45. Computer Assisted Structural Interpretation of Mass Spectral Data[1])

by **Neil A.B. Gray, Armand Buchs[2]), Dennis H. Smith** and **Carl Djerassi**

Department of Chemistry, Stanford University, Stanford, California 94305

Department of Chemistry, University of Geneva, 16 Bd. d'Yvoy, CH-1211 Geneva 4

(20.XII.80)

Summary

An interactive system of programs has been developed to assist in structure elucidation based on mass spectral data. The program relies on algorithms for generating all the structural isomers that constitute alternative explanations of the observed data and it associates relative plausibility values with the different isomers. The structure assembly part of the program allows for the use of overlapping substructural components, such as substructures inferred from the appearance of particular ion patterns in the spectrum of an unknown compound. Mass spectrum interpretation procedures used with this structure assembly process could exploit any form of spectrum-substructure correlation scheme. In this work, the emphasis has been on the use of detailed and class specific spectrum-substructure correlations. Applications of the program are illustrated by means of an example analysis of the mass spectra of a variety of marine sterols.

Introduction. – We have recently described computer programs that can exploit structure-spectrum relationships to determine which of a set of plausible candidate structures provides the best rationalization of an observed mass spectrum [2] [3]. Though of general utility, these programs are limited by the prerequisite of a set of candidate structures. They can be most appropriately applied to problems where spectral and chemical evidence provide sufficient structural constraints for an isomer generating program [4] to produce a reasonably limited number of plausible candidates. Other applications are to cases where the unknown is related to a known structure and where plausible candidates can be generated by a program applying, for example, biosynthetic reactions [5].

There remains considerable scope for the development of an approach in which mass spectral data provide substructural constraints for a structure generating

[1]) Part XXXVI of the series 'Application of Artificial Intelligence for Chemical Inference'. For part XXXV see [1].

[2]) On sabbatical leave from the University of Geneva, Switzerland. Author to whom correspondence should be addressed.

0018-019X/81/2/0458-13$01.00/0

program such as CONGEN [4] or the more recently developed GENOA program [6]. A number of earlier programs, such as the Heuristic DENDRAL program [7] and the PLANNER program [8], did incorporate simple 'Inference' or 'Planning' functions that could infer structural constraints from mass spectra. However, the approaches employed in those programs were limited in their versatility and in their applicability, especially by the close coupling of the spectrum-interpretation process and structure-generation procedure. The lack of a clear boundary between data-dependent interpretive procedures and structure-generating functions makes it particularly difficult to extend such programs to handle new classes of compounds.

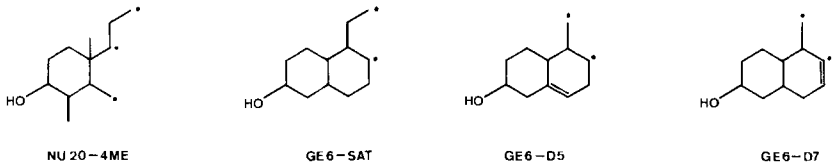
Recent development of structure generating algorithms has led to a new program, GENOA, possessing considerable generality and flexibility [6]. Two features of this second-generation structure builder are of particular interest in the context of processing substructures inferred from mass spectra. First, GENOA allows the use of potentially overlapping substructures whereas CONGEN could only accept separate and distinct substructures as constraints. Indeed, substructures that might be inferred from the analysis of a mass spectrum are particularly likely to overlap. An atom of a parent structure may be found in many of the ions resulting from different fragmentation processes. Any structure generator that is to use constraints inferred from fragment ions must allow for the possibility of overlap. The second

Table 1. *Part of the index to the marine sterol data-base*

Nominal mass: 161	Formula: $C_{12}H_{17}$	
The following substructures are associated with this composition:		
NU20-4ME	161 $C_{12}H_{17}$	Intensity: 10–40%
	Compositions and intensities of satellite ions:	
	147 $C_{11}H_{15}$	Intensity: 5–30%
	149 $C_{11}H_{17}$	Intensity: 10–50%
	163 $C_{12}H_{19}$	Intensity: 5–30%
GE6-SAT	161 $C_{12}H_{17}$	Intensity: 10–30%
	Compositions and intensities of satellite ions:	
	159 $C_{12}H_{15}$	Intensity: 2–10%
	175 $C_{13}H_{19}$	Intensity: 5–20%
GE6-D5	161 $C_{12}H_{17}$	Intensity: 5–30%
	Compositions and intensities of satellite ions:	
	143 $C_{11}H_{11}$	Intensity: 5–30%
	145 $C_{11}H_{13}$	Intensity: 10–50%
	159 $C_{12}H_{15}$	Intensity: 10–50%
GE6-D7	161 $C_{12}H_{17}$	Intensity: 5–25%
	Compositions and intensities of satellite ions:	
	143 $C_{11}H_{11}$	Intensity: 2–10%
	145 $C_{11}H_{13}$	Intensity: 10–30%
	159 $C_{12}H_{15}$	Intensity: 5–25%

