Tetrahedron Vol. 29, pp. 3117 to 3134. Pergamon Press 1973. Printed in Great Britain

APPLICATIONS OF ARTIFICIAL INTELLIGENCE FOR CHEMICAL INFERENCE—X

INTSUM. A DATA INTERPRETATION AND SUMMARY PROGRAM APPLIED TO THE COLLECTED MASS SPECTRA OF ESTROGENIC STEROIDS†‡

D. H. SMITH, B. G. BUCHANAN, W. C. WHITE, E. A. FEIGENBAUM, J. LEDERBERG and CARL DJERASSI*

Contribution from the Departments of Chemistry, Computer Science and Genetics, Stanford University, Stanford, California 94305

(*Received in USA* 17 *January* 1973; *Received in UK for publication* 29 *May* 1973)

Abstract—A method for systematic interpretation and summary of evidence found for all possible mass spectral fragmentations of a molecule or set of related molecules is described. The method is embodied in a computer program (INTSUM) which interprets, in terms of fragmentation processes, mass spectral data collected on known compounds. Utilizing high resolution mass spectra from 47 estrogenic steroids, the method is verified and new findings are discussed. Finally, the method is used to explore the fragmentations of equilenins and several acetate and benzoate ester derivatives.

Interpretation of the mass spectra of known compounds to elucidate mechanisms of fragmentation has been performed manually for many years. The results of such interpretations have resulted in a significant body of empirical rules relating freatures of molecular structures to the ways in which the structures fragment subsequent to ionization. When more than a few related structures are involved, the task of manual examination of the spectra to determine related modes of fragmentation can become complex and tedious. The advent of high resolution mass spectra to limit compositional ambiguities inherent in low resolution mass spectral data has, to some extent, eased the problem of interpretation. Several techniques have been suggested which are designed to present high resolution spectra in a format that aids interpretation.[1-3] It can be argued, however, that detailed mechanistic interpretation of data in a single high resolution spectrum, and, particularly, comparison of several spectra of related compounds is still a difficult task utilizing these techniques.

This particular area of data interpretation seems well suited for study by techniques of artificial intelligence using the heuristic search paradigm.[4] The search in this case is over the space of possible fragmentation processes. The heuristics, or rules, used to guide the search are those empirical rules about fragmentation probabilities which are used routinely in manual data interpretation. These rules are sometimes very good, but sometimes inadequate. The use of such judgemental knowledge without guarantees of success characterizes artificial intelligence programs as a whole.

Interpretation of large amounts of data is a task well suited for a computer. It can explore the space of possible interpretations much more systematically and tirelessly than a chemist can. Computer techniques related to this approach have recently been utilized in mechanistic interpretation of individual low resolution mass spectrum/structure pairs.[5]

A set of 65 complete high resolution mass spectra of estrogenic steroids§ was available to test the performance of the computer program (termed INTSUM) written for data interpretation and summary. The first goal was to verify the performance of the program by comparing its results with manual interpretation, using a set of 47 compounds closely related to those studied previously.[6]

This verification step offered the advantage of ensuring that generalizations developed previously,[6] utilizing low resolution mass spectra, were correct. This step also offered the possibility of investigation, across a wide variety of compounds, of certain fragmentation processes which were of limited generality when incorporated into a program for automatic structure elucidation.[7] The remainder of the set of compounds was utilized to fulfill the second goal, namely, use of INTSUM to explore fragmentation processes for compounds (equilenins and several acetate and benzoate esters)

†This paper is dedicated to Professor Edgar Lederer of Institut de Chimie des Substances Naturelles, Gif-sur-Yvette, on the occasion of his sixty-fifth birthday.

‡For Part IX, see D. H. Smith, B. G. Buchanan, R. S. Engelmore, H. Adlercreutz and C. Djerassi, *J. Am. Chem. Soc.* submitted.

§Derivatives of the 1,3,5(10)-estratriene skeleton.

whose mass spectra have been subject to little or no prior investigation.

The program is described in the context of operation with high resolution mass spectral data. It is capable of analysis of low resolution mass spectra, with possible elemental compositions for each fragment ion limited only by the nominal mass and the heteroatom content of the molecule. As with manual interpretation there is a considerable increase in ambiguity of explanatory hypotheses when low resolution spectra are analyzed. The philosophy underlying development of this program, its importance to automatic theory formation, a long-range goal of this research, and an overview of requirements for generality and thoroughness have been presented.[8]

METHOD

Data interpretation and summary. The INTSUM program performs three basic tasks:

(1) Given the basic skeleton (superatom*) common to the set of related compounds, a nonredundant list of all possible fragmentations of this skeleton which result in smaller, unique fragments is produced. This list is called "ALLBREAKS".†

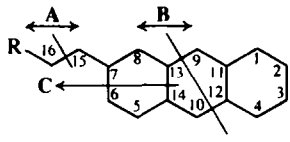
(2) Each structure/spectrum pair is then interpreted in turn as the program seeks evidence for each fragmentation in ALLBREAKS with transfer of hydrogens in or out of the charged fragment or without hydrogen transfers.

(3) Evidence for all structure/spectrum pairs is collected and correlated. Evidence for common fragmentation modes is grouped together and a summary output is provided.

The summary output shows which fragmentation modes are common to the entire set of molecules and which modes are more dependent on substituent placement, and to what extent they are dependent. The three tasks are described in more detail below. Additional details of the method are available.[8]

(1) *ALLBREAKS*

There are several points concerning generation of the ALLBREAKS list which deserve more detailed comment. The program is cognizant of the identities of each atom in the superatom. Thus fragmentation processes which result in fragments formally containing the same numbers of atoms but actually representing unique portions of the original skeleton are saved as separate processes. It is important to note that for an acyclic skeleton, every process of single bond cleavage separates the molecule into two smaller fragments. For wholly cyclic skeletons, however, every process in ALLBREAKS must consist of at least two single bond cleavages, as cleavage of one bond results only in a modified form of the molecular ion. In addition, if a cyclic skeleton is viewed as a graph structure, a fragmentation will be found in the ALLBREAKS list if and only if it begins and ends outside‡ the graph. These considerations are combined if the basic skeleton contains both cycles and chains. As an example, consider the alkyl-substituted perhydroanthracene skeleton (Scheme 1). The fragmentation of the alkyl chain, formally



Scheme 1

depicted as process **A** (15‖16), would be included in ALLBREAKS with charge placement on either of the two resultant fragments. Similarly, process **B** (Scheme 1, 10‖12, 9‖13) is an allowable process as it divides the structure into two fragments (begins and ends outside the graph structure). Process **C** (Scheme 1, 13‖14, 6‖7), however, is not an allowable process. Because it does not begin and end outside the graph, it does not split the skeleton into two fragments.

The ALLBREAKS list may be extended and/or restricted in a number of ways, completely under the control of the chemist analyzing the data, who can make as many or as few simplifying assumptions as he wishes. The following heuristics (rules) may be used in any combination:

(A) Cleavage of aromatic ring bonds and/or isolated double or triple bonds can be forbidden.

(B) Cleavage of two or more C–C bonds to the same carbon atom can be forbidden.

(C) A minimum number of skeletal atoms in the charged fragment may be specified.

