easily managed by REACT, which can accurately and systematically monitor transformations of the precursor into products, follow the isotopic labels throughout a reaction sequence and detect the formation of equivalent structures and labelling patterns. We stress that this is not an exercise in "paper chemistry", but a systematic way to investigate all the possible aspects of a proposed experiment before devoting valuable time and resources to an experiment which leads to ambiguous results.

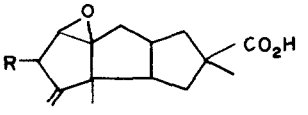
An example which illustrates our method is the exploration of biosynthetic pathways leading to formation of a family of fungal metabolites[16] which includes illudin-S **9**, illudin-M **10**, illudol **11**, hirsutic acid **12**, complicatic acid **13**, marasmic acid **14** and coriolins **15-17**.



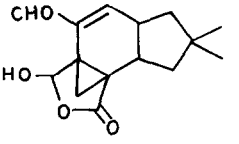
9: ILLUDIN S R = OH
10: ILLUDIN M R = H



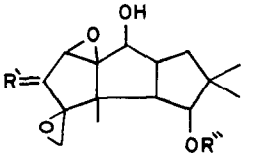
11: ILLUDOL



12: HIRSUTIC ACID R = OH
13: COMPLICATIC ACID R = O



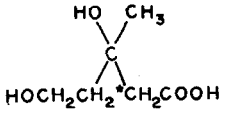
14: MARASMIC ACID



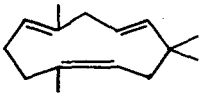
CORIOLINS

15: R = O, R = H
16: R = OH, H, R = COC_7H_{15}
17: R = O, R = $COCHOHC_6H_{13}$

Initial biosynthetic work with [2-14C]-mevalonate **18** demonstrated incorporation of the labeled precursor into **9, 10, 11** and it was suggested[16] that the humulene skeleton **19** is an important intermediate.



18



19

In studies[17,18] of the biosynthesis of hirsutic acid **12** and complicatic acid **13**, labeled [1-13C]- and [2-13C]-acetate were introduced as precursors. 13C NMR spectroscopy was used to determine which carbon atoms in the products were enriched by 13C. The observed pattern of labelling[17,18] suggests that a humulene precursor undergoes a further cyclization followed by 1,2-alkyl (or hydrogen) shifts to yield the skeletons of hirsutic **12** and complicatic **13** acids. 13C NMR spectroscopy was also utilized in the elucidation of biosynthetic pathways leading to the coriolins[19] **15-17**. The method used in this experiment was the incorporation of [1,2-13C]-acetate and the observation of 13C–13C coupling in labeled coriolins. This single experiment served to select the correct pathway from among three which were previously suggested.[18]

Results

We used REACT to simulate the biochemical conversion of the labelled humulene **20** precursor to the different skeletons which are presumed to arise via humulene. As a starting material we used the labelled humulene skeleton shown in Fig. 9, with a label distribution which results from incorporation of [1-*C]-acetate (*C can be either 13C or 14C). We followed earlier suggestions[18] and assumed that the first step involves cyclization to form the dimethylcyclopentane subunit, as shown in Fig. 9. The resulting structure **21** was subjected to a sequence of reactions detailed in Fig. 9, initiated by a double cyclization (Fig. 9) to yield tricyclic carbonium ions. The ions were allowed to undergo one or two steps of rearrangement. At each step, any of the possible Wagner-Merwein 1,2-alkyl or 1,2- hydride shifts triggered by the carbonium ion were allowed to occur.

The cyclizations yielded eight carbonium ions (Fig. 9). The first rearrangement step produced 41 ions, which on quenching the carbonium ion site yielded 26 unique skeletons. (Structures which differed only in the placement of the positive charge are removed as duplicates by REACT after quenching.) The second step of rearrangement yielded 191 ions which on quenching yielded 136 structures. The REACT program assisted us in determining that 26 of the 136 formally retain an all head-to-tail isoprene linkage, and 14 of the 26 can be

