Applications of Artificial Intelligence for Chemical Inference. VIII.[1,2] An Approach to the Computer Interpretation of the High Resolution Mass Spectra of Complex Molecules. Structure Elucidation of Estrogenic Steroids

D. H. Smith,* B. G. Buchanan, R. S. Engelmore, A. M. Duffield, A. Yeo, E. A. Feigenbaum, J. Lederberg, and Carl Djerassi

Contribution from the Departments of Chemistry, Computer Science, and Genetics, Stanford University, Stanford, California 94305. *Received December 20, 1971*

Abstract: An extension of the Heuristic DENDRAL program has been used for automatic interpretation of the complete high resolution mass spectra of estrogenic steroids. The program has been structured for facile adaptation to the general problem of analysis of the high resolution mass spectra of other classes of complex organic compounds. The operation of the program and its performance in interpretation of the spectra of 43 estrogen-related compounds are described. Its performance is comparable to the quality and surpasses in speed the performance of trained mass spectroscopists.

Intelligent computer programs for chemical inference about structure elucidation from mass spectral and other data have yielded promising results.[1b,3] The Heuristic DENDRAL program interprets the low resolution mass spectra of a large variety of saturated, aliphatic, monofunctional compounds in terms of molecular structure. Other approaches to analysis of low resolution mass spectra have included library matching procedures, recently reviewed by Hertz, *et al.*,[4] procedures involving a combination of matching and interpretation,[5,6] and totally empirical learning machines.[7] In this paper we introduce the systematic use of high resolution mass analysis and of metastable ions and their application to a more complex set of structures, the estrogenic steroids.

The interpretive methods used in Heuristic DENDRAL differ from these other methods in several respects: (a) the Heuristic DENDRAL program uses a set of empirical mass spectrometry rules (hereafter referred to as a "theory") which is nearly comparable in complexity to the theory used by human chemists; (b) the program's process of reasoning from data to explanation, using the theory of mass spectrometry, emulates the idealized reasoning process of an experienced mass spectroscopist; (c) the Heuristic DENDRAL program is capable of examining all possible molecular structures within a class of compounds, or fitting an observed molecular formula, in order to choose the best explanation of the data; (d) the program is able to reduce the number of structures actually considered by referencing the data, including other data such as nmr, when available. Using rules of mass spectrometry developed from a small set of spectra of standard compounds, the Heuristic DENDRAL program is able to analyze large numbers of mass spectra correctly.

Because the Heuristic DENDRAL program uses a theory of mass spectrometry in much the same way that mass spectroscopists do, it is possible for chemists to understand the reasoning steps of the program. Thus they can suggest extensions or alternative steps when the program fails to analyze some spectra correctly (a difficulty of some statistical approaches[7]). This is a great advantage for building a powerful program in incremental steps.

Research in most disciplines of organic chemistry involves complex, generally polyfunctional molecules that cannot be analyzed easily by existing computer programs. One direction of future computer research in chemistry will certainly be to develop programs that reason about molecules of greater complexity. Considering automated structure elucidation based, at least in part, on mass spectral data, the multiplicity of possible elemental compositions in low resolution mass spectra of complex molecules is an effective deterrent to the successful implementation of the various approaches mentioned above. It is apparent that the specificity of complete high resolution mass spectra, where elemental compositions of all ions are determined, is required.[8a,9b]

High resolution mass spectra have been utilized in conjunction with computer programs to aid in the structure elucidation of a limited number of types of com-

[1] (a) Financial support for these investigations was provided in part by the National Institutes of Health, Grant No. RR-612-01, the Advanced Research Projects Agency, Grant No. SD-183, and the National Aeronautics and Space Administration, Grant No. NGR-005-020-004; (b) for part VI, see A. Buchs, A. B. Delfino, A. M. Duffield, C. Djerassi, B. G. Buchanan, E. A. Feigenbaum, and J. Lederberg, *Helv. Chim. Acta*, 53, 1394 (1970).

[2] For part VII, see A. Buchs, A. Delfino, C. Djerassi, A. M. Duffield, B. G. Buchanan, E. A. Feigenbaum, J. Lederberg, G. Schroll, and G. L. Sutherland, *Advan. Mass Spectrom.*, 5, 314 (1971).

[3] (a) A. M. Duffield, A. V. Robertson, C. Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg, *J. Amer. Chem. Soc.*, **91**, 2977 (1969); (b) A. Buchs, A. M. Duffield, G. Schroll, C. Djerassi, A. B. Delfino, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg, *ibid.*, **92**, 6831 (1970); (c) G. Schroll, A. M. Duffield, C. Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg, *ibid.*, **91**, 7440 (1969).

[4] H. S. Hertz, R. A. Hites, and K. Biemann, *Anal. Chem.*, 43, 681 (1971).

[5] L. R. Crawford and J. D. Morrison, *ibid.*, **43**, 1790 (1971).

[6] D. H. Smith, *ibid.*, **44**, 536 (1972).

[7] P. C. Jurs, **43**, 1812 (1971), and references cited therein.

[8] (a) A. Mandelbaum, P. V. Fennessey, and K. Biemann, *Proc. Annu. Conf. Mass Spectrom. Allied Topics, 15th*, 111 (1967); (b) R. Venkataraghavan, F. W. McLafferty, and G. E. VanLear, *Org. Mass Spectrom.*, **2**, 1 (1969); (c) W. J. Richter, B. R. Simoneit, D. H. Smith, and A. L. Burlingame, *Anal. Chem.*, **41**, 1392 (1969).

[9] (a) M. Senn, R. Venkataraghavan, and F. W. McLafferty, *J. Amer. Chem. Soc.*, **88**, 5593 (1966); (b) K. Biemann, C. Cone, B. R. Webster, and G. P. Arsenault, *ibid.*, **88**, 5598 (1966).

pounds,[8] most notably, peptides.[9] A submolecular group analysis method, wherein certain combinations of elements are assumed to represent certain portions of molecules, has been proposed as a more general approach.[10] However, there has been little effort toward a systematic approach to computerized interpretation of high resolution mass spectra. The availability of the facilities for structure generation and manipulation in the DENDRAL algorithm,[11] and the ability to encode knowledge about mass spectral fragmentation processes (heuristics, rules) in a simple format that can be handled by the computer, permit an implementation of such a systematic approach. As will be shown in a subsequent publication, utilization of metastable ion data can permit the analysis of mixtures of compounds without prior separation of the constituents.

The program described below is an extension of the Heuristic DENDRAL program previously mentioned. Given information about the basic structural unit of a class of compounds, the fragmentation mechanisms general to that class, a high resolution mass spectrum and metastable ion information, the program attempts to determine a molecular structure(s) to explain the data. One will note that this program is tied more rigidly to the data than the chemist, since it cannot bring to bear on the problem chemical experience for which it has not been programmed. On the other hand, the program is often much more thorough in its systematic search through the data and in its consideration of all combinations of evidence uncovered by it.