(D) A specified atom or group of atoms may be transferred into or out of the charged fragment. This is subject only to the valence constraint that there must be a sufficient number of atoms to transfer. Transfer processes may be restricted to hydrogen atoms. The source and destination of hydrogen atoms are not specified.§

(E) Loss of and/or fragmentations within substituents on the superatom can be explored.

*As described in previous publications in this series, a superatom is defined as a structural subunit having at least one free valence. A free valence is a bond on an atom in the superatom to which another atom (e.g., hydrogen) or superatom may be connected.

†The program has sufficient flexibility to allow the chemist to input a selected list of possible fragmentations to be investigated rather than considering ALLBREAKS.

‡This concept is defined more precisely in Ref 8.

§It is recognized that many hydrogen transfer processes are relatively site-specific. However, because of formidable synthesis problems, a complete series of specifically deuterated analogs is seldom available, making it pointless to attempt specification without additional data.

(F) Multiple step processes (two-step, three-step,...) can be considered to a specified level (level two, level three,...).

(G) Analysis of a set of spectra with respect to a given list of processes is also possible. In this case ALLBREAKS is completely specified by the chemist.

Considering (F) in more detail, unless additional data are available (e.g., metastable defocussing), a multiple step process of level n is not allowable if the same fragment can be generated at level $n-1$. In other words, all processes are considered as concerted rather than stepwise, if possible.* For example, with reference to Scheme 1, process **B** with charge retention to the left of **B** followed by process **A** with charge retention to the right of **A** is a legitimate two-step process resulting in a fragment comprising C-10, 14, 13, 8, 7, 6, 5 and 15. Process **B** with charge retention to the left of **B**, followed by process **A** with charge retention to the left of **A** is not considered a legitimate two-step process as it can be fully explained by the single process **A** with charge retention to the left of **A**.

(2) *Fragmentation evidence*

Data input to the interpretation section of the program consists of structure/spectrum pairs. The structure is specified by the superatom used to create ALLBREAKS followed by modifiers representing substituent placements about the skeleton. For example, an hydroxyl group at carbon 1 (Scheme 1) would be specified as (SUBSTOH 1). The mass spectrum consists of a list of elemental compositions with their associated intensities expressed as a percentage of total ionization (%Σ).

Each structure/spectrum pair (a single pair may be considered if desired) is analyzed by searching for supporting evidence in the spectrum for each entry in ALLBREAKS. The appropriate elemental composition is sought through ALLBREAKS' knowledge of the identity of skeletal atoms in each proposed fragment and the specification of substituent placements in the input structure. For example, the ion C7H12 would be evidence for fragmentation **B** (Scheme 1) with charge retention to the right of **B** in any derivative of that particular superatom unsubstituted at C-1, 2, 3, 4, 9, 11 and 12. With an hydroxyl substituent at C-1, the ion C7H12O would be evidence for process **B** as described.

This part of the program performs another important task by grouping alternative explanations together. Because the elemental composition of an ion does not specify the portion of the skeleton from which it arose, this ion may be explicable by more than one of the processes in ALLBREAKS. The alternatives may be eliminated if appropriate isotopically labeled or substituent labeled molecules are available. But in general there may be several alternative explanations for an ion, all of which are saved.

(3) *Summary output*

The program summarizes all the evidence found for each fragmentation process. Each process for which any evidence is found is presented together with an ordered (by %Σ) list of those molecules displaying evidence for the process. This list also contains all alternative explanations for the ions assigned to this process.

Spectrum by spectrum output is also provided which includes:

(a) all explanations of each ion in each spectrum (within the restrictions for ALLBREAKS);

(b) a list of ions in each spectrum not explained by any process in the ALLBREAKS list;

(c) the percent of total ion current which remains unexplained.†

In addition, fragmentation processes differing only by hydrogen transfers can be grouped together as single processes if desired. These lists provide a check on how well the chosen set of ALLBREAKS has explained the spectrum.

RESULTS AND DISCUSSION

The number of proposed unique fragmentation processes to be considered for a complex molecule is formidable, even when the operation of ALLBREAKS is constrained at the discretion of the chemist, by employing heuristics described previously. This is amply illustrated for the subject of this study, the estrogen superatom (Fig 1). The number of processes (entries in the ALLBREAKS table) for various degrees of restriction are shown in Table 1. Note that the heuristics that the aromatic ring not be cleaved and only H atoms can be transferred are common to all entries in Table 1. These are heuristics which apply to most classes of molecules.‡ In particular, no structurally significant processes involving aromatic ring cleavages or group migration have been noted for estrogens.[6] Loss or fragmentation of substituents (heuristic E)

*This is an application of Occam's Razor. We feel that, confronted with a choice between simple and complex hypotheses, with no additional data, most scientists would choose the simpler one.

†A reviewer suggests that this value may be misleading if a number of low intensity ions at high mass have no explanation. We agree. The facility for examination of the data in this detail is not included in the present program. Manual examination of the spectrum by spectrum output quickly reveals this information.

‡There are, of course, exceptions to these heuristics. Predominantly aromatic compounds frequently undergo aromatic ring cleavage.[9] Also, trimethylsilyl ethers appear particularly prone to undergo group migrations during fragmentation.[10,11]

These heuristic limitations can be removed when dealing with a class of molecules wherein these mechanisms of generally lower probability may be operative.

Table 1. The number of unique fragmentation processes for the estrogen superatom. Letter designations of heuristic restrictions to ALLBREAKS common to all entries are defined in the KEY

Heuristic	Number of processes* No H transfer	Allow H transfer† −2, −1, 0, 1, 2
Two-step processes:		
(1) A, D, E, allow up to two-step processes	501	2500
(2) As (1), but forbid cleavage of two bonds to the same carbon atom	96	470
(3) As (2), but forbid fragments containing less than six skeletal carbon atoms	79	390
One-step processes:		
(4) A, D, E, allow only one-step processes	135	460
(5) As (4), but forbid cleavage of two bonds to the same carbon atom	49	210
(6) As (5), but forbid fragments containing less than six skeletal carbon atoms	37	175

KEY A—do not cleave aromatic ring bonds.
D—allow only hydrogen atoms to be transferred (no larger groups).
E—do not consider loss or fragmentation of substituents.
*Includes the identity process (molecular ion).
†A (−) sign indicates hydrogen transfer away from the charged fragment. The results in this column are not exact 5× multiples of the previous column because of valence constraints.

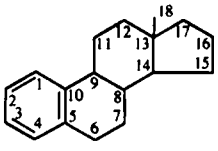


Fig 1. The estrogen superatom.

are not considered in this tabulation as the increase in numbers of processes are a function of the size and nature of substituents on each particular molecule.

One of the heuristics with greatest restrictive power is that which forbids cleavage of two C–C bonds to the same carbon atom (compare (1) to (2) and (4) to (5), Table 1). Processes of this type, resulting in charge retention on the portion of the skeleton retaining the C atom to which two bonds were broken, formally result in generation of a charged carbene species, processes which are generally regarded as unfavorable.* Cleavage of two C–C bonds to the same C atom involved in an eliminated neutral species may be a favorable process, however, particularly when the neutral species is carbon monoxide 13, 14 or another stable neutral molecule.