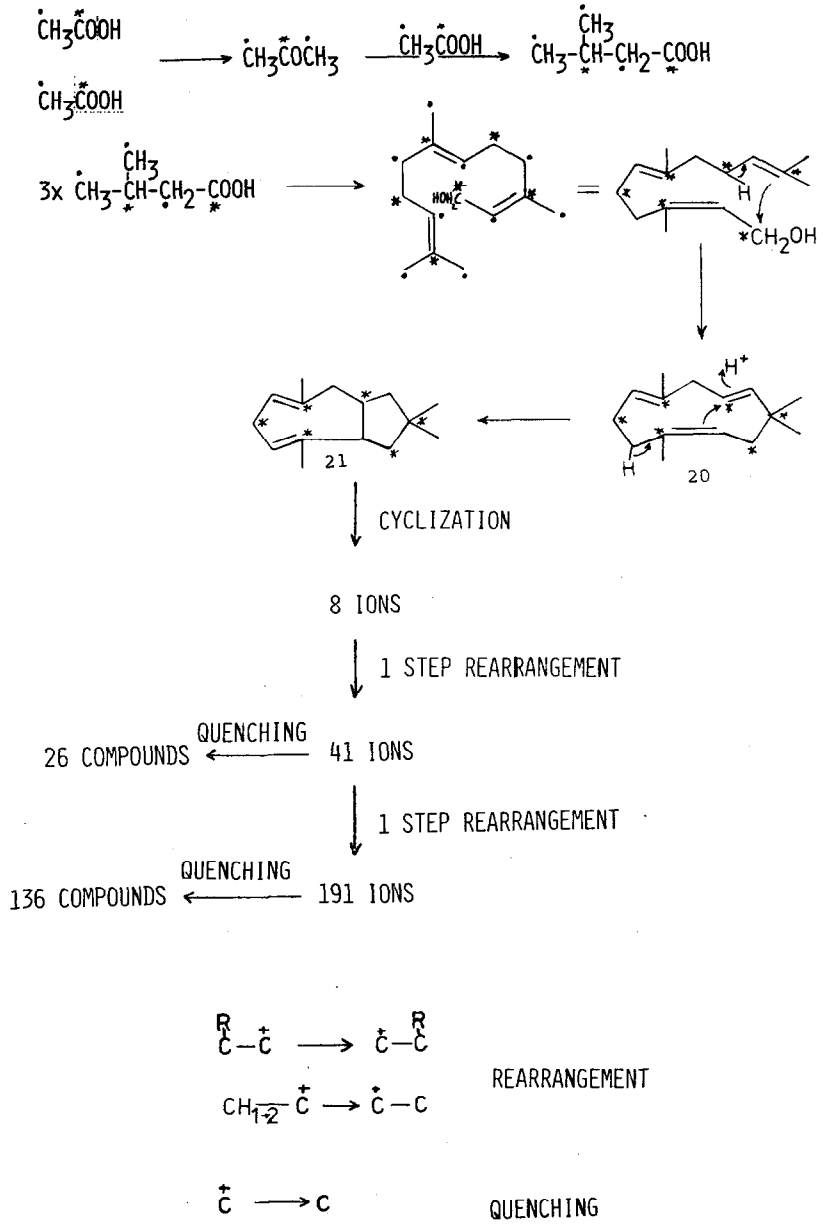


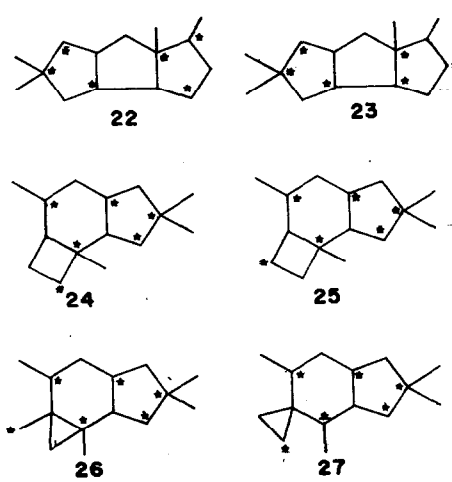
Fig. 9. Formation of tricyclic-sesquiterpene skeletons from labeled acetate through the labeled humulene skeleton.

obtained directly by cyclization of a labelled farnesyl precursor without further skeletal rearrangement.

In this single experiment with REACT we can determine whether members of the family of structures **9-17** can be obtained by the proposed reaction sequence and determine the pathways leading to each one. Six skeletons **22-27** produced by REACT formally represent the four carbocyclic skeletons of the structures **9-17** we chose for investigation.

Thus, the sequence of reactions is sufficient to predict at least one pathway to each of the structures of interest.