The present approach differs from the method of the Heuristic DENDRAL program in at least two important respects. First, this work does not encompass a systematic program for enumerating molecular structures comparable to the DENDRAL acyclic structure generator.[11] The reason for this is that no structure-generating program of sufficient scope and flexibility has yet been written. Second, this work does not encompass a predictor like that of the Heuristic DENDRAL system which allows ranking the final structures. Essentially, the work described here is an extension of the Heuristic DENDRAL system's planning phase.

Methods

The three main parts of the program, labeled ANALYSIS, SYNTHESIS, and FILTER, are described below and shown schematically in Figure 1. Briefly, the overall operation is as follows: (I) identify the molecular ion or ions (if dealing with a mixture); for each molecular ion, find, using the given mass spectrometry theory, all associations of substituents and fragments for which evidence exists in the spectrum; (II) generate all molecular structures consistent with the information in I; (III) apply plausibility criteria to each structure in order to find the best explanation(s) of the data.

Many minor steps are omitted from this discussion, as are all programming details, in the interest of conceptual clarity. A detailed description of the Heuristic DENDRAL structure manipulation algorithms will appear in a subsequent publication.

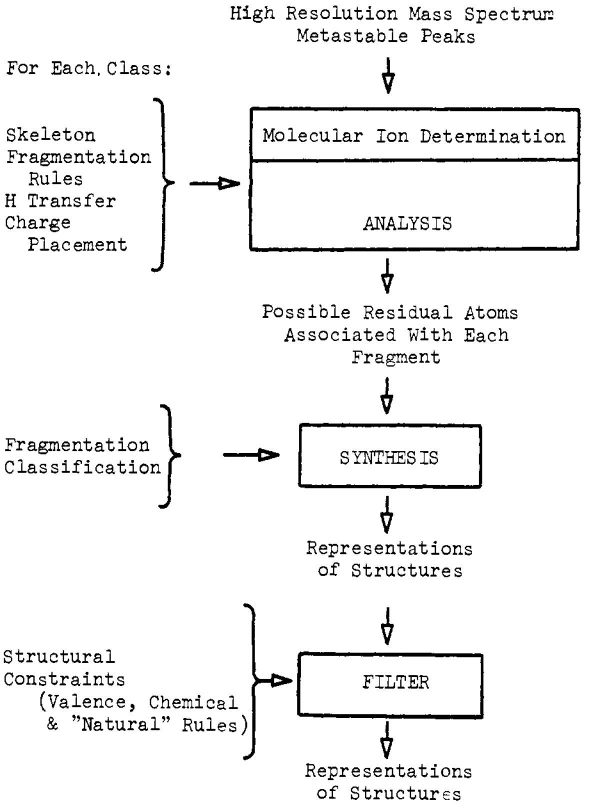


Figure 1. A summary of the major steps involved in the high resolution mass spectral analysis program.

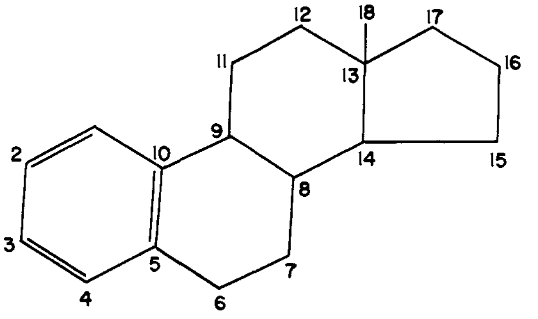
I. ANALYSIS. The required information for the first phase of the program's operation, termed ANALYSIS, is summarized in Figure 1. The spectrum-specific information consists of a complete high resolution mass spectrum in the form of a table of accurate masses, relative ion abundances, and associated elemental compositions, along with available metastable ion information, presently determined by manual examination of a conventional low resolution mass spectrum or by metastable defocusing techniques. The program makes the usual corrections for naturally occurring isotopes (^{13}C, ^{18}O...) on the ion intensities and eliminates ions considered to be in the noise level. The requisite class-specific information (Figure 1) consists of the basic structural skeleton and fragmentation rules general to this class of compounds. First the program determines the molecular ion(s) indicated by the data. Then it searches for substituent placement evidence in the spectrum for each molecular ion.

A. Molecular Ion Determination. Given the structural skeleton, the program calculates the empirical formula, mass, and degree of unsaturation of this skeleton. As an example, the skeleton supplied to the program for estrogens is depicted in Figure 2. It has the indicated empirical formula $C_{18}H_{24}$ (*m/e* 240) and thus seven degrees of unsaturation (rings plus double bonds). Any ion in the high resolution mass spectrum of this mass or higher mass with at least this number of carbon atoms is considered a molecular ion candidate.

A candidate molecular ion is eliminated if its ion intensity can be wholly accounted for either as loss of hydrogens from another candidate at higher mass, or as contributions from isotopes of carbon, oxygen, and so forth, from another candidate at lower mass.

[10] A. Kundered, R. B. Spencer, and W. L. Budde, *Anal. Chem.*, **43**, 1086 (1971).

[11] J. Lederberg, G. L. Sutherland, B. G. Buchanan, E. A. Feigenbaum, A. V. Robertson, A. M. Duffield, and C. Djerassi, *J. Amer. Chem. Soc.*, 91, 2973 (1969).



ESTROGEN SKELETON $C_{18}H_{24}$

Figure 2. The basic estrogen skeleton supplied to the ANALYSIS phase of the program.

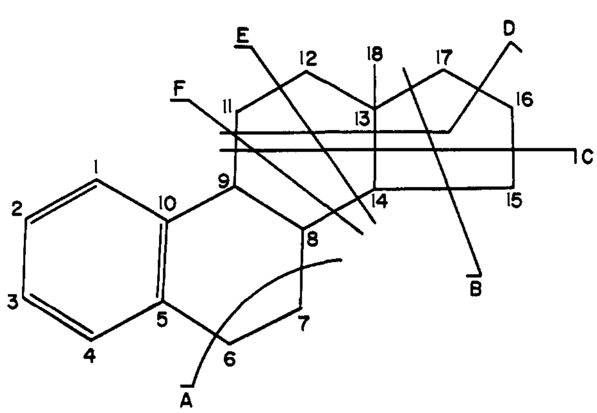


Figure 3. Symbolic representation of the fragmentation rules employed by the program for estrogens.