The optional restriction of specifying a minimum number of skeletal C atoms in the charged fragment (compare (2) to (3) and (5) to (6), Table 1) reduces the complications of having the program consider low mass ions which retain little or no structural information, at least in the case of estrogens.

The question of the level of processes to be considered for a class of molecules is an important one. Consideration of up to two-step processes for the estrogen superatom increased the size of the ALLBREAKS list by about a factor of four (compare (1) to (4), Table 1). This increase occurs despite the restriction (see Method section) that a two-step process is not considered if a single-step process can explain the fragment.

By forbidding cleavage of two C–C bonds to the same carbon atom, the size of the two-step process lists is reduced significantly resulting in about a

*Several instances of formal cleavage of a C–C bond and a C–H bond to the same C atom have been noted in studies of the fragmentation of other classes of steroids.[12] Because the program is insensitive to the source of hydrogen atoms, this type of process is formally cleavage of a single C–C bond with hydrogen transfer. Although carbon monoxide expulsion is sometimes noted in the fragmentation of derivatives of estrone (7), there is no definitive evidence for other processes involving cleavage of two C–C bonds to the same C atom for the simple estrogens (below).

factor of two more processes than for the single-step series (compare (2) to (5) and (3) to (6), Table 1). The subsequent discussion will be concerned with up to two step processes as these serve quite well to explain the majority of processes important to structure elucidation. This is, however, a characteristic of estrogens and may not be true in general.

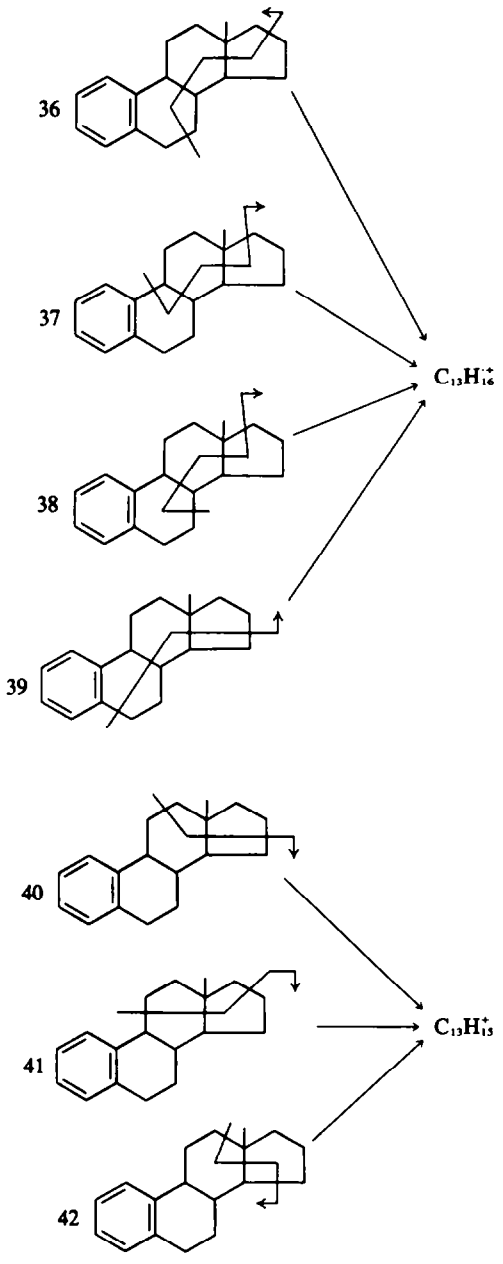
To provide an indication of the ambiguity of the high resolution data and the extent of alternative explanations for a datum, processes for generation of $C_{13}H_{16}^+$ and $C_{13}H_{15}^+$ fragment ions in the unsubstituted estrogen superatom (Fig 1) are presented in Scheme 2. These species are the only two possible which contain 13 C atoms for single-step processes involving no hydrogen atom transfers (Table 1, Case 4). Note that processes 37, 38 and 42 would not be considered if cleavage of two carbon-carbon bonds to the same carbon atom is forbidden. Also, the seven processes in Scheme 2 can yield the same ion, $C_{13}H_{15}^+$, if a single H atom transfer (−1, Table 1) is allowed in addition to zero H atoms transferred as specified in Scheme 2.

The specification of fragmentation processes produced by the program is highly symbolic (Scheme 2). The program does not attempt to delineate the course of fragmentation in any further detail as it is felt that, without additional data (e.g. deuterium labeling to specify H transfers), such detail would represent only speculation.

For simplicity the set of 65 compounds (see Table 2) was divided into two groups. The first group, comprising 47 compounds, contains a series of compounds substituted in various positions with what might be termed "simple" substituents, for example, hydroxyl, methoxyl, oxo, halogen, alkyl and so forth. These compounds are closely related to those studied previously[6] and comprise the set used for verification of the program. The second group (18 compounds) contains those compounds whose spectra have not been subjected to scrutiny, the equilenins,[6] and several acetate[15] and benzoate esters. This group comprised the set of compounds used to test the program on new data.

"Simple" estrogens

The spectra of the 47 compounds were analyzed employing the heuristics summarized in condition (3), Table 1, allowing from −2 to +2 hydrogen transfers. Because previous work[6] indicated no significant processes involving substituents and because verification of the program's operation was the primary goal in analysis of the 47 compounds, substituent processes were not considered. The right-hand column of Table 2 indicates the percentage of the total ion current explained by processes in the list of ALLBREAKS. Generally 70–90% of the ions in each spectrum could be explained with the remaining ion current distributed over a variety of low intensity, low mass ions which represent complex fragmentations including loss of substituents.



SCHEME 2

Exceptions are **13**, which displays intense loss of the C-6 and C-7 OH substituents as water, and to a lesser extent, 7-ketoestrone (**17**) which displays an intense series of ions of composition $M^{+\cdot}—C_nH_{2n+1}O_2$ ($n = 1$–3), processes which remain unexplained at this time.

From the standpoint of structure elucidation, the most important processes are those which occur in all or nearly all of the compounds. The data presented in Table 3 represent a portion of the output

of the program. This includes processes which comprise ten or more skeletal C atoms and for which evidence was found in the spectra of more than 40 out of the 47 compounds (85% or more).* The intensity data represent results after H transfers were combined. The most frequent H transfers are noted. A symbolic description of each process is included in the Table. The following general comments may be made about these data.

Ambiguity

There is significant ambiguity in the results, with the number of alternative explanations generally increasing as the size of the charged fragment decreases particularly with eight or fewer skeletal carbon atoms. The complete set of results, only part of which is summarized in Table 3, confirms this observation by specifying the ambiguity of every peak in every spectrum. There are, for example, four explanations for ions comprising eight skeletal C atoms, eight explanations for seven, and eight explanations for six skeletal C atoms. It has been a general assumption of mass spectroscopists that smaller fragments of molecular ions generally specify less information about molecular structure. These results place this assumption on a quantitative footing. In most cases where only a finite number of compounds are available, ambiguity will be the rule rather than the exception. For a relatively complex skeleton such as the estrogen skeleton a series of derivatives with representative substituents at every skeletal atom is seldom available. For many molecules the pattern of substitution will not allow differentiation of possible processes. Thus a single ion may have alternative explanations. Ambiguities can only be resolved by examination of the results obtained for compounds possessing substituents on skeletal atoms which are

*The program does not impose a lower limit on intensity of the peaks used as evidence. However, although peak intensities vary in principle through a continuum, the sensitivity of the mass spectrometer imposes a lower limit. Ions over a dynamic range of at least 100:1 but usually not more than 200:1 were recorded for all samples. This is a useful intensity range for most investigations of "significant" fragmentation processes, but it must be recognized that the operating conditions of the mass spectrometer create an artificial threshold determining whether a given molecule shows or does not show ions due to a given fragmentation process.