major advantage of GENOA over CONGEN is that it allows the use of constraints requiring the presence of only one member from a set of alternative substructures. CONGEN, on the other hand, required the constraints to be completely specific as to form and number of instances of each substructure. Since it is frequently possible to interpret a particular pattern of fragment ions in terms of more than one substructure, the flexibility of GENOA is required. For example, as discussed in more detail later and summarized in *Table 1* and *Scheme 1*, ions at *m/z* 161

Scheme 1. *Substructures associated with the ion patterns of* Table 1



($C_{12}H_{17}$) and *m/z* 159 ($C_{12}H_{15}$) occur in the spectra of sterols with nuclei incorporating any one of the substructures NU20-4ME, GE7-SAT, GE6-D5 or GE6-D7 illustrated in *Scheme 1*[3]). The appearance of such ions in the spectrum of a sterol may thus be interpreted in terms of the presence of any one of these substructures. In such cases it is appropriate to specify all alternative substructural interpretations as a single, composite constraint in structure generation. As will be illustrated later, GENOA further allows different weighting scores to be associated with the different members of a set of alternative substructures used as a single constraint. These weighting scores allow the relative plausibilities of the different substructural interpretations to be distinguished. For example, the data summarized in *Table 1* suggest that if the $C_{12}H_{15}$ ion is more abundant than the $C_{12}H_{17}$ ion then substructures such as GE6-D5 and GE6-D7 are slightly more probable than NU20-4ME or GE7-SAT. Generated candidate structures acquire scores based on a combination of the scores associated with the substructures that they incorporate. These overall candidate scores then can be used to eliminate those structures which are ranked poorly because they combine many low plausibility features.

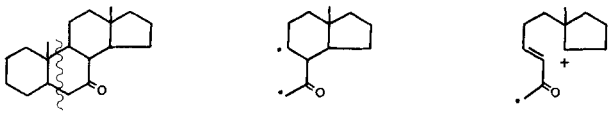
We present here an approach to the interpretation of mass spectral data that takes advantage of the flexibility of the GENOA program. This program could exploit any form of spectrum-substructure correlations, such as those which have been summarized by *McLafferty* [9]. We have chosen to use a data base containing substructures associated with mass spectral data. Such a data base is of course inherently limited to the analysis of structures similar to those used in its construction. *But, since the interpretation and structure generation procedures are completely distinct and general, application of the method to a new class of structures requires only the construction of a new data base.* The possibilities and the limitations of our approach are illustrated in relation to the characterization of steroids, for which the data necessary for the construction of the data base were available from earlier studies [3] [10].

[3]) The naming system for the substructures is based on the fragmentation rule schemes reported by *Lavanchy et al.* [3].

Method. – Essentially, the spectrum-interpretation process employed by the program involves the retrieval from a data base of reference substructures which serve as rationalizations for specific features appearing in the mass spectrum of an unknown compound. These substructures are then used as constraints in the structure-generation procedure. In addition to any substructural constraints, the generation procedure can also utilize other data, such as for example data from the interpretation of ^{1}H- or ^{13}C-NMR. spectra.

The current interpretation functions rely on the existence of a class-specific-data base which contains definitions of substructures together with some form of spectral signature, *e.g.* a characteristic pattern of ions or neutral fragment losses associated with each substructure. The substructures represent those fragments of a molecule which are associated with the occurrence of a particular ion; they do not generally represent actual or hypothesized ion structures. Thus, for example, the structure associated with the occurrence of an abundant ion at *m/z* 178 in the spectra of androstan-7-one (**1**) and also of its A-ring-functionalized derivatives, would be represented in a data base as **2**[4]), despite the fact that the actual ion is thought to be best represented as (**3**) [11].