Finally, the program eliminates any candidate which shows no metastable transition to any daughter. That is, if there is no metastable evidence that a candidate is a parent of some other ion found in the high resolution mass spectrum, that candidate is eliminated. These general rules are usually strong enough to determine the molecular ions in the spectrum. Occasionally, spurious molecular ions will be inferred from a given metastable transition because of an accidental numerical relationship between two ions. On the other hand, the program has correctly inferred, from the high resolution mass spectrum, that a sample, originally thought to be a pure compound, was in fact a mixture (see compounds **20a** and **21a** below).

B. Search for Substituent Placements. The most important phase of the program is the analysis of the spectrum in light of the given structural skeleton and its fragmentation pattern. Each fragmentation (break) of the skeleton is considered in turn. Any evidence in the spectrum for a given break, including the allowed hydrogen rearrangements, or transfers, is saved. The breaks which the program considers for estrogens are shown schematically in Figure 3, and associated hydrogen transfers are summarized in Table I and discussed in detail in the Results and Discussion.

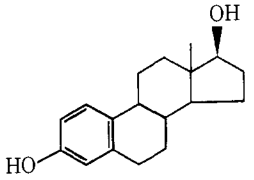
Table I. Class-Specific Input to ANALYSIS for Estrogens

Estrogen skeleton

C=C · C=C · C · C · C · C · C· ·C C · C · C· ·C · C · C · C C
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
(1 10) (5 10) (5 10) (8 14) (13 17)

Estrogen breaks		Transfer	Place
Name	Break bonds	hydrogens	charge
B	(14·15) (13·17)	(−1 0)	14
C	(9·11) (14·13) (15·16)	(−1 0)	9
D	(9·11) (14·13) ·(1617)	(−2 −1)	9
E	(11·12) (8·14)	(−1 0)	11
F	(9·11) (8·14)	(−1 0)	9

Since the objective of the program is to determine the location of the substituents on the skeletal frame, this information is, of course, not yet known. Consequently, the correct ion in the spectrum, corresponding to a particular break and its associated hydrogen transfers, has the empirical formula of at least the skeletal fragment and at most the skeleton plus all residual atoms and unsaturations. For example, the break labeled B in Figure 3 indicates loss of three carbons, six hydrogens, and any substituents, resulting in a charged fragment encompassing 15 carbons, 18 hydrogens ($C_{15}H_{18}$), and the remaining substituents. Estradiol (**24**),[12] a simple example, has the molecular formula $C_{18}H_{24}O_2$. By difference from the skeleton, $C_{18}H_{24}$, the substituents that must be placed are two oxygens. If break B is assumed to occur as depicted in Figure 3 along with an associated loss of a hydrogen atom (see Table I), the following ion series must be considered by the program as providing evidence for substituent placement based on break B: ion series 1, $C_{15}H_{18}$; $C_{15}H_{17}$ (no substituents on the charged fragment); ion series 2, $C_{15}H_{18}O_1$; $C_{15}H_{17}O_1$ (one oxygen substituent on the charged fragment); ion series 3, $C_{15}H_{18}O_2$; $C_{15}H_{17}O_2$ (both oxygen substituents on the charged fragment).



24

Since, in this example, there are no extra carbon atoms in the molecular formula, and no extra degrees of unsaturation, the program has a short list of possibilities to consider. The list grows rapidly as the number of substituents increases.

Typically, more than one ion can be found as evidence for a given fragmentation. Several ions in an ion series may occur, and more than one ion series may be indicated. The program must decide which of the possible substituents are strongly enough indicated to be saved as alternative results for the break. It does this by summing the ion intensities of all ions in each ion series and ordering the ion series by decreasing total intensities. For some breaks, only the most strongly indicated ion series will be saved. For most, however, every ion series whose total intensity is above a calculated threshold is saved. (The ion series thresh-

[12] Common names. For the systematic names of the parent compounds see Table V.

old for a break is currently one-third of the intensity of the strongest ion series.)

If there is no evidence for a break, even with any of the indicated hydrogen transfers, the program continues its analysis without that break information. Trouble can come when an incorrect ion of high intensity masquerades as the ion resulting from a break while the correct ion is absent. If the correct ion is also present, however, the program can usually determine the correct structure. The information from the masquerading ion may cause the program to produce alternative structures for its final answer, but often this ion information is inconsistent with information from the other breaks and is effectively discarded.

For each break, the program saves the possible substituents with the positions on the skeleton where the substituents might be placed. For example, if ion series 1 above were below the calculated threshold, leaving series 2 and 3 as possibilities for break B in estradiol (**24**), the program would save the following three items of information: (i) BREAK B; (ii) (O2 $X\%$) or (O1 $Y\%$); (iii) skeletal fragment atoms = (C-1–C-14, C-18). Later, the program may decide to reconsider ion series 1 if the possible substituents for all fragments (results of all breaks) cannot be placed consistently.

II. SYNTHESIS. **Combination of Possible Substituents.** After the break-by-break analysis, the program has a list of possible substituents for each of the fragments. It must now consider the consequences of every combination of substituents, taking one set of substituents from each break result. The procedure by which this is accomplished is termed SYNTHESIS, and is pictured in a highly schematic fashion in Figure 1. This phase of the program operates on the lists of possible substituent placements from ANALYSIS. The only class-specific input is a break classification, specifying use of either the most intense ion series for a given break, or use of all evidence for this break.

As the number of possibilities for each break increases, the number of combinations increases rapidly. Therefore, any heuristics that will reduce the number of possibilities for any break will save considerable computing effort. One such heuristic is to give very high priority to the most intense ion series for some breaks (break classification). However, this can be done only if the mass spectral theory assures the priority of those breaks. Another heuristic is to set a high threshold for the total intensity of the ion series to be saved. But this, too, works only when the breaks are known to be preferred (if not the most preferred) a large part of the time.

Still another way of reducing the number of possibilities is to apply the FILTER rules (discussed in the subsequent section) at this stage. The natural rules, discussed under FILTER, are presently used in this manner.

The result of combining substituents and attachment sites within the given constraints is a list of molecular models, each indicating a possible placement of substituents around the skeleton.

III. FILTER. **Exclusion of Structures.** The reason for the large number of possible structures, particularly as the number of substituents increases, is that ANALYSIS and SYNTHESIS are immanently unrestricted. Such restrictions are saved for the final phase of the program, termed the FILTER. The FILTER (Figure 1) contains a variety of rules, some of which are general in nature and others which are very specific to the compound class. These rules can shorten drastically the list of structures presented as possible solutions.