Table 2. The set of sixty-five estrogens

Compound	Name	% explanation*
1	desoxyestrone	87
2	2-hydroxyestradiol	79
3	2-methoxyestradiol	78
4	2-hydroxyestrone	78
5	2-methoxyestrone	77
6	estriol	75
7	estrone	87
8	estradiol	81
9	1-methylestradiol	77
10	16-hydroxyestrone	81
11	estrone 3-methyl ether	77
12	17α-methylestradiol	77
13	1-methyl-6α-7α-dihydroxyestrone	44
14	17-vinylestradiol 3-methyl ether	73
15	$\Delta^{7,8}$-estrone (equilin)	74
16	$\Delta^{6,7}$-estrone	74
17	7-ketostrone	63
18	17α-ethinylestradiol	77
19	17α-ethinylestradiol 3-methyl ether	71
20	11α-hydroxyestradiol	70
21	6-ketoestradiol	81
22	15α-hydroxyestradiol 3-methyl ether	72
23	$\Delta^{9,11}$-estradiol	72
24	$\Delta^{6,7}$-estradiol	76
25	16-ketoestradiol	82
26	15α-hydroxyestrone	74
27	11-keto-9β-estrone	67
28	estradiol 3-methyl ether	78
29	6-methylestrone	84
30	11α-hydroxyestradiol 3-methyl ether	71
31	11α-hydroxyestrone 3-methyl ether	70
32	1-methylestrone 3-methyl ether	79
33	$\Delta^{6,7}$-1-methylestrone	77

Table 2—*Continued*

Compound	Name	% explanation*
34	1-methylestrone	80
35	$\Delta^{9,11}$-1-methylestrone 3-methyl ether	62
36	$\Delta^{6,7}$-6-methylestrone	74
37	1,2-dimethylestrone	78
38	$\Delta^{6,7}$-1,2-dimethylestradiol	70
39	$\Delta^{6,7}$-1-methylestradiol	72
40	16,16-difluoroestrone 3-methyl ether	75
41	$\Delta^{6,7}$-1,2-dimethylestrone	72
42	13-propyl-18-nor-estrone	81
43	16β-fluoroestrone 3-methyl ether	78
44	16-bromoestrone 3-methyl ether	83
45	$\Delta^{15,16}$-estrone 3-methyl ether	77
46	$\Delta^{8,9}$-estrone 3-methyl ether	71
47	1-methyl-2-bromoestradiol	65
48	equilenin	90
49	equilenin 3-methyl ether	94
50	13-ethyl-18-nor-equilenin-14β	90
51	13-ethyl-18-nor-equilenin-14β 3-methyl ether	90
52	13-ethyl-18-nor-equilenin 3-methyl ether	92
53	1-methylestradiol 17-monoacetate	94
54	equilenin 3-acetate	92
55	$\Delta^{6,7}$-6-methylestradiol 3-benzoate	97
56	$\Delta^{9,11}$-estradiol 3-methyl ether 17-benzoate	94
57	estriol 3,16,17-triacetate	82
58	$\Delta^{6,7}$-estrone 3-acetate	93
59	estradiol 3-benzoate	99
60	2-methylestrone 3-benzoate	87
61	$\Delta^{6,7}$-estrone 3-benzoate	98
62	$\Delta^{6,7}$-1-methylestradiol 3,17-diacetate	92
63	$\Delta^{6,7}$-estradiol 3,17-diacetate	92
64	$\Delta^{16,17}$-estradiol 3-methyl ether 17-acetate	97
65	$\Delta^{6,7}$-1,2-dimethylestrone 3-acetate	91
66	16,16-d_2-equilenin 3-methyl ether	

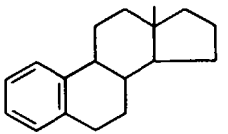
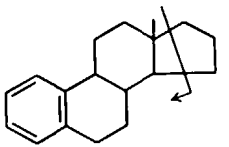
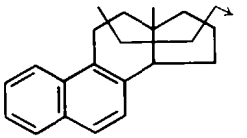
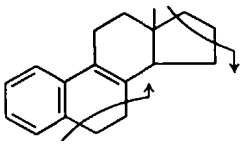
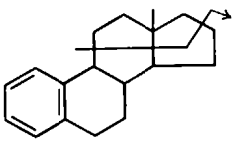
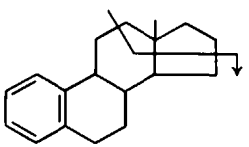
*This column presents the amount of data explained by the processes included in ALLBREAKS (100% = total). Processes involving loss of or fragmentation within substituents were considered for **48–65**, but not for **1–47**.

lost in one process but retained in an alternative process. The alternative explanations are by definition different processes and will differ from one another by at least two skeletal atoms for cyclic skeletons. This method of ambiguity resolution was manually applied to the program's output. Strong weight was attached to processes which were unambiguous explanations for at least some compounds which permit differentiation. This method has the potential drawback that the particular substituent label may itself direct the fragmentation of the molecule along a different pathway. This occurrence is revealed in the spectrum by spectrum outputs which were manually examined in an attempt to avoid this possibility. The results of this procedure are summarized in Table 4.

*An alternative explanation is loss of C-16 and C-17 with the C-17 oxo group. As this process is not significant in any other estrone-related compounds, it is not a plausible explanation for **27**.

Verification. The performance of the program is verified and results serve to extend previous knowledge. Mechanisms corresponding to processes 2L (cleavage of the C-5,6 C-7,8 bonds resulting in expulsion of C-6 and C-7 as ethylene), 10L, 18L, 17L, 7L and 6L have been proposed previously.[6] These mechanisms are verified as general to the skeleton, with the exception of 2L. As has been noted previously,[7] this process does not yield a significant ion in several instances, the most notable being C-17 hydroxyl compounds (generally displaying a less abundant ion than corresponding C-17 oxo compounds) and C-11 hydroxyl or oxo compounds, which yield no ions corresponding to process 2L. The C-11 substituted compounds (**20**, **27**, **30** and **31**) instead display evidence for loss of a two carbon fragment comprising C-11 and C-12 (process 9L). The resulting ion comprises 13·4%Σ for 11-keto-9β-estrone (**27**).* Process 9L may occur to a significant extent in other derivatives also, where it cannot be distinguished from process

Table 3. Processes[a] shown by ≥85% (≥40/47) of the simple estrogens, including ten or more skeletal carbon atoms