Scheme 2



1 **2** **3**

Substructures in the class-specific-data base are those associated with major cleavage processes. Typically, such major cleavages are accompanied by several characteristic hydrogen-transfer reactions, neutral losses and secondary cleavages leading for example to the loss of acyclic appendages. Thus, substructures in the data base are in general characterized by a pattern of ions, with related masses and relative abundances, that all result from the same basic cleavage process. It is these ion patterns that are used as the spectral signatures for the substructures. The data bases are indexed by the nominal mass and by the elemental composition of all of the ions that they include.

The substructures included in the data base can be derived either by conventional analysis of mass spectral data, or by application of the INTSUM program to the mass spectra of known compounds [10]. The process of employing such a data base to aid in the analysis of an unknown mass spectrum involves: 1) selection of a specific ion in the observed mass spectrum, 2) use of this ion to index the data base and retrieve associated substructures, 3) estimation of the relative plausibilities of alternative substructures, 4) use of the retrieved substructures and plausibility values to construct and score candidate molecular structures. This entire process is repeated for each ion selected from the observed mass spectrum.

---

[4]) Species such as **2** and the substructures in *Scheme 1* are charged; the charge is not indicated in the diagram.

The investigator, using the mass-spectral-interpretation program, selects individual ions from the observed mass spectrum that appear structurally significant on the basis of relative abundance or distinctive elemental composition, and enters them into the program. The program retrieves all substructures indexed by the selected ion and displays summaries of those substructures' spectral signatures. The data in *Table 1* would be displayed if a $C_{12}H_{17}$ ion were used to index the marine sterol data base; the other ions forming the spectral signatures for each of the four substructures indexed by this composition are listed as 'satellite' ions; further data include the relative abundances of these ions.

The investigator must estimate the degree of consistency between the mass spectrum of the unknown and the ions with their relative-intensity patterns which form the spectral signature of each of the retrieved substructures. Although standardized schemes can be employed, this estimate of compatibility is basically subjective. The estimate must be given to the program as an integer score in the range -100 for complete mismatch to $+100$ for perfect agreement. Typically, many of a set of spectral signatures will match the observed data to some degree, though changes in relative ion abundances may make some substructures considerably more plausible than others. Thus, intrinsically, the result of analyzing a few features in the observed mass spectrum is a set of alternative substructural interpretations.

It is of course possible for the unknown compound to include novel substructural features that do not correspond to any of the substructures in an existing data base. A particular fragment ion in the observed spectrum may be produced by some fragmentation process induced by a novel substructural feature in the unknown compound. All rationalizations given in the existing data base for that ion could then be invalid in the context of the unknown structure. Consequently, it is generally necessary to include a 'null' alternative in any set of retrieved substructures.

The set of alternative substructures and associated plausibility scores are then used in the structure generation part of the program. The structure generating routines examine all existing partially assembled structures or 'CASES' [6]. If a particular substructure is found to exist within a given case, then the program merely updates the overall score associated with the case to reflect the plausibility of that substructure. If a substructure is not present in a given case, but could be assembled from the constituent parts of the case in one or more ways, then the constructive program creates new cases incorporating assembled versions of the substructure [6].

Each ion composition selected from the observed spectrum may either be used as such to index the data base or, in association with a second ion composition for an assumed parent ion, it may be used to identify possible neutral losses. Substructures corresponding to plausible neutral losses can thus be included in these data bases and used to aid the analysis of unknowns.

Application to marine sterols. - In the course of extended studies on marine sterols carried out in our laboratories [12], a substantial set of high-resolution mass spectra of known sterols have been recorded and analyzed. This set of sterol

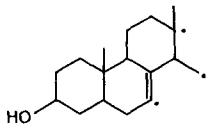
spectra was used to create a data base. Typical data contained in this data base are illustrated in *Table 1* which shows a portion of the data base relating to ions of composition $C_{12}H_{17}$ or of nominal mass 161.

The four substructures, shown in *Scheme 1*, are all associated with an ion of this composition. For each of these substructures, the ion $C_{12}H_{17}$ is just one of a group of ions resulting from a particular type of cleavage of a sterol accompanied by different neutral losses or hydrogen-transfer reactions. Substructure NU20-4ME in *Scheme 1*, for example, corresponds to that part of a 4-methyl-sterol nucleus left after cleavage through rings B and C. In the spectra of 4-methyl-sterols the ion at *m/z* 161, $C_{12}H_{17}$, results from this cleavage accompanied by loss of H_2O and H_2; the satellite ions at *m/z* 163, 149 and 147 involve the same cleavage but with loss of H_2O, $(H_2O + CH_3)$ and $(H_2O + CH_3 + H_2)$, respectively.