Valence considerations are checked in the last phase of the program with one set of filtering rules. These rules, which are quite general, prohibit chemically implausible combinations of substituents placed on various skeletal atoms (analogous to BADLIST[11]). The illegal structures are filtered out retrospectively by testing each structure against a list of rules. For example, the SYNTHESIS phase of the program does not check initially to see that double bonds have some atom at the other end; unaccompanied double bonds are screened out by FILTER. When in the future a general cyclic structure generator becomes available, analogous to the acyclic structure generator,[11] this set of rules will be unnecessary as a retrospective filter.

A second set of FILTER rules is termed "chemical." These rules can be structured to encompass chemical information, if available, from a variety of sources. Knowledge of the chemical synthesis or isolation procedures for a given sample in many cases permits clustering of various substituents in the substituent lists (analogous to GOODLIST[11]). For example, if a compound bearing a hydroxyl group was acetylated, one would expect the atoms $-OCOCH_3$ to be present together at a skeletal atom. Similarly, ir, uv, or nmr information can be encoded in these rules to aid in identifying substituent clusters. Isotopic labeling information can also be included. For example, the facility exists for specifying the number of labile hydrogens, *e.g.*, –OH, associated with each molecular ion, presently determined by a low resolution mass spectrum of the deuterium-exchanged compound or mixture of compounds. Again, structures which do not meet the requirements imposed by this additional chemical information are filtered out retrospectively.

A third set of FILTER rules defines properties of naturally occurring compounds. It may be known, for example, that for the class of molecules under consideration, oxygen substituents, when present, are always found on places x, y, and z; similarly, it may be known that carbon substituents are never found on places v or w.

Feedback Loops in the Program. If the program fails to construct even one structure, it shares the dilemma of the mass spectroscopist who uses a certain set of fragmentation rules. Either the rules must be modified or additional evidence sought. In this case, there is opportunity for modification of the rules. The first modification is, presumably, the first modification that would be made under manual interpretation. In a second pass through SYNTHESIS, the program considers evidence originally rejected by the break classification. If this second pass still yields no structures from FILTER, the evidence criteria are relaxed still further by returning to ANALYSIS for a third pass and lowering the calculated minimum total intensity which an ion series must have to be considered. The minimum may be any nonzero sum, which allows every initial possibility to be considered. Because there are so many combinations of possibilities, this should be regarded as a last resort. If no structures result from this third pass, the program would admit defeat.

It should be noted that the program as now constituted terminates at the end of any pass when the list of filtered structures is not empty. Also, the probability of finding structures satisfying the data in addition to or in place of the correct structure increases as additional passes are made and low intensity ion series allowed in as evidence. The latter problem, termination with an incorrect structure, is not necessarily unique to the program. Chemists, as well as the program, can make incorrect choices when the data are ambiguous (see Performance section below).

Flexibility of the Program. The computer program has been written to allow for easy extensions to other classes of molecuels. In order to switch from analysis of one class of compounds to another, it is necessary only to change the skeleton and fragmentation rules given to the program and indicate new break priorities and structural constraints. Experiments are under way to evaluate the performance of the program for other classes of compounds in addition to the estrogens described below in order to test this hypothesis.

In addition, the following general features of the program should be noted, which allow extension to other classes of molecules. (a) The program does not assume that the data are derived from a pure compound. It looks for as many molecular ions as can satisfy its criteria. (b) The unknown compound may contain any elements, not just C, H, and O. (c) The skeleton may be of any complexity and may contain heteroatoms at any position. (d) Functional groups, or substituents, about the basic skeleton may be of any complexity, providing that the mass spectral theory remains valid. (e) The data need not show every break specified by the theory. The program continues its analysis with as much information about the structure as it can find in the data. (f) Additional information from other spectroscopic or chemical techniques may be encoded with relative ease. This information, though helpful, is not essential to the operation of the program.

It is possible also for the chemist to change the operation of the program by exercising any of several options. He can alter the weight given to metastable information, *e.g.*, he can instruct the program to ignore all ions except those which show a metastable transition from the molecular ion, or he can require that the program ignore metastable information in its consideration of breaks. The chemist can reset the thresholds for deleting ions in the noise level and for ignoring ion series with low combined intensities. It is possible also for the chemist to change the rules governing fragmentation, hydrogen transfer, priority of breaks, and exclusion of structures.

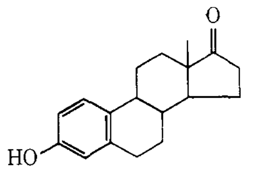
Results and Discussion

Estrogenic steroids represent the first class of relatively complex molecules whose high resolution mass spectra have been analyzed by the previously described program. This class of compounds was chosen both because of its biological importance and because previous reports concerning the fragmentation mechanisms of estrogens indicated strong correlations between mass spectra and structure.[13,14]

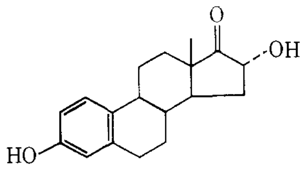
Input. The basic estrogen skeleton considered by the program is shown in Figure 2 and has the indicated composition $C_{18}H_{24}$. One may note that the generally expected oxygen placements at C-3 and C-17 are not included in the skeleton, so as to accommodate corresponding modifications of an estrogen. If desired, this restriction could be implanted, and would make the program more efficient in analyses where these oxygens can be postulated *a priori.* The formal representation of the basic fragmentations initially considered by ANALYSIS is presented in Figure 3 (fragmentation A has been found to be unreliable and is not presently used). This representation, based on mechanisms postulated from previous studies,[13,14] is sufficient to characterize the fragmentations until such time as the more detailed postulates can be proven.

Any carbon atom that may be effectively "severed" from the skeleton by the operation of one or more breaks (B–F, Figure 3) is a unique position at which a substituent can be placed unambiguously. For estrogens, these positions are C-11, C-14, C-15, C-16, and C-17. For C-12, 13, 18, and C-1 through C-10 the situation is quite different. Mass spectrometry alone, given this set of postulated mechanisms, is not capable of differentation, for example, among C-1 through C-10, so that this group of atoms must be labeled collectively.

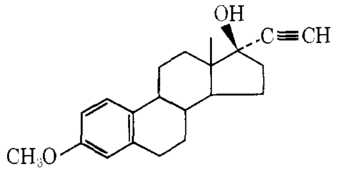
A summary of the above input data may be found in Table I. The string of 18 carbon atoms given for the skeleton is the DENDRAL notation for cyclic molecules[15] and represents a linear chain of 18 atoms with ring closures noted as extra bonds. The strings of numbers after each skeleton break (B–F listed in Table I) indicate the bonds that are broken, hydrogens transferred (negative value indicating away from the charged fragment), and charge location, which may be any atom remaining on the charged fragment. As an example, break B implies cleaving the C-13–C-17 and C-14–C-15 bonds, with associated transfer of either zero or one hydrogen atom and the charge remaining somewhere on the fragment containing C-14. This representation of structure and fragmentations is kept as simple as possible to permit facile change either to test the program's performance employing additional or different rules, or to change the compound class considered.