Process label[b]	Observed/ total	Retention of n skeletal carbon atoms n =	Symbolic description	Ambiguity[c]	Alternative explanations	% Σ range[d]	Most frequent hydrogen transfers
0	47/47	18 (molecular ion)		—	—	31·4–3·1	—
10L	46/47	15		No	—	26·7–0	0, −1
20L	43/47	14		Yes	2L/11L	3·7–0	−1, −2
2L/11L	45/47	14		Yes	20L	6·4–0	0, −1, −2
18L	45/47	13		Yes	19L 2L/10L	14·4–0	−1, −2
19L	44/47	13		Yes	18L 2L/10L	14·4–0	−1, −2

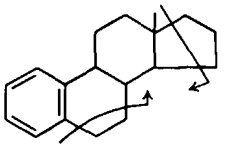
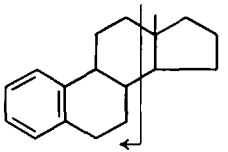
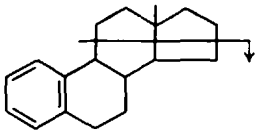
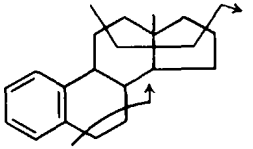
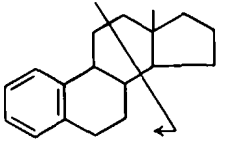
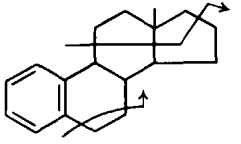
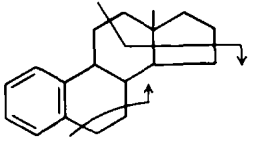
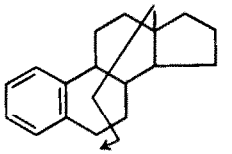
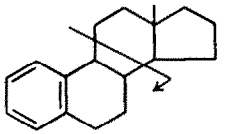
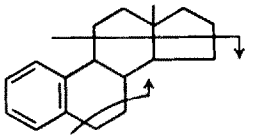
2L/10L	45/47	13		Yes	18L 19L	14·7–0	0, −1, −2
8L	45/47	12		Yes	17L 2L/20L	10·5–0	−1, −2
17L	46/47	12		Yes	8L 2L/20L	11·2–0	0, −1, −2
2L/20L	40/47	12		Yes	8L 17L	8·3–0	−1, −2
7L	46/47	11		Yes	2L/18L 2L/19L	19·8–0	0, −1, −2
2L/18L	40/47	11		Yes	7L 2L/19L	19·9–0	+1, 0, −1
2L/19L	46/47	11		Yes	7L 2L/18L	19·9–0	+1, 0, −1, −2

Table 3—*Continued*

Process label[b]	Observed/ total	Retention of n skeletal carbon atoms n =	Symbolic description	Ambiguity[c]	Alternative explanations	% Σ range[d]	Most frequent hydrogen transfers
16L	45/47	10		Yes	6L 2L/17L	15·0–0	0, −1, −2
6L	47/47	10		Yes	16L 2L/17L	24·6–0·1	+1, 0, −1
2L/17L	45/47	10		Yes	16L 6L	24·6–0	+2, +1, 0, −1

[a]The processes are ordered by decreasing size of the charged fragment.

[b]An "L" in the process label specifies charge retention on the portion of the molecule which contains the lowest numbered skeletal position of those positions involved in the bond cleavages. The numbering of the positions for the estrogen skeleton, or superatom, is specified in Fig 1. For example, process 10 cleaves the 13/17 and 14/15 bonds. Process 10L specifies charge retention on the portion of the molecule retaining C-13. An "H" specifies the reverse. Two-step processes are designated by two labels separated by a "/" mark.

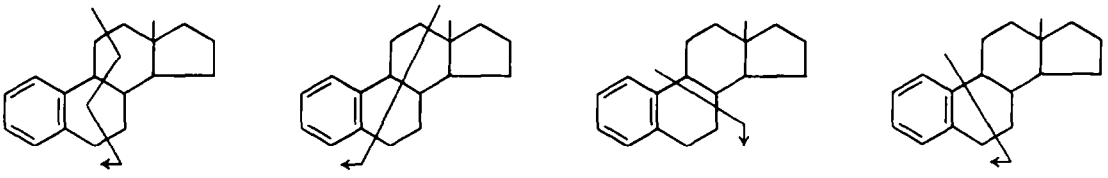
[c]A "no" indicates that there was no process which was a frequent alternative explanation. A "yes" indicates that one or more processes were frequent alternative explanations. Frequent was arbitrarily defined to be more than 50% of the time.

[d]In all cases a relatively smooth decline of % Σ values was observed between the high and low values given in the column. Thus the average value of % Σ for a process is approximately the mid-point of the indicated range.

Table 4. Processes, general to the skeleton, which remain after manual removal of ambiguities[a]

Process	Description	Compounds not displaying process
0	Molecular ion	—
10L	Retention of 15 skeletal carbon atoms (Rings A, B, C)	13
20L[b]	Retention of 14 skeletal carbon atoms	13, 26, 43, 44
2L/11L[b]		26, 41
18L	Retention of 13 skeletal carbon atoms	13, 44
2L/10L[b]		27, 31
17L	Retention of 12 skeletal carbon atoms	13
7L	Retention of 11 skeletal carbon atoms	13
6L	Retention of 10 skeletal carbon atoms (Rings A, B)	—
14L[b]	Retention of 9 skeletal carbon atoms	27, 31
15L		27, 31
5L[b]	Retention of 8 skeletal carbon atoms (including Ring A)	—
4L[b]	Retention of 7 skeletal carbon atoms (including Ring A)	13

[a]See Table 3 and Scheme 3 for process descriptions.
[b]Previously unreported processes suggested by INTSUM.



*Most frequent hydrogen transfers.

SCHEME 3

2L in the absence of isotopic or substituent labeling.

New fragmentation processes. The output of the program combined with analysis of ambiguities (Table 4) suggests that additional and alternative mechanisms are operative. Evidence supports both 20L and 2L/11L, and 2L/10L as an alternative to process 18L (Table 4). Stepwise processes involving 2L are alternative explanations for other processes in Table 4 as well, but not with sufficient frequency to define them as general. As an example, extensive metastable defocussing experiments performed utilizing estrone methyl ether[16] indicate that the fragment of *m/e* 256 (M^{+}–C-6,7 as C_2H_4) is a detectable but not extensive contributor to ions resulting primarily from 17L, 7L and 6L, in the first field-free region.

The new processes suggested in Table 4 (14L or 15L, 5L, 4L) serve to define ever decreasing portions of rings B and C. Because of their common occurrence they should facilitate structure elucidation. High resolution mass spectral data will increase the utility of these mechanisms. This region of estrogen spectra is generally quite complex, and there are many doublets unseparable by low resolution techniques.

Other studies[17,18] have led to the proposal of a mechanism formally corresponding to 15L for generation of fragment ions retaining nine skeletal carbon atoms comprised of, in part, ring A. Process 14L, however, is as plausible an explanation from the data at hand without invoking other considerations of fragmentation probability such as preferential cleavage of benzylic bonds. Process 4L is particularly helpful as it serves to define the substituents on the aromatic ring or at C-6. This process will allow more precise specification of molecular structure, as past efforts[7] could place substituents only somewhere on ring A or B based on mass spectral data alone.