When processing the marine-sterol spectra, a predefined scheme was used for estimating the relative plausibilities of alternative retrieved substructures. In this scheme the score obtained by a substructure was the sum of scores for each ion specified in its pattern. A score of +4 was given to an ion observed within the specified intensity limits; ions specified in the pattern associated with a substructure but not present with appropriate intensities in the observed spectrum contributed −8 to that substructure's score. Increased emphasis was given to the more diagnostic high mass ions by doubling the scores for substructures incorporating more than twenty C atoms.

The Δ^5- and Δ^7-nuclei are of course the most difficult to discriminate because they exhibit similar sets of ions with similar intensities. One of the few discriminating features appears to be the pattern of ions, associated with the substructure shown in *Scheme 3*, which we have observed only in the spectra of Δ^7-sterols used in the training set.

Scheme 3. *Substructure and ion patterns characteristic of the Δ^7-sterols*



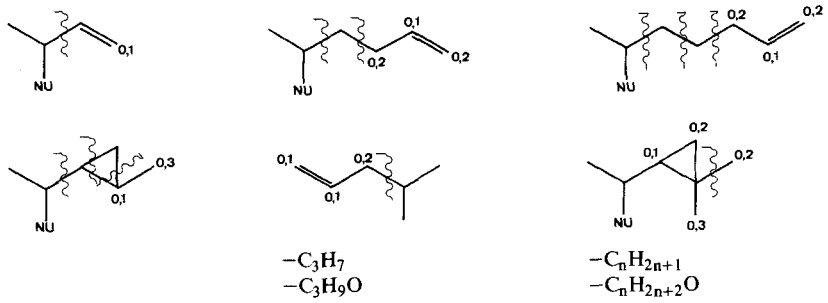
GEII-D7

$C_{17}H_{26}O$	7–35%
$C_{17}H_{24}$	2–10% ($-H_2O$)
$C_{16}H_{23}O$	8–20% ($-CH_3$)
$C_{16}H_{21}$	8–20% ($-H_2O-CH_3$)

Scheme 4 shows the main-fragmentation modes of the side chains which were utilized for the construction of the data base[5]). The ions which correspond to these fragmentations are formed, in most cases, through additional losses of H_2O, CH_3, H_2, H or combinations of those losses. *Scheme 4* also shows two side chain substructures which would be retrieved if the investigator specifies that the ion being processed should be used to identify the loss of a neutral fragment from a predefined precursor ion. If the fragment which was lost corresponds to C_3H_7

[5]) The digits which appear next to some of the C atoms in *Scheme 4* represent the H ranges for these atoms; 0, 2 for example means that at the time of structure generation the C atom will be allowed to bond to 0, 1 or 2 H atoms.

Scheme 4. *Main side chain fragmentation modes*

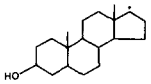
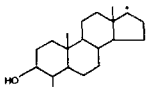
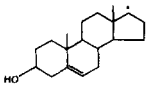
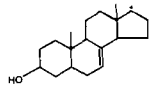
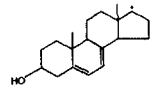


for example, the substructure in *Scheme 4* possessing the isopropyl group would be retrieved from the data base.

The data shown in *Table 2* for the various Δ^{22}-unsaturated sterols illustrate that it is not always possible to consider separately the fragmentation modes of the side chains and those of the nuclei [13]. The extent of cleavage of the C(20)–C(22) vinylic bond depends on the degree of unsaturation of the sterol nucleus. This dependency probably relates to the initial charge localization; the side chain cleavage is most important with the two saturated skeletons and completely absent from the spectra of the sterols possessing a $\Delta^{5,7}$-nucleus.