Two skeletons **22** and **23**, which differ only in the label distribution are related to the structures of hirsutic and complicatic acids; the same holds for the illudol skeleton (compare **11** with **24, 25**). Thus, even before investigation of pathways of formation we can already determine how many different distributions of label in a product are possible for a given set of reactions.



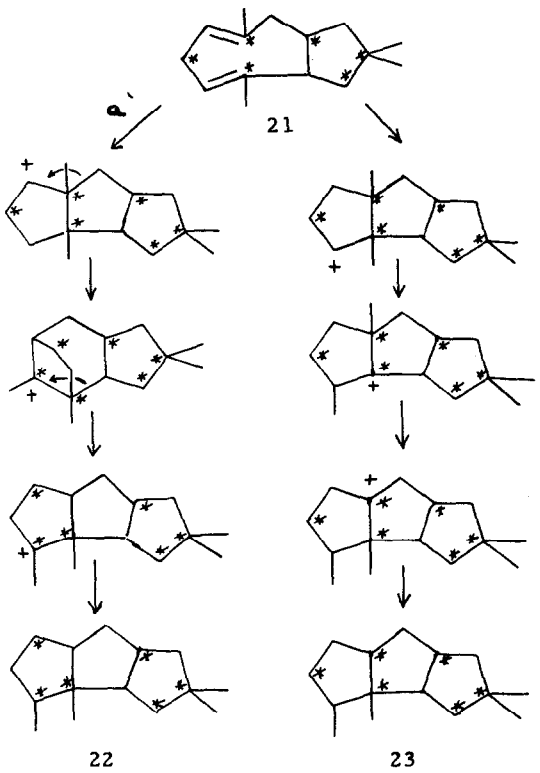


Fig. 10. Computer generated biosynthetic pathways for the formation of the skeletons of Coriolins (15–17) and Hirsutic (12) and complicatic (13) acids.

Considering now pathways for conversion of **21** to **22** and **23**, the carbocyclic skeletons corresponding to hirsutic **12** and complicatic **13** acids and coriolins **15-17**, can be obtained from two different pathways under our reaction constraints, each of which yields a unique label distribution (Fig. 10). Previous experiments[19] demonstrated that path **a**, leading to skeleton **22**, is the correct one. The remaining structures **9-11**, and **14** have not been studied as thoroughly. Our use of REACT to study pathways of cyclization and rearrangement of **21** predicts the following. The marasmic acid skeleton **26** is obtained by only one pathway under our constraints, and should have the label distribution shown in Fig. 11. The illudin skeleton **27** can be obtained from four different pathways as shown in Fig. 11. Each pathway produces the same skeleton with the same label distribution. Thus, this particular choice of labelled precursor would be a good choice for studying formation of marasmic acid, but would be a poor choice for a definitive answer about the pathway to illudin.

Two differently labelled skeletons **24** and **25** (Fig. 12), representing the illudol structure **11** are obtained via the pathways shown in Fig. 12. One skeleton **25** is obtained on simple cyclization of the precursor. A more complex rearrangement (Fig. 12) yields the same skeleton but with a different label distribution **24**. The absence of other pathways indicates that a laboratory experiment with labeled acetate should provide a determination of the correct pathway.

CONCLUSION

We have shown that emulation of cyclization and rearrangement reactions in a computer program can