Intensity variation. The processes summarized in Table 4 in general yield some variation in ion abundances as a function of particular substituents at particular positions. In many cases effects of substituents on these general skeletal processes are small. In certain molecules, however, the effect of

substituents on fragmentation is significant. Although the scope of this report does not permit an extensive summary of these variations, the most important are the following. The influence of the 6,7-dihydroxyl functionality was mentioned previously, as was the influence of a C-11 hydroxyl or keto functionality. The presence of additional skeletal double bonds usually results in diminished importance of processes involving cleavages of these bonds, as expected.

There are instances where the influence of unsaturations in Rings B, C, and D would probably not be predicted. Examples are the enhancing effect of a C-7,8 double bond on process 7L, and a C-9,11 double bond on process 10L. These observations may indicate double bond isomerization subsequent to ionization (Ref 19, p. 276). Methyl substitution in Ring A severely diminishes the importance of the characteristic Ring D cleavage (process 10L, Table 3), presumably a result of an increased population of molecular ions formed by ionization at the aromatic ring A.

Equilenins

Equilenins, estrogens with an aromatized Ring B, were treated as a separate sub-class. The five examples (**48–52**) are mentioned in Table 2 together with the percentage of data explained by ALLBREAKS (about 90% or more). The improvement over the degree of explanation for the simple estrogens is partially due to consideration of processes involving loss or fragmentation of substituents. The mass spectra of equilenins have been partially interpreted previously based on low resolution mass spectral data.[6,19]

The processes common to the set of five compounds are summarized in Table 5, along with processes related to specific substituents. There are processes common to the equilenins which are not summarized in the Table due to low ion abundances, which preclude their use for structure elucidation.

The processes outlined in Table 5 serve quite well to indicate origins of the significant ions in the spectra of the equilenins which have not been discussed previously. For example, with reference to the spectrum of equilenin methyl ether, **49**, Fig 2, *m/e* 223 arises from process 10L with one hydrogen loss and *m/e* 196 from 19L with H loss (supported by the lack of shift of *m/e* 196 in the spectrum of 16,16-d_2-equilenin methyl ether, **66**). There are several important processes involving substituents for the equilenins. For example the important hydrocarbon ions *m/e* 178 and 179, *m/e* 165, and *m/e* 152 and 153 arise from loss of the C-3 methoxyl substituent together with process 20L, 19L, and 8L or 17L, respectively (SUB3L processes, Table 5).

As expected, processes 18L, 17L and 7L (Table 5) all of which involve cleavage of a bond adjacent to aromatic Ring B, are much less important for equilenins than for the simple estrogens (Table 4). Process 6L (Ring C and D loss) is insignificant for the equilenins.

It is clear from a careful comparison of Figs 2 and 3 in combination with the processes summarized in Table 5 that important ions in the spectrum of **49** remain unexplained, *m/e* 237 and *m/e* 211. The mechanism postulated for genesis of *m/e* 237, 11L (Table 5), equivalent to that suggested by Budzikiewicz[19] appears plausible, but cannot be operative as the 16,16-d_2 analog of equilenin methyl ether (**66**) does not lose C-16[6] as illustrated in Figs 2 and 3, where *m/e* 237 is observed to shift to *m/e* 239. This observation is supported by examination of the intensities displayed by the resulting ion for **48–52**. Compounds **48** and **49**, which possess a methyl group at C-13, display intense ions at M^+–C_2H_3O of 8·9 and 7·7% Σ respectively. However, compounds **50–52**, which possess an ethyl group at C-13, yield ions of only about 1% Σ for process 11L. This implies that the angular Me or Et substituent is lost in the process, along with the elements of carbon monoxide. For compounds **50–52**, this loss would be equivalent to (ambiguous with) Ring D loss accompanied by loss of an H atom so that it cannot be stated unequivocally that the process is operative. A process of this type was not postulated by the program as it formally involves cleavage of two C–C bonds to the same C atom (C-13). Examination of the proposed origin of *m/e* 211 (Fig 2), process 20L, in light of the spectrum of 16,16-d_2-equilenin methyl ether (**66**, see Fig 3) reveals that it is operative only to a minor extent at least in the spectrum of **49**. The peak remains unshifted indicating loss of C-16 in the process (*m/e* 211, Figs 2, 3). There appears to be no obvious alternative explanation that holds true for all equilenins studied. There in fact appear to be two distinct processes operative, one generating M^+–C_4H_7O for **48** and **49**, which may be ring D loss accompanied by loss of the angular Me group, yielding *m/e* 211, Fig 2. This process is of lesser importance in the 13-ethyl-18-nor compounds **50–52** (the low resolution spectrum of **52** is presented in Ref 6). The other process is analogous to 20L with loss of two additional H atoms, is common to **48–52** and results in loss of C-16 (results, for **49**, in *m/e* 209, Fig 2). Because loss of the C-13 Me or Et group is implicated, the process would correspond to loss of C-13, 18, 16 and 17. Any mechanistic interpretation would be speculation without additional data. Further deuterium labelling experiments are in progress to clarify the origins of these particular ions.

Acetate and benzoate esters

The previous examples illustrated several fragmentation processes of structural significance. The acetate and benzoate esters, however, (Table 2, compounds **53–65**) exhibit fragmentation processes

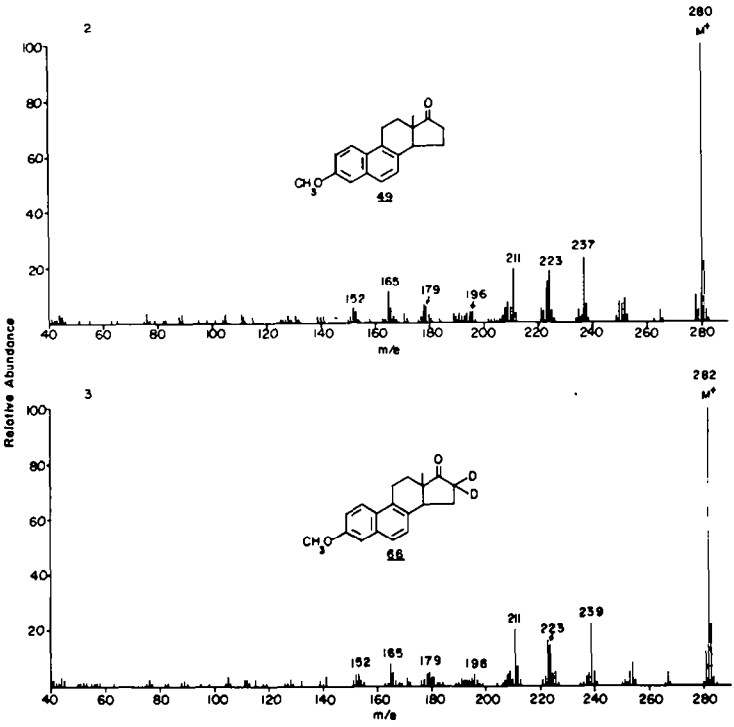


Fig 2. The low resolution mass spectrum (70 ev) of equilenin 3-methyl ether, 49.