This data base was employed in the interpretation of the mass spectra of seven sterols not included in the reference set used for its creation. The analyses were

Table 2. *Ion patterns which result from the cleavage of the C(20)–C(22) bond in the different nuclei*

Nuclei	Ions observed
	$C_{21}H_{34}O$ 10–100% (–H) $C_{20}H_{31}O$ 5– 50% ($-H-CH_3$)
	$C_{22}H_{36}O$ 40–100% (–H) $C_{21}H_{33}O$ 10– 40% ($-H-CH_3$) $C_{22}H_{37}O$ 1– 20%
	$C_{21}H_{32}O$ 20– 50% (–H) $C_{20}H_{27}$ 2– 20% ($-H-H_2O-CH_3$)
	$C_{21}H_{32}O$ 20– 50% (–H) $C_{21}H_{31}O$ 1– 20% ($-H_2$) $C_{20}H_{29}O$ 1– 10% ($-H-CH_3$)
	No such cleavage

carried out using an interactive mode of the program with the chemist selecting the ions to be processed and scoring the different proposed substructures. In addition to the constraints inferred from the mass spectra, the structure-generation procedure was further limited by constraints derived from biosynthetic considerations. The candidate structures were restricted to those incorporating one of the common marine sterol nuclei shown in *Table 2* and the forms allowed to side chains were limited to a subset of those generated by the REACT program employing biosynthetically plausible transforms of the desmosterol side chain (**4**). This subset included all side chains previously identified in known marine sterols along with some additional, particularly probable side chains generated by the REACT program.

Table 3 shows an example of the dialogue during the interpretation of the high resolution mass spectrum of stelliferasterol (**5**). First, the molecular formula is defined and then files containing appropriate side chains are specified. In this example, the relevant files contained twenty different side chains; the program used these to create the initial 'CASES' [6], *i.e.* partially assembled structures being considered. The next step involved the specification of the five standard alternative marine-sterol skeletons; as there was no prior evidence favoring any particular nucleus, these were all assigned equal starting scores. Generation using the nuclei and the side chains resulted in thirty-six cases.

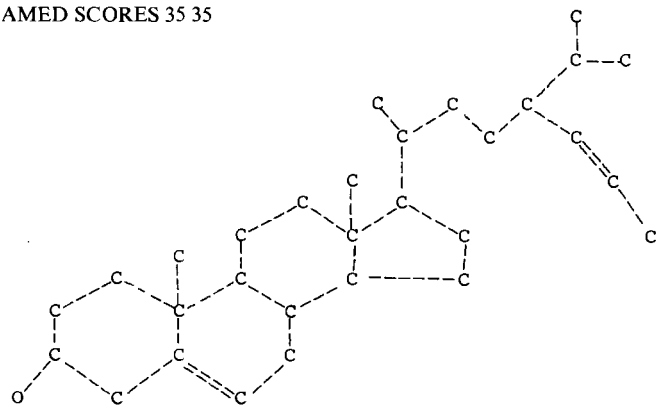
These initial cases included examples in which additional methylene groups occured between C(17) of the nucleus and C(20) of the side chain. Such cases were for example generated from a C_{30} molecular formula with a C_{10} side chain and a C_{19} nucleus. These cases were eliminated through the use of the constraint involving substructure 'OVLAP' (**6**) which defines the correct attachment of side

Table 3. *Portion of the dialogue during the interpretation of the high resolution mass spectrum of* **5**

```
$ RESTORE STEROL. INT[a])
  sterol. int restored[b])

$ DEFINE MOLECULAR FORMULA
  molecular formula: C 30 H 50 O

$ PRESUPPOSE
  file with substructures: C10. SIDECHAINS
  file with substructures: C11. SIDECHAINS
  file with substructures:
  .....
  20 cases were obtained

$ ALTERNATIVES
  alternative substructures and scores: SAT-SKEL 1 D5-SKEL 1 D7-SKEL 1 D57-SKEL 1 4ME-SKEL 1
  . +.. +.. +.. +. +. +.. +.. +.. +.. +.. +.. +. +.. +. +.. +.. +.. +.. +.. +.. +.
  36 cases were obtained

$ CONSTRAINTS
  substructure name: OVLAP
  range of occurrence: EXACTLY 1
  + + + + + + + + + + + + + +. + + + + + + + + + +. +. +. +. +. +. +. +. +. +. +. +. +. +. +. +. +.
  16 cases were obtained
```