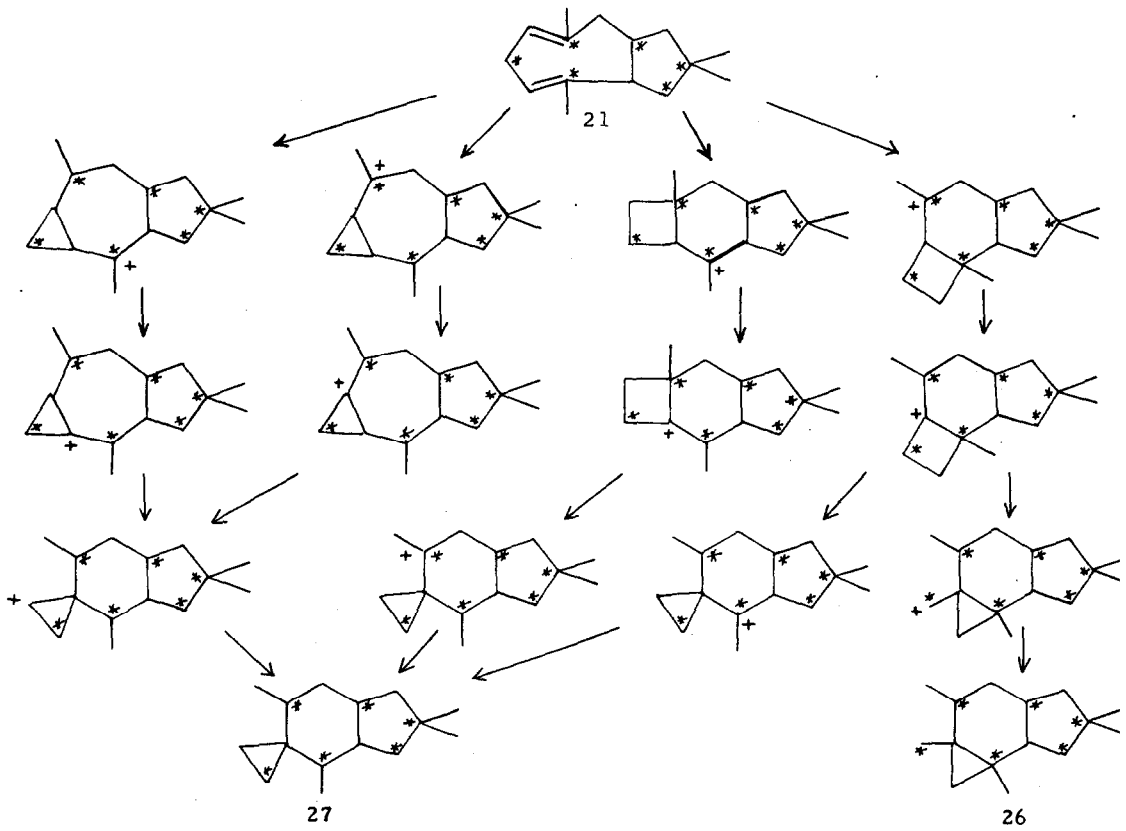


Fig. 11. Computer generated biosynthetic pathways for the formation of the skeletons of Marasmic acid (14) and Illudins (9-10).