Fig 3. The low resolution mass spectrum (70 ev) of 16, 16-d_2-equilenin 3-methyl ether, 66.

at least in part characteristic of the substituent rather than the estrogen skeleton. These derivatives are thus not as suitable for structural identification as the parent sterols. By requesting that the program consider fragmentation processes involving substituents, these processes can be investigated. A brief summary is provided to indicate some of the major cleavages and rearrangements involved. The data on acetate esters support and extend the findings of a previous study based on low resolution mass spectra.[15]

Acetyl derivatives

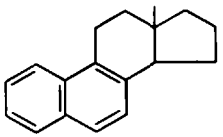
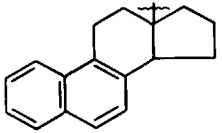
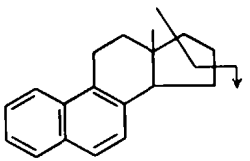
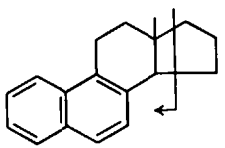
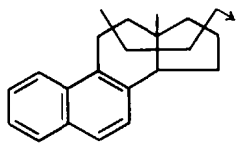
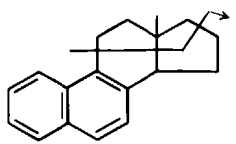
The spectra of all C-3 acetates, (**54, 57, 58, 62, 63,** and **65**), regardless of whether they are diacetates (**62, 63**) or triacetates (**57**), are dominated by loss of ketene from the C-3 acetyl group. This process, characteristic of acetates derived from aromatic hydroxyl functionalities,[20] is followed by processes summarized previously, such as Ring D loss, in the case of monoacetates **54, 58** and **65**. The diacetates **62** and **63** appear to follow loss of ketene with loss of the C-17 acetyl function as acetic acid. Other fragments are generated in further decompositions of these ions. Estriol triacetate (**57**) fragments in an analogous manner with loss of ketene accompanied by loss of two additional molecules of ketene, two molecules of acetic acid and all combinations thereof. Further decomposition of the ions so generated is complex and gives rise to many low abundance fragment ions.

1-Methylestradiol 17-monoacetate (**53**) displays the characteristic fragmentations noted for the simple estrogens, but with diminished intensity. Only minor elimination of acetic acid is noted. $\Delta^{16,17}$ estradiol 3-methyl ether 17-acetate (**64**), on the other hand, yields a mass spectrum exhibiting many of the features of the C-3-acetyl derivatives mentioned above. There is a strong ion (base peak) for elimination of ketene followed by processes reminiscent of estrone 3-methyl ether which, however, yield ions of greatly reduced intensity, particularly Ring D loss.

Benzoate esters

The C-3 benzoates (**55, 59, 60** and **61**) yield distinctive spectra consisting of the molecular ion, the benzoyl ion (C_6H_5CO+) and the phenyl ion (C_6H_5+) which account for 70%Σ for comound **55** and 90–95%Σ for **59, 60** and **61**. The C-17 benzoate, **56**, however, displays an intense ion from loss of benzoic acid, a process analogous to loss of acetic acid from the C-17 acetates. This is accompanied by an intense benzoyl ion, loss of benzoic acid plus Me radical and other ions of low abundance and obscure origins.

Table 5. Fragmentation processes common to equilenins[a]

Process label	Observed/ total	Retention of n skeletal carbon atoms n =	Symbolic description	Ambiguity	Alternative explanations	% Σ range	Most frequent hydrogen transfers
O	5/5	18 (molecular ion)		—	—	33·7–20·7	—
1L	5/5	17		No	—	10·7–0·8	0
11L	5/5	16		No	—	8·9–0·6	−1
10L	5/5	15		No	—	16·9–10·2	0, −1
20L	5/5	14		No	—	12·4–3·7	0 (for **48**, **49**) −1, −2 (see OC3•1L/10L)
18L	5/5	13		Yes	19L	4·7–1·8	−1

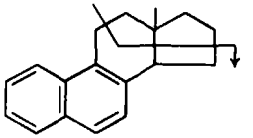
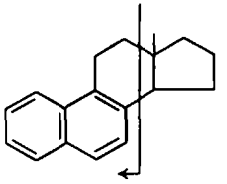
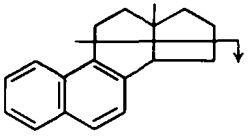
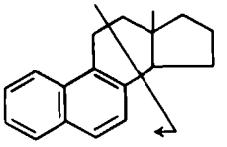
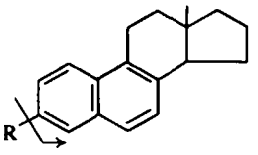
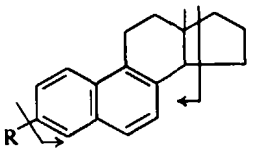
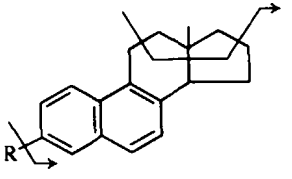
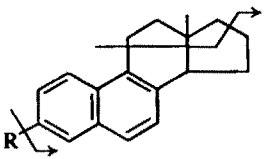
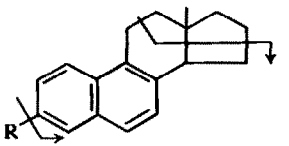
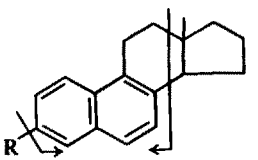
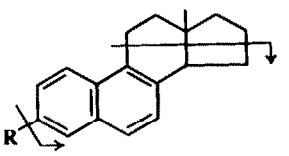
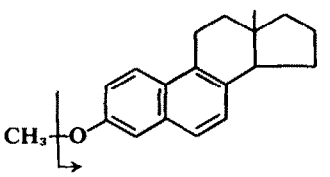
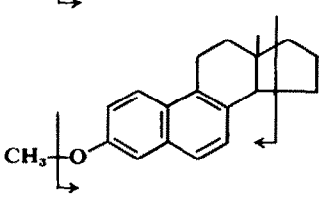
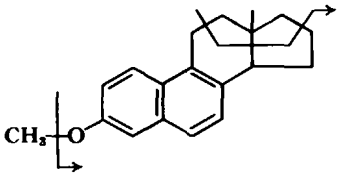
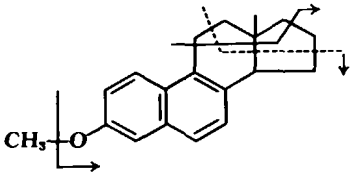
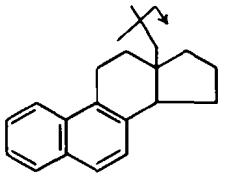
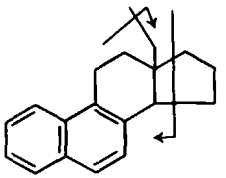
19L	5/5	13		Yes	18L	4·7–1·8	−1
8L	5/5	12		Yes	17L	1·7–0·6	0, −1, −2
17L	5/5	12		Yes	8L	1·4–0·8	+1, 0, −1
7L	5/5	11		No	—	1·2–0·7	+1
SUB3L	5/5	18 (loss of C-3 substituent)		No	—	2·4–0·2	0 (+1 for **49**, **51**, **52**)
SUB3L/10L	5/5	15		No	—	2·2–0·3	0, −1, −2
SUB3L/20L	5/5	14		No	—	4·7–2·3	0, −1, −2