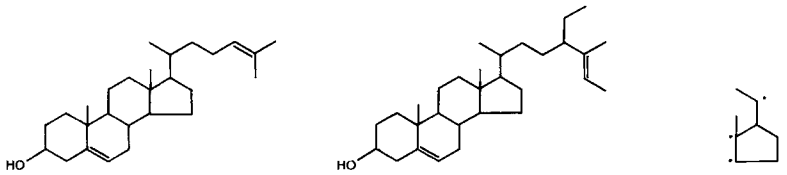
Table 3. (continued)

```
$ MSI
      ---functions for interpretation of mass spectra---
  which is the file containing your data base: AB. STEROL
  low resolution data: NO
  type: ION
  composition: C 23 H 36 O
  observed intensity: 29
  possible substructures for that composition together with their ion patterns:
  sd8a-d7          2-25%
  sd8a-d5          2-25%
  sd346-d7         5-30%
  sd346-d5         5-30%
  Assign relative confidence ratings to the different substructures (score = 0→ ignore substructure)
  sd8a-d7          score: -10
  sd8a-d5          score: -10
  sd346-d7         score:  10
  sd346-d5         score:  10
  none of these    score:   1
  +. +. +. +. +. +. +. +. +. +. +. +. +. +. +. +. +.
  16 cases were obtained
$ MSI
  continue? YES
  type: ION
--- etc --- etc --- etc ---
$ SHOW SCORES
 case index number and scores:
   1     1     2    -18     3    -18     4     1
   5    -7     6     -7     7    -29     8    -19
   9    33    10     33    11     35    12     33
  13    35    14     33    15     33    16     35
$ FORGET 7
  forgetting 7
  shall I really forget? YES
$ DRAW ATNAMED SCORES 35 35
  ≠ 11
```



[a]) Capital letters = chemist.
[b]) Lower-case letters = computer.

Scheme 5



4 5 6

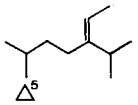
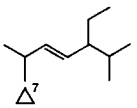
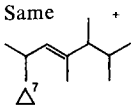
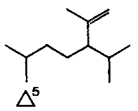
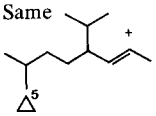
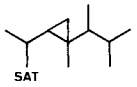
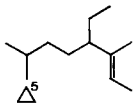
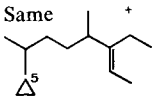
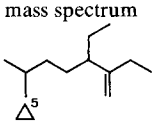
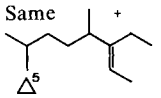
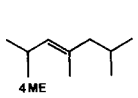
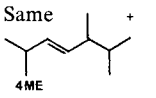
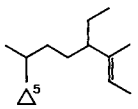
chain and nucleus. As shown in *Table 3*, this constraint eliminated 20 of the initial 36 cases.

Once a set of candidate structures compatible with the applied constraints has been created, the spectrum-interpretation process can be initiated by giving the 'MSI' command to the program. The mass spectral-analysis functions require the appropriate data base to be identified, in this case the file 'AB.STEROL', and a processing mode selected. The program works either in high-resolution mode, in which all ions and losses are processed as elemental compositions, or in low resolution mode, in which only nominal masses are considered. Each ion subsequently entered by the investigator may be processed as such or used in order to identify a neutral loss. The program takes the elemental composition, or the nominal mass, of each ion or loss and uses this to index the data base, presenting the investigator with details of all associated substructures and their characteristic spectral features. In the example shown in *Table 3*, the selected ion $C_{23}H_{36}O$ was found to be associated with four substructures characterizing different side-chain fragmentations. Here, the fragmentation process involving these substructures are simple, without additional neutral losses or hydrogen-transfer reactions so, consequently, the spectral patterns for the substructures consist of the single ion $C_{23}H_{36}O$ with appropriate intensity ranges. The observed ion occured with an intensity appropriate to substructures sd346-d7 and sd346-d5 so these received favorable scores, but was too abundant to be appropriate for fragmentation processes relating to sd8a-d7 or sd8a-d5 so these substructures were given unfavorable scores. It is of course possible that the actual fragmentation process leading to this ion in the unknown is induced by structural features not yet represented in the data base. Consequently, a null substructure, 'none of these', is also an allowed interpretation.

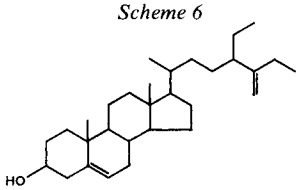
The program then checked the alternative substructures, sd8a-d7 etc., against existing cases. Here, no new cases could arise and the processing merely updated the scores associated with existing cases. The chemist can inspect intermediate results and summaries of the status of the problem solving process at any stage. It is also possible to eliminate low scoring cases from further consideration, as illustrated in *Table 3* where case 7, with a very unfavorable overall score of -29 was discarded. Structures associated with scores in any particular range can be displayed if desired.