1, O2 DOT2



14, O3 DOT2



39, C3 O2 DOT4

[13] H. Budzikiewicz, C. Djerassi, and D. H. Williams, "Structure Elucidation of Natural Products by Mass Spectrometry," Vol. 2, Holden-Day, San Francisco, Calif., 1964.

[14] C. Djerassi, J. M. Wilson, H. Budzikiewicz, and J. W. Chamberlain, *J. Amer. Chem. Soc.*, **84**, 4544 (1962).

[15] Y. M. Sheikh, A. Buchs, A. B. Delfino, G. Schroll, A. M. Duffield, C. Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg, *Org. Mass Spectrom.*, **4**, 493 (1970).

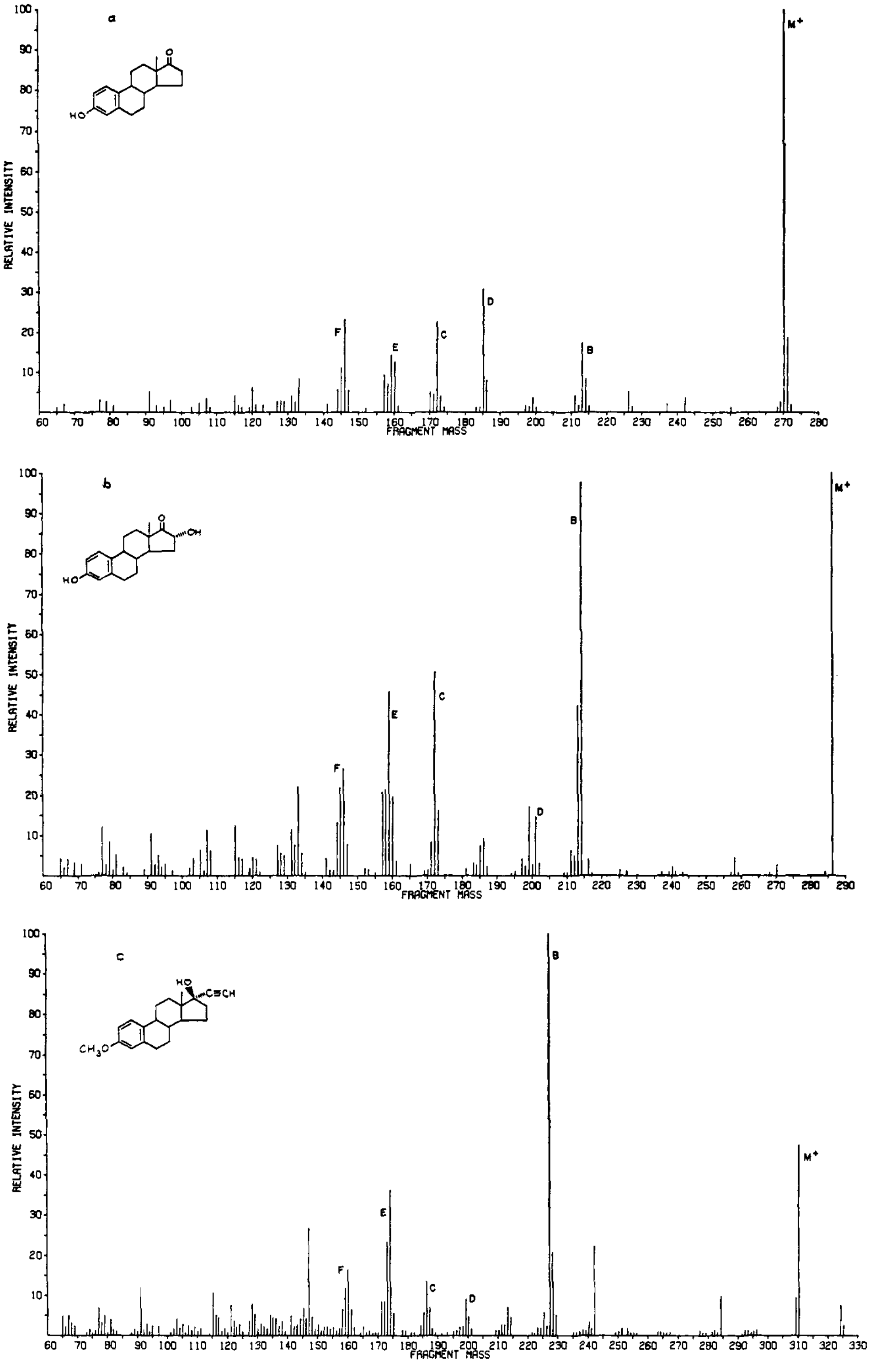


Figure 4. (a) Low resolution mass spectrum of estrone (1); (b) low resolution mass spectrum of 16α-hydroxyestrone (14); (c) low resolution mass spectrum of 17α-ethinylestradiol 3-methyl ether (39).

The conclusions of each of the three phases of the program are best illustrated with the use of examples. Data from three samples of varying complexity are presented, those for estrone (**1**), 16α-hydroxyestrone (**14**), and 17α-ethinylestradiol 3-methyl ether (**39**).[12]

ANALYSIS. The substituents that must be placed on the estrogen skeleton are given in association with structures **1**, **14**, and **39**. These values have, of course, been obtained by subtraction of the skeleton composition, $C_{18}H_{24}$, from the inferred molecular formula for each compound. Note that a hydrogen deficiency is designated by a DOT, which is regarded as a substituent that must be placed just as any other atom. The low resolution mass spectra, obtained from appropriate summation of all ions from the high resolution mass spectra, are presented in Figure 4. In each spectrum the ions resulting from the general fragmentations B–F, as determined by accurate mass measurement, are labeled accordingly.