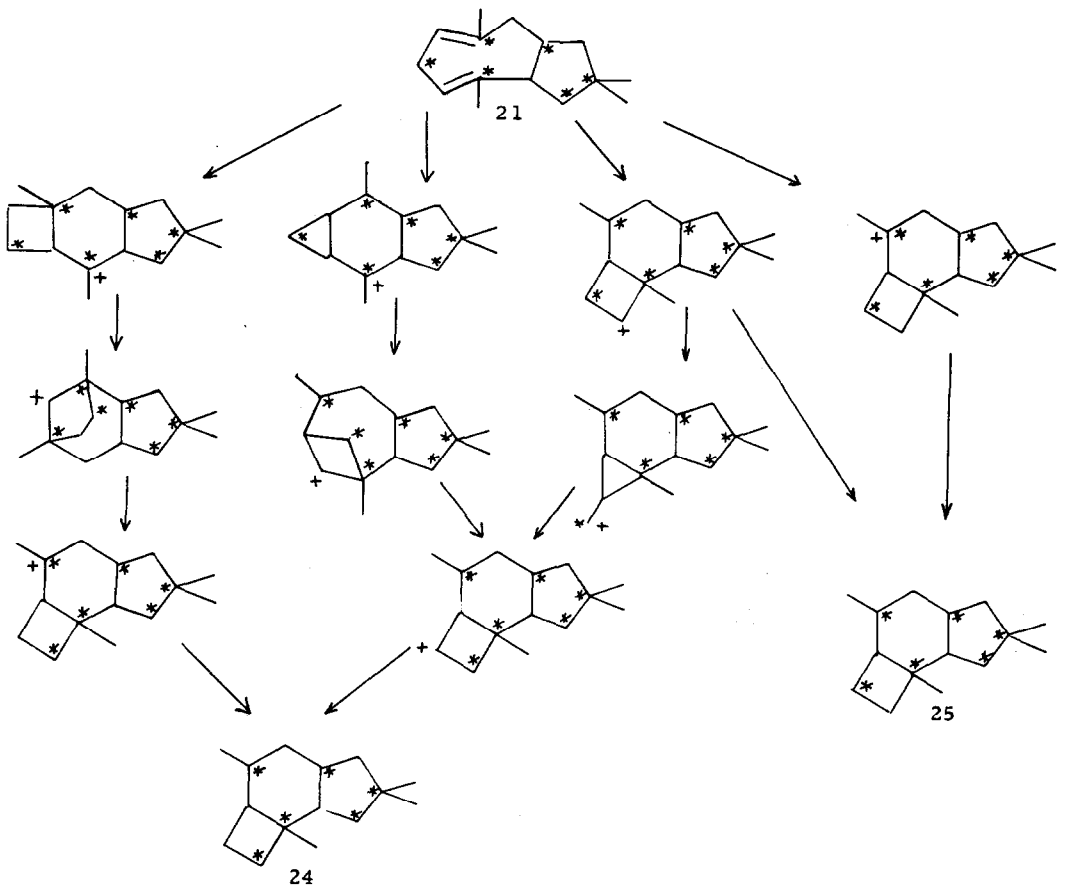


Fig. 12. Computer generated biosynthetic pathways for the formation of the skeleton of Illudol (11).

provide useful information for the study of biosynthetic reactions. The computer is especially useful for those problems where a complex sequence of reactions can potentially produce a myriad of products. Observation in nature of a particular product together with the results of REACT permit elucidation of pathways of formation. Alternatively, the set of products produced by REACT represents a set of plausible structures on which to base a search for new compounds.

REACT should have a significant impact on design of experiments. The program can be used as a "dry" laboratory to test hypotheses. Labelling experiments are expensive and time consuming. A computer-assisted preview of the different, plausible biochemical pathways can help the chemist in choosing a labelled precursor which will provide the maximum information in a real experiment.

REFERENCES

[1]Part XXVII of the series: *Applications of Artificial Intelligence for Chemical Inference.* For Part XXVI see G. M. Schwenzer, *J. Org. Chem.* in press.

[2]We wish to thank the National Institute of Health for the support of this work [RR 00612 and GM 20832 and GM 06840] and for the support of SUMEX computer facility [RR 00785] on which the REACT program is available to interested persons.

[3]T. H. Varkony, R. E. Carhart and D. H. Smith, *Computer-Assisted Organic Synthesis* (Edited by W. T. Wipke), American Chemical Society, Washington, District of Columbia, p. 188 (1977).

[4]R. E. Carhart, T. H. Varkony and D. H. Smith, *Computer-Assisted Structure Elucidation* (Edited by D. H. Smith), American Chemical Society, Washington, District of Columbia, p. 126 (1977).

[5]T. M. Gund, P. v. R. Schleyer, P. H. Gund and W. T. Wipke, *J. Am. Chem. Soc.* **97**, 743 (1975).

[6]S. A. Godleski, P. v. R. Schleyer, E. Osawa, Y. Inamoto and Y. Fujikura, *J. Org. Chem.* **41**, 2596 (1976).

[7]R. E. Carhart, T. H. Varkony and D. H. Smith, unpublished results.

[8]E. J. Corey and W. T. Wipke, *Science* **166**, 178 (1969).

[9]R. E. Carhart, D. H. Smith, H. Brown and C. Djerassi, *J. Am. Chem. Soc.* **97**, 5755 (1975).

[10]S. Popov, R. M. K. Carlson, A. Wegmann and C. Djerassi, *Steroids* **28**, 699 (1976).

[11]C. Djerassi, R. M. K. Carlson, S. Popov and T. H. Varkony, *Marine Natural Products Chemistry* (Edited by D. J. Faulkner and W. H. Fenical), p. 111. Plenum, New York (1977).

[12]T. H. Varkony, R. E. Carhart and D. H. Smith, presented at the meeting of the American Society for Mass Spectrometry, Washington, District of Columbia (1977).

[13]E. Lederer, *Quart. Rev., Chem. Soc.* **23**, 453 (1969).

[14]F. F. Knapp, J. B. Greig, L. J. Goad and T. W. Goodwin, *J. Chem. Soc.*, Chem. Comm. 707 (1971).

[15]J. D. BuLock, *The Biosynthesis of Natural Products*, p. 94. McGraw Hill, New York (1965).

[16]G. A. Cordell, *Chem. Rev.* **76**, 425 (1976).

[17]G. Mellows, P. G. Mantle, T. C. Feline and D. J. Williams, *Phytochemistry* **12**, 2717 (1973).

[18]T. C. Feline, G. Mellow, R. B. Jones and L. Phillips, *J. Chem. Soc.* Chem. Comm., 63 (1974).

[19]M. Tanabe, K. T. Suzuki and W. C. Jankowski, *Tetrahedron Letters* 2271 (1974).