For six of the seven sterols analyzed, the interpretation was based on high-resolution data acquired in our laboratories, while the seventh, strongylosterol (7),

Table 4. *Results for seven marine sterol mass spectra*

Correct structure	Best ranked structures	Correct structure	Best ranked structures
	Same		Same 
	Same 		Same
	Same 	Low resolution mass spectrum 	Same 
	Same 		

Scheme 6



7

used a partial low resolution mass spectrum (*m/z* 213 and above) published in the literature [14]. Results for these analyses are summarized in *Table 4*.

In each case, the correct structure was ranked either first or tied for first with closely related compounds; the nucleus was always unambiguously identified. However, the structure identifications obtained are not complete. Although identification of the nucleus is specific, side chains are really characterized as being in one of the mass spectrometrically non equivalent classes defined by the fragmentation rules of *Scheme 4*. In some cases there may be only one previously encountered representative of such a class of side chains, more typically several structures are possible. Because they do not exhibit any distinctive differences in their fragmentations, all saturated side chains with a given number of C atoms form a mass spectrometrically equivalent class. Though in this sense incomplete, the structure identifications that can be achieved are quite useful. This potential utility is well illustrated by the correct inference that the novel compound strongylosterol (**7**) was a 3-hydroxy-Δ^5 sterol, branched at C(24) and with a double bond at C(25). This inference was made on the basis of just a few ion patterns in the upper mass range of a low-resolution mass spectrum. It should be noted that for most interpretations only about ten ions in the upper part of the mass range need to be considered in order to obtain the correct result.

The analyses of the marine sterols were carried out using the interactive mode of the program with the chemist selecting the ions to be processed. Once the quality of the data base has been proven and, as in this application, some standard scoring scheme developed, it is appropriate for the program to operate in a more automated mode. One current option allows the program to list the fragment ions expected for particular substructures, or to identify the ions that would result from the losses associated with the presence of a given substructure, still leaving it to the chemist to check these patterns in the observed spectrum. Fully automated operation with the program directly checking the spectrum and scoring the structures is also possible.

Conclusions. – The combination of constraints based on grounds of biological plausibility and of substructures inferred from detailed mass-spectral patterns can be quite effective in determining the structures of moderately complex molecules such as the marine sterols. If the ion structures in the data base were characterized by some definitive property, and if it were practical to test the nature of the ions of the unknown with respect to this property, then any statements about substructural constraints inferred could be made with greater certainty. An appropriate ion property might be the collision-activation mass spectrum or the MIKES spectrum [15]. The equivalence of the MIKES spectra of ions is sometimes employed in structure elucidation as, for example, in the paper originally reporting the structure of strongylosterol [14] and studies of related sterols [16]. There, the equivalence of the MIKES spectra of the ions at *m/z* 271 and *m/z* 314 in the spectra of strongylosterol and fucosterol was used to establish that the former compound incorporated a 3-β-hydroxy-Δ^5-sterol skeleton identical to cholesterol up to atom C(22). The use of MIKES spectra could equally serve to characterize smaller fragment ions incorporating atoms from the sterol side chain. The potential value of such indirect determinations of structures was realized from the earliest work on collision-activated mass spectra [17]. Our current programs provide a means for exploiting such sources of structural data and assembling the inferred substructures into complete structures. Given a sufficiently comprehensive data base of substructures and associated MIKES spectra, it would be possible to analyze both more complex molecules and also examples with fewer biological or chemical constraints.

In our studies, we have emphasized the use of class-specific data bases and comprehensive spectral data for the unknown, such as ion compositions from mass spectra. The same structure-assembly mechanisms could also be used in association with other more general purpose systems for abstracting substructural data from mass spectra. One such system, *McLafferty's* STIR program [18] [19] has already undergone extensive development and is in reasonably widespread use. For typical unknown compounds, STIRS is capable of suggesting two or more correct substructural components along with maybe one or two inappropriate suggestions. Since substructural constraints inferred by STIRS are based in part on statistical arguments, our program's capabilities for allowing alternative interpretations of data, and for assigning plausibility weightings, would obviously be helpful. The practicality of using substructures inferred by STIRS is however still not proven.