The conclusions of ANALYSIS for the above three compounds are summarized in Table II. The substituent lists following each break are printed by the pro-

Table II. Conclusions of ANALYSIS for Compounds **1**, **14**, and **39**

Compd	Substituents	Break	Substituents on charged fragment	Summed rel abundance, %
1	O2 DOT2	B	O1	23
		C	O1	23
			O1 DOT2	9
		D	O1	35
		E	O1	25
			O1 DOT2	16
		F	O1	33
14	O3 DOT2	B	O1	133
		C	O1	60
		D	O2	16
			O1	15
			O1 DOT2	5
		E	O1	58
			O1 DOT2	40
		F	O1	45
39	C3 O2 DOT4	B	C1 O1	104
			C2 O1	23
			O1	10
		C	O1	30
			C1 O1	18
			C2 O1 DOT2	11
			O1 DOT2	9
			C3 O1 DOT2	9
			C1 O1 DOT2	8
			C2 O1	5
			C3 O1	4
		D	C3 O1	104
			O1	18
			C1 O1	12
			C2 O1	10
		E	C1 O1	56
			O1	26
			C2 O1	6
			C3 O1	1
		F	C2 O1	56
			C1 O1	26
			O1	10
			C3 O1	6

gram to show that evidence has been found in the mass spectrum for those substituents on the charged fragment. In the case of estrone (**1**), for example, ions possibly resulting from break B have been found that represent the presence of a single substituent, an oxygen. By subtraction, the other oxygen and the unsaturation (two DOT's) have been lost to the neutral fragment from this break. For break C in **1**, however, there are two possibilities for substituent placement, either an oxygen, or an oxygen and an unsaturation on the charged fragment. As mentioned previously, as the number of substituents increases, so do the possibilities for a particular break, as evidenced by the list of eight choices for break C in compound **39** (Table II).

SYNTHESIS. The substituent lists show all candidate substituent placements for all breaks, under the constraints of ANALYSIS described previously. There is, however, no consistency check on these results. The number of combinations of substituent placements considered by SYNTHESIS based on ANALYSIS conclusions to attempt to generate a structure are: for **1**, 4; for **14**, 6; for **39**, 1536. The combinations for compounds such as **39** can be carried out efficiently only with the use of a computer. At this point, a chemist would seek additional information or make educated guesses to limit the problem, whereas the computer can accomplish an exhaustive combination of the possibilities very quickly. Fortunately, not all of the combinations will, in general, yield a structure.

The only class-specific input to the SYNTHESIS phase for estrogens is a classification of break B. Break B, formal scission of ring D from the structure, was initially assumed from previous experience[14] to be the most reliable fragmentation of estrogens. Reliability in this case refers to the generally high abundance of the break B fragment ions relative to other fragment ion abundances. As a first approximation, then, SYNTHESIS considers only the top entry (highest summed relative ionic abundances) in the break B substituent list from ANALYSIS (Table II). All evidence for breaks C–F appearing in the ANALYSIS substituent list is considered. The structures generated by SYNTHESIS for compounds **1**, **14**, and **39** are listed in Table III (only six of the 24 structures generated for **39** are presented to conserve space). There are considerably fewer structures than there are combinations of break evidence. For example, only one combination of break evidence for **1** yields a structure. The output format for these structures consists of a substituent followed by a carbon number placement. The structure for **1** then reads: "an oxygen on C-1 through C-10, an oxygen on C-17, and an unsaturation on C-17."

Structures exemplified by **39**, which contain extra carbon atoms, present ambiguities. The pair of breaks C–D differs effectively by a single carbon atom, as does

Table III. SYNTHESIS Conclusions. Possible Structures for Compounds **1**, **14**, and **39**

Compd	Structure no.	Substituents	Placement
1	1	O1	C-1–C-10
		O1 DOT2	C-17
14	1	O1	C-1–C-10
		O1 DOT2	C-17
		O1	C-16
	2	O1	C-1–C-10
		O2 DOT2	C-17
	3	O1	C-1–C-10
		O2	C-17
		DOT2	C-16
39	1	C1 O1	C-1–C-10
		O1 DOT4	C-17
		C2	C-15
	5	C1 O1	C-1–C-10
		O1 DOT4	C-17
		C2	C-16
	9	O1	C-1–C-10
		O1 DOT4	C-17
		C2	C-15
		C1	C-12, 13, 18
	13	O1	C-1–C-10
		O1 DOT4	C-17
		C1	C-15
		C1	C-16
		C1	C-12, 13, 18
	17	C1 O1	C-1–C-10
		C2 O1 DOT4	C-17
	21	O1	C-1–C-10
		C1 O1 DOT4	C-17
		C1	C-16
		C1	C-11

the pair E–F. In the spectrum of any estrogen possessing extra carbon atoms (>18), evidence can be found in the high resolution data for placement of these extra carbons at a wide variety of positions. As an example, evidence for an extra carbon associated with the charged fragment from break F immediately causes a redundancy with break E, as the same ions will represent E without the extra carbon. Thus, the substituent placements for **39** (Table II), derived from breaks E and F, which differ by a single carbon, *i.e.*, C1 O1 *vs.* C2 O1, O1 *vs.* C1 O1, C2 O1 *vs.* C3 O1, arise from the same ions, as indicated by the identical summed relative abundances. This ambiguity allows generation of structures for **39** with carbon substituents scattered about the skeleton, as shown in Table III. There is, as yet, no simple test by which the correct choice for substituent placement for each break can be made. It is clearly not the most intense ion series, as is evidenced by the entries for breaks, C, D, and F in compound **39**, Table II.

FILTER. These structures (Table III) are passed to FILTER for review. In addition to valence and chemical rules that may be brought to bear in examination of these structures, there are some "natural" rules that may be used. The naturally occurring estrogens isolated during the course of *in vitro* or *in vivo* studies represent a very restricted set of structural types.[16] In particular, positions of biological methylation are at C-2 and C-3 only (OH → OCH_3). Because synthetic estrogenic contraceptives are functionalized with extra carbons at C-17 (17α-ethinyl derivatives), however, this position can also be considered for placement of extra carbons under a "natural" rule when dealing with estrogens isolated from females who have taken these contraceptives. The most powerful "natural" rule, then, passes through FILTER only those structures with all extra carbons placed on the C-1 through C-10 or C-17 atoms. This rule is, of course, relaxed if one is dealing with a synthetic sample from an unknown source.

Employing the above rules with mass spectrometric data only (*i.e.*, no chemical rules derived from other techniques) for the examples **1**, **14**, and **39** yields the results summarized in Table IV. In each case, the SYNTHESIS output is reduced to a single correct structure. Note that structures 2 and 3 for **14** (Table III) are removed by valence considerations, the former because the only method for placing two oxygens and an unsaturation at C-17, a carbon with only two free valences, is a three-membered ring including two oxygens. The latter is removed because there is no additional substituent at C-16 to which an unsaturation can be attached. For **39**, the natural rules for carbon placement in addition to valence rules serve to remove all structures but number **17** (Table III), the correct one.

Table IV. Conclusions of FILTER for Compounds **1**, **14**, and **39**

Compd	Structure no.	Substituents	Placement
1	1	O1	C-1–C-10
		O1 DOT2	C-17
14	1	O1	C-1–C-10
		O1 DOT2	C-17
		O1	C-16
39	1	C1 O1	C-1–C-10
		C2 O1 DOT4	C-17

These results are characteristic of many of the estrogen standards studied. In some cases, however, structures cannot be obtained due to the rigid classification of break B or the diminished intensity of ions arising from breaks C–F. For example, although break B normally yields an intense ion or pair of ions in the mass spectra of estrogens, certain derivatives, *e.g.*, 11-oxo (**11**) or 11-hydroxy (**12**, **30**, **31**) compounds, yield mass spectra wherein break B yields ions of greatly diminished intensity.