Table 5—*Continued*

Process label	Observed/ total	Retention of *n* skeletal carbon atoms *n* =	Symbolic description	Ambiguity	Alternative explanations	% Σ range	Most frequent hydrogen transfers
SUB3L/18L	5/5	13		Yes	SUB3L/19L	6·4–4·4	−1
SUB3L/19L	5/5	13		Yes	SUB3L/18L	6·4–4·4	−1
SUB3L/8L	5/5	12		Yes	SUB3L/17L	4·0–2·7	0, −1
SUB3L/17L	5/5	12		Yes	SUB3L/8L	4·1–2·8	+1, 0
OC3*1L	3/3[b]	18		Yes	See 20L	1·3–0·8	0
OC3*1L/10L	3/3	15		Yes	See 20L	11·2–9·1	+2, 0 (see 20L)

OC3*1L/20L	3/3	14		Yes	18L 19L	3·3–2·9	0, −1, −2
OC3*1L/18L OC3*1L/19L	3/3	13		Yes	8L 17L	1·3–0·9	+2, 0
SUB18L	3/3[c]	18		No	—	1·2–0·8	0
SUB18L/10L	3/3	15		No	—	10·7–9·1	0, −1, −2

[a]See Table 4 for explanation of terms. Processes yielding ions of $<1\%\,\Sigma$, for all compounds, are not included.
[b]These processes involve loss of methyl from a methoxyl substituent, so that only compounds **49**, **51** and **52** are included.
[c]These processes involve loss of the substituent at C-18, so that only compounds **50**, **51** and **52** are included.

CONCLUSIONS

The computer program for data interpretation, INTSUM, has been shown to be a powerful aid to the interpretation of large quantities of high resolution mass spectral data.

The program's representation of knowledge of molecular structure and mass spectrometry is sufficiently flexible and general to suggest potential wide applicability. The output of INTSUM is a valuable aid to chemists in determining firm rules of fragmentation which can then be used in studies of related but unknown compounds. The program is presently in routine use in studies of the fragmentation of other classes of compounds, including other steroids and alkaloids.

EXPERIMENTAL

High resolution mass spectra were determined utilizing both AEI-MS9 and Varian-MAT 711 mass spectrometers. The former instrument was operated with an ionizing voltage and current of 70 ev and 500 ua, respectively. Scans were recorded at a scan rate of 34 sec/decade in mass. The latter instrument was operated with an ionizing voltage and current of 70 ev and 1·6 ma, respectively, and scan rates of either 38 or 22 sec/decade. Samples were introduced via the direct insertion probe in both instruments. Data were recorded on-line to a Digital Equipment Corp. PDP-11 interfaced directly to the ACME computer facility.

Data on defocused metastable ions* were obtained utilizing the AEI-MS9 and Varian-MAT 711 instruments with ion source conditions and sample introduction as outlined above.

Available samples of estrogens were subjected to mass spectral analysis without further purification. Two samples found to be mixtures were not included in this study.

The INTSUM program is written in the Stanford 360/LISP language and runs in batch mode on the Stanford IBM 360/67 machine or on Stanford Medical School's IBM 360/50 (the ACME facility). On the faster of the two machines, the summary of 47 estrogen spectra took roughly 15 min. Programming details are omitted here for the sake of brevity but can be obtained upon request from the authors.

Acknowledgements—Financial support from the National Institutes of Health (RR-612-02) and the Advanced Research Projects Agency (SD-183) is gratefully acknowledged.

*These experiments were carried out utilizing ion decompositions in the first field-free region of a double focussing mass spectrometer. See Refs 21–23.

REFERENCES

[1]K. Biemann, P. Bommer and D. M. Desiderio, *Tetrahedron Letters* 1725 (1964)
[2]A. L. Burlingame and D. H. Smith, *Tetrahedron* **24**, 5749 (1968)
[3]R. Venkataraghavan and F. W. McLafferty, *Analyt. Chem.* **39**, 278 (1967)
[4]E. A. Feigenbaum, "*Information Processing 68*," North Holland, Amsterdam (1968)
[5]A. B. Delfino and A. Buchs, *Helv. Chim. Acta* **55**, 2017 (1972)
[6]C. Djerassi, J. M. Wilson, H. Budzikiewicz and J. W. Chamberlin, *J. Am. Chem. Soc.* **84**, 4544 (1962).
[7]D. H. Smith, B. G. Buchanan, R. S. Engelmore, A. M. Duffield, A. Yeo, E. A. Feigenbaum, J. Lederberg and C. Djerassi, *Ibid.* **94**, 5962 (1972)
[8]B. G. Buchanan, E. A. Feigenbaum and N. S. Sridharan, *Machine Intelligence* 7 (Edited by B. Meltzer and D. Michie), Edinburgh University Press, Edinburgh (1972)
[9]H. Budzikiewicz, C. Djerassi and D. H. Williams, *Mass Spectrometry of Organic Compounds*, Holden-Day, San Francisco (1967)
[10]W. B. Weber, R. A. Felix and A. K. Willard, *J. Am. Chem. Soc.* **92**, 1420 (1970)
[11]J. A. McCloskey, R. N. Stillwell and A. M. Lawson, *Analyt. Chem.* **40**, 233 (1968)
[12]L. Tökés and C. Djerassi, *J. Am. Chem. Soc* **91**, 5017 (1969)
[13]J. H. Beynon, G. R. Lester and A. E. Williams, *J. Phys. Chem. A*, **63**, 1861 (1959)
[14]J. H. Bowie, D. W. Cameron, R. G. F. Giles and D. H. Williams, *J. Chem. Soc.* B, 335 (1966)
[15]R. A. Okerholm, S. J. Clark and H. H. Wotiz, *Analyt. Biochem.* **44**, 1 (1971)
[16]D. H. Smith, A. M. Duffield and C. Djerassi, *Org. Mass Spectrom.* in press.
[17]M. Spiteller-Friedmann and G. Spiteller, *Fortschritte der Chemischen Forschung*, **12**, 440 (1969)
[18]V. I. Zaretskii, N. S. Wulfson, V. L. Sadovskaya, S. N. Ananchenko and I. V. Torgov, *Tetrahedron* **24**, 2339 (1968)
[19]H. Budzikiewicz, *Biochemical Applications of Mass Spectrometry* (Edited by G. R. Waller), p. 251. Wiley, New York (1972)
[20]R. H. Shapiro and K B. Tomer, *Org. Mass Spectrom.* **2**, 579 (1969)
[21]M. Barber and R. M. Elliott, *presented at the Twelfth Annual Conference on Mass Spectrometry and Allied Topics*, Montreal, June (1964)
[22]K. R. Jennings, *J. Chem. Phys.* **43**, 4176 (1965)
[23]J. H. Beynon, R. A. Saunders and A. E. Williams, *Nature* **204**, 67 (1964)