The combination of STIRS with structure generators of the CONGEN/GENOA type is currently under investigation [20].

We wish to thank the *National Institutes of Health* (grants RR-00612 and AM-4247) for their generous financial support. Computer resources were provided by the SUMEX facility at *Stanford University* under *National Institutes of Health* (grant RR-0785).

Experimental Part

These programs are implemented in the ALGOL-like BCPL [21] programming language on a Digital Equipment Corporation KI-10 computer at the SUMEX-AIM facility in Stanford. The programs are available to an outside community of users *via* an international computer network.

REFERENCES

[1] *N.A.B. Gray, J.G. Nourse, C.W. Crandell, D.H. Smith & C. Djerassi,* Org. Magn. Reson., submitted.
[2] *N.A.B. Gray, R.E. Carhart, A. Lavanchy, D.H. Smith, T. Varkony, B.G. Buchanan, W.C. White & L. Creary,* Anal. Chem. *52*, 1095 (1980).
[3] *A. Lavanchy, T. Varkony, D.H. Smith, N.A. B. Gray, W.C. White, R.E. Carhart, B.G. Buchanan & C. Djerassi,* Org. Mass Spectrom. *15*, 355 (1980).
[4] *R.E. Carhart, D.H. Smith, H. Brown & C. Djerassi,* J. Am. Chem. Soc. *97*, 5755 (1975).
[5] *T.H. Varkony, D.H. Smith & C. Djerassi,* Tetrahedron *34*, 841 (1978).
[6] *R.E. Carhart, D.H. Smith, N.A.B. Gray, J.G. Nourse & C. Djerassi,* in preparation.
[7] *A. Buchs, A.B. Delfino, A.M. Duffield, C. Djerassi, B.G. Buchanan, E.A. Feigenbaum & J. Lederberg,* Helv. *53*, 1394 (1970).
[8] *D.H. Smith, B.G. Buchanan, R.S. Engelmore, A.M. Duffield, A. Yeo, E.A. Feigenbaum, L. Lederberg & C. Djerassi,* J. Am. Chem. Soc. *94*, 5962 (1972).
[9] *F.W. McLafferty,* 'Mass Spectral Correlations', Advances in Chemistry Series No. 40, American Chemical Society, Washington 1963.
[10] *D.H. Smith, B.G. Buchanan, W.C. White, E.A. Feigenbaum, J. Lederberg & C. Djerassi,* Tetrahedron *29*, 3117 (1973).
[11] *H. Budzikiewicz, C. Djerassi & D.H. Williams,* 'Structure Elucidation of Natural Products by Mass Spectrometry', Vol. II, Holden-Day, San Francisco 1964.
[12] *C. Djerassi, N. Theobald, W.C.M.C. Kokke, C.S. Pak & R.M.K. Carlson,* Pure Appl. Chem. *51*, 1815 (1979).
[13] *S.G. Wyllie & C. Djerassi,* J. Org. Chem. *33*, 305 (1968).
[14] *M. Bortolotto, J.C. Braekman, D. Daloze & B. Tursch,* Bull. Soc. Chim. Belge *87*, 539 (1978).
[15] *F.W. McLafferty,* in 'High Performance Mass Spectrometry: Chemical Applications', p. 47, M.L. Gross (Ed.), American Chemical Society, Washington 1978.
[16] *A. Maquestiau, Y. van Haverbeke, R. Flammang, H. Mispreuve, M. Kaisin, J.C. Braekman, D. Daloze & B. Tursch,* Steroids *31*, 31 (1978).
[17] *F.W. McLafferty, R. Kornfeld, W.F. Haddon, K. Levsen, I. Sakai, P.F. Bente, S.C. Tsai & H.D.R. Schuddemage,* J. Am. Chem. Soc. *95*, 3886 (1973).
[18] *K.S. Kwok, R. Venkataraghavan & F.W. McLafferty,* J. Am. Chem. Soc. *95*, 4185 (1973).
[19] *H.E. Dayringer, G.M. Peysna, R. Venkataraghavan & F.W. McLafferty,* Org. Mass Spectrom. *11*, 529 (1976).
[20] *In Ki Mun,* personal communication.
[21] *M. Richards & C. Whitby-Stevens,* 'BCPL – The language and its compiler', Cambridge University Press, Cambridge 1979.