Upon failure to generate any plausible structures on the first pass, the next step is to send the program through the feedback loops as described in the Methods section. The break B classification is revised first. If no structures are obtained, all information for all breaks is considered. This sequence of operations yields the performance summarized below.

Performance. The performance of the program based on analysis of 43 estrogen standards is summarized in Table V.

The designation of hydrogen deficiencies as DOT's permits flexibility in specification of structures containing unsaturations. Substituents containing double bonds or rings are inferred by the occurrence of DOT's in pairs. For structures possessing intramolecular double bonds, the SYNTHESIS and FILTER phases of the

[16] (a) H. Adlercreutz, *J. Endocrinol.*, **46**, 129 (1970); (b) H. Adlercreutz and T. Luukkainen, *Ann. Clin. Res.*, **2**, 365 (1970).

Table V. Summary of Results Obtained for Estrogen Standard Compounds

Derivatives	No. of structures	One correct?	How obtained[a]
Of estrone (3-hydroxy-1,3,5(10)-estratrien-17-one)			
1, Estrone	1	Yes	Normal
2, 2-Hydroxy-	1	Yes	Normal
3, 2-Methoxy-	1	Yes	Normal
4, 3-Methoxy-	1	Yes	Normal
5, 1-Methyl-	1	Yes	Normal
6, 1-Methyl-3-methoxy-	1	Yes	Pass 2
7, 1,2-Dimethyl-	1	No[d]	Normal
8, 1-Methyl-6α,7α-dihydroxy-	2	No[d]	Pass 2
9, 6-Methyl-	1	Yes	Pass 2
10, 7-Oxo-	2	Yes	Pass 2
11, 11-Oxo-9β-	5	Yes	Pass 3
12, 3-Methoxy-11α-hydroxy-	1	No[d]	Pass 2
13, 15α-Hydroxy-	2	Yes	Pass 3
14, 16α-Hydroxy-	1	Yes	Normal
15, 6-Dehydro-	1	Yes	Normal
16, 7-Dehydro- (equilin)	1	Yes	Normal
17, 1-Methyl-6-dehydro-	1	Yes	Normal
18, 6-Methyl-6-dehydro-	1	Yes	Normal
19, 6-Dehydro-8-dehydro- (equilenin)	0		Pass 3
20a, 9,11-Dehydro- (mixture with **20b**)	5	Yes	Pass 2
20b, Estrone	1	Yes	Normal
21a, 3-Methoxy-9,11-dehydro- (mixture with **21b**)	4	Yes	Normal
21b, 3-Methoxy-	1	Yes	Pass 2
22, 1-Methyl-3-methoxy-9,11-dehydro-		*b*	
23, 17-Deoxo-	1	Yes	Pass 2
Of estradiol (1,3,5(10)-estratriene-3,17β-diol)			
24, Estradiol	1	Yes	Normal
25, 2-Hydroxy-	1	Yes	Normal
26, 2-Methoxy-	1	Yes	Normal
27, 3-Methoxy-	1	Yes	Normal
28, 1-Methyl-	1	Yes	Normal
29, 6-Oxo-	1	Yes	Normal
30, 11α-Hydroxy-	2	Yes	Normal
31, 3-Methoxy-11α-hydroxy-	1	No[d]	Normal
32, 3-Methoxy-15α-hydroxy-	1	No[d]	Pass 2
33, 16-Oxo-	3	Yes	Pass 3
34, 17α-Methyl-	1	Yes	Normal
35, "17α-Acetyl-"[c]	2	*c*	Pass 2
36, 1-Methyl-17α-acetyl-	3	Yes	Pass 2
37, 3-Methoxy-17α-vinyl-	1	Yes	Normal
38, 17α-Ethinyl-	1	Yes	Normal
39, 3-Methoxy-17α-ethinyl-	1	Yes	Normal
40, 1-Methyl-6-dehydro-	1	Yes	Normal
41, 1,2-Dimethyl-6-dehydro-	1	No	Normal
42, 9,11-Dehydro-	3	Yes	Normal
43, Estriol (1,3,5(10)-estratriene-3,16α,17β-triol)	1	Yes	Normal

[a] Normal: standard, one pass processing through ANALYSIS, SYNTHESIS, FILTER. Pass 2: recycled through SYNTHESIS with break B classification relaxed. Pass 3: recycled through entire program to include all evidence for all breaks. [b] See text for description. [c] Compound is not estradiol 17α-acetate. The program indicates there is an extra unsaturation, possibly in ring C. The true identity of the sample is not known at this time. [d] Incorrect structures could be removed by more powerful FILTER rules. See text for discussion.

program, by determining which skeletal atoms are adjacent to one another and have the appropriate free valence, will allow candidate structures with single DOT's on adjacent atoms. Thus, correct structures for compounds **20a**, **21a**, and **42** are inferred by the presence of single DOT's at C-1–C-10 (C-9) and C-11, respectively. Compound **22**, although possessing evidence in its spectrum for the correct structure, was not processed past the point of inference of an intramolecular double bond associated with correct placement of the remaining substituents (evaluation of the more than 100,000 possible structures would have resulted in excessive use of computer time).

These results, Table V, fall into some distinct categories, depending entirely on the quality of the mass spectrometry rules supplied to the program. Although these rules (breaks B–F) are reasonably general for this entire class, as evidenced by the large number of compounds processed correctly in the normal, single pass operation of the program, they are not completely general. The mass spectra of several compounds (those obtained by pass 2, Table V) yielded no structures on normal operation, but a second pass after relaxation of the break B classification yielded structures. Evidence for some of the breaks was severely diminished in several spectra, *e.g.*, compounds **11**, **13**, and **33**, so that a third pass was required to generate any structures. There is one example where no structure was obtained, for compound **19**. In addition, there are six compounds for which incorrect structures were obtained, **7**, **8**, **12**, **31**, **32**, and **41**, because an incorrect placement of substituents violated no FILTER rules, even though the spectra of **7**, **12**, **31**, **32**, and **41** contain evidence, albeit weak, for a correct structure. Additional filter rules could eliminate these incorrect structures in all cases but **41** to permit further analysis of the data. In particular, isotopic labeling with deuterium (see above) coupled with an algorithm to construct and test structural fragments from substituent lists would yield improved performance for these compounds. This more comprehensive solution would be provided by a general cyclic structure generator.

Manual examination of the spectra of the compounds yielding anomalous results indicates that competitive fragmentations are operative. In the case of several of the hydroxyl-substituted compounds such as **8**, **12**, **31**, and **32**, it would appear that the loss of the elements of H_2O followed by other fragmentations effectively suppresses one or more of the breaks B–F. Interestingly, break B is also suppressed by alkyl substitution in ring A. Thus, with two extra carbons present, ions from an intense break D masquerade as ions from break B. This results in an incorrect structure for **7** and **41**, with the extra carbon atoms placed at C-17. It would seem that manual interpretation of these spectra, based on present knowledge of estrogen fragmentation rules, will yield the same incorrect conclusions as the program.

In the case of **19**, break F, for example, produces no ions in the spectra due to the low probability of breaking the two bonds adjacent to the aromatic ring B. Although incorrect evidence is present for these breaks, no structures consistent with the data can be constructed.

An interesting feature of the computational procedures is that even when breaks are suppressed in a spectrum, leading to associations of incorrect substituent placements with certain breaks, few consistent structures compared to the possible combinations of evidence can be generated. Even fewer of these (frequently none) can pass rather basic FILTER rules.

Difficulties. Some of the difficulties which the program has encountered thus far are difficulties which chemists also face. Some are easily avoided by humans, but not by the program. The following are difficulties which chemists also face.

(1) When there are a large number of ions in the high resolution mass spectrum which could have resulted from one of the specified fragmentations, the analysis must consider many if not all of the possibilities. Even in the case where the most intense ion series usually indicates the correct set of substituents on the charged fragment (break B), it may be necessary to consider a less intense ion series as the correct one.

(2) If the theory for the class of molecules is not accurate, the results of the analysis will be unreliable. When this investigation was started, the program was instructed to consider break A (losing C-6 and C-7, Figure 2) as a firm fragmentation rule.[14] However, not every estrogen spectrum shows evidence for break A. Thus the program would often find spurious ions for this break and would have trouble finding a consistent way of allocating substituents to the skeleton using information from all breaks. The program's performance improved considerably when break A was omitted from its list of rules.

(3) Mixtures containing very low concentrations of one or more of the compounds will be difficult to analyze. Many important ions in the mass spectra of the low concentration compounds will be in the noise level, if present at all. Therefore, many of the program's choices for evidence ions will be spurious. Moreover, for some breaks the program may find no evidence at all.

(4) If the data are unreliable, for any of a number of reasons, the resulting analysis will also be unreliable. Chemists are more adept than the program at detecting this situation, but there is very little processing that one can do to convert bad data into reliable conclusions.

(5) Sufficient experimental work has not yet been done to allow differentiation between epimeric, underivatized estrogens. (Trimethylsilyl derivatives have been used for differentiation by low resolution mass spectrometry[17].)

Two difficulties which are faced by artificial but not human intelligence are the following.

(1) The program's algorithm for identifying molecular ions depends upon the existence of appropriate metastable ions. This general algorithm has not, in fact, been enhanced with class-specific information. For example, chemists would have little trouble recognizing the molecular ion of an estrogen. For other classes of compounds or mixtures of estrogens, both the computer and the chemist may be faced with similar problems.

(2) The program now operates with a rather limited theory of mass spectrometry. It does not now have, for example, rules allowing it to consider the effects of loss of neutral molecules such as water. However, it is not clear that even chemists would be able to carry out correct structural analyses of estrogens when mechanisms such as significant loss of H_2O are operative, as the mass spectrometry rules may change under these conditions.

Conclusions

Results obtained in this initial effort to employ intelligent programs for structure elucidation of complex organic molecules are very promising. The performance of the program is comparable with the performance of a trained mass spectroscopist, although much faster. It is not clear at this point whether, with the limited information available about the mass spectral fragmentation of estrogens,[13,14] a trained mass spectroscopist can do significantly better than the program. It is apparent that the operation of the program can be improved considerably by the inclusion of a cyclic structure generator to allow proper representation of GOODLIST and BADLIST.[11] In addition, the availability of a large collection of high resolution mass spectra of standard compounds will permit development of a more complete set of fragmentation rules to further improve performance. For example, rules that are strongly dependent on substituent placement can be employed retrospectively. The mass spectrum of a candidate structure showing this placement can be examined to determine if the rules are followed and the structure saved or discarded accordingly.

Work on estrogen mixture analysis is now under way. The operation of the program proceeds in the same manner whether one is dealing with a single compound or a mixture of compounds. All inferred molecular ions are processed one at a time to attempt structure determination. With a complex mixture, however, considerably more metastable ion information is desirable, both to identify molecular ions and to match daughter ions with their respective parents. Simple examples encountered in this work are 9,11-dehydroestrone (**20a**) and 9,11-dehydroestrone 3-methyl ether (**21a**).[12] Both compounds were inadvertent mixtures with estrone and estrone 3-methyl ether, respectively, based on mass spectral evidence. In each case, two molecular ions were found and structures of each determined properly in the expected manner (see Table V) in these simple cases without any additional metastable information.

This approach, with its flexibility and speed, has great potential for handling diverse classes and mixtures of compounds of timely importance in many disciplines of chemistry.

Experimental Section

High resolution mass spectra were acquired with either an Associated Electrical Industries MS-9 or a Varian MAT-711 mass spectrometer operated under the following conditions: 70-eV ionization voltage; 32 sec/decade in mass scan rate; 10,000 resolving power; mass measurement accuracy better than 10 ppm. The mass spectrometers were interfaced to the IBM 360/50 time-shared computer, Stanford University Medical School ("ACME" facility), utilizing a Digital Equipment Corp. PDP-11 as a preliminary data processing device.

Spectra may be transferred from the ACME computer facility to the campus computer center *via* the PDP-11 and telephone lines. Although this transmission is quite slow (15–20 min), results of an analysis can be obtained in less than 30 min from time of sample introduction to final output.

The computer program described here is written in LISP for the IBM 360/67 computer at Stanford University and runs in batch mode. Execution times and memory requirements vary from about 10 sec and 300,000 bytes for the simpler structures to about 1000 sec and 600,000 bytes for the most complex structures examined thus far.

Acknowledgments. We wish to thank Mr. Robert Ross and Miss Annemarie Wegmann for operation of the mass spectrometers, Mr. Mark Stefik and Mr. Robert Stillman for programming assistance and data acquisition, and Mr. Walter Reynolds for interfacing. We are grateful to Syntex Corporation for provision of many of the estrogen standards employed in this study.

[17] H. Adlercreutz and T. Luukkainen in "Gas Chromatography of Hormonal Steroids," R. Scholler and M. F. Jayle, Ed., Dunod, Paris, 1968, p 93.