

- (7) W. Fuller and M. J. Waring, *Ber. Bunsenges. Phys. Chem.*, **68**, 805 (1964).
- (8) D. M. Neville, Jr., and D. R. Davies, *J. Mol. Biol.*, **17**, 57 (1966).
- (9) J. Cairns, *Cold Spring Harbor Symp. Quant. Biol.*, **27**, 311 (1962).
- (10) (a) D. M. Crothers, *Biopolymers*, **6**, 575 (1968); (b) W. R. Bauer and J. Vinograd, *J. Mol. Biol.*, **47**, 419 (1970).
- (11) E. F. Gale, E. Cundliffe, P. E. Reynolds, M. H. Richmond, and M. J. Waring, "The Molecular Basis of Antibiotic Action", Wiley, London, 1972, pp 188-220.
- (12) G. T. Morgan and F. H. Burstall, *J. Chem. Soc.*, 1498 (1934).
- (13) G. M. Intille, Ph.D. Dissertation, Syracuse University, Syracuse, N.Y., 1967.
- (14) G. M. Intille, C. E. Pfluger, and W. A. Baker, Jr., *J. Cryst. Mol. Struct.*, **3**, 47 (1973).
- (15) F. Basolo, H. B. Gray, and R. G. Pearson, *J. Am. Chem. Soc.*, **82**, 4200 (1960).
- (16) C. H. Langford and H. B. Gray, "Ligand Substitution Processes", W.A. Benjamin, New York, N.Y., 1965, pp 18-54.
- (17) P. J. Bond, R. Langridge, K. W. Jennette, and S. J. Lippard, *Proc. Natl. Acad. Sci. U.S.A.*, **72**, 4825 (1975).
- (18) G. Schwarz, S. Klose, and W. Balthasar, *Eur. J. Biochem.*, **12**, 454 (1970).
- (19) G. P. Kreishman, S. I. Chan, and W. Bauer, *J. Mol. Biol.*, **61**, 45 (1971).
- (20) (a) G. Thomas and B. Roques, *FEBS Lett.*, **26**, 169 (1972); (b) D. M. Crothers, private communication.
- (21) N. S. Angerman, T. A. Victor, C. L. Bell, and S. S. Danyluk, *Biochemistry*, **11**, 2402 (1972).
- (22) (a) R. S. Osborn and D. Rogers, *J. Chem. Soc., Dalton Trans.*, 1002 (1974); (b) D. Rogers, private communication.
- (23) (a) W. Müller and D. M. Crothers, *Eur. J. Biochem.*, **54**, 267 (1975); (b) W. Müller, H. Büneemann, and N. Dattagupta, *ibid.*, **54**, 279 (1975); (c) E. J. Gabbay, R. E. Scofield, and C. S. Baxter, *J. Am. Chem. Soc.*, **95**, 7850 (1973), and references cited therein.
- (24) (a) J.-B. LePecq, M. LeBret, J. Barbet, and B. Roques, *Proc. Natl. Acad. Sci. U.S.A.*, **72**, 2915 (1975); (b) E. S. Canellakis, Y. H. Shaw, W. E. Hammers, and R. A. Schwartz, *Biochim. Biophys. Acta*, **418**, 277 (1976).
- (25) R. D. Feltham and R. G. Hayter, *J. Chem. Soc.*, 4587 (1964).
- (26) W. E. Parham, H. Wynberg, and F. L. Ramp, *J. Am. Chem. Soc.*, **75**, 2065 (1953).
- (27) Programs for IBM 360-91 computer used in this work: UMAT, the local version of the Brookhaven diffractometer setting and cell constant and orientation refinement program; ORABS, the local version of the absorption correction program by D. J. Wehe, W. R. Busing, and H. A. Levy, adapted to FACS-I geometry; XDATA, the Brookhaven Wilson plot and scaling program; FOURIER, the Zalkin Fourier program; CUOLS, the local version of the Busing-Martin-Levy structure factor calculation and least-squares refinement program (ORFLS) modified by Ibers and Doedens for rigid-body refinement; ORFFE, the Busing-Martin-Levy molecular geometry and error function program; TRACER II, the Lawton lattice transformation-cell reduction program; and ORTEP II, the Johnson thermal ellipsoid plotting program; in addition to various local data processing programs.
- (28) B. G. Segal and S. J. Lippard, *Inorg. Chem.*, **13**, 822 (1974).
- (29) J. T. Gill and S. J. Lippard, *Inorg. Chem.*, **14**, 751 (1975).
- (30) "International Tables for X-Ray Crystallography", Vol. III, Kynoch Press, Birmingham, England, 1962, pp 162-163.
- (31) A. J. C. Wilson, *Nature (London)*, **150**, 151 (1942).
- (32)  $R_1 = \sum |F_o| - |F_c| / \sum |F_o|$  and  $R_2 = [\sum w|F_o| - |F_c|]^2 / \sum w|F_o|^2$  <sup>1/2</sup>, where  $w = 4F_o^2 / \sigma^2(F_o^2)$ . The function  $\sum w|F_o| - |F_c|$  was minimized.
- (33) Atom scattering factors (and anomalous dispersion corrections for Pt and S atoms): "International Tables for X-Ray Crystallography", Vol. IV, Kynoch Press, Birmingham, England, 1974, Tables 2.2A (and 2.3.1).
- (34) D. W. J. Cruickshank in "Computing Methods of Crystallography", J. S. Rollett, Ed., Pergamon Press, New York, N.Y., 1965, pp 112-115.
- (35) (a) S. Castellano, H. Gunther, and S. Ebersole, *J. Phys. Chem.*, **69**, 4166 (1965); (b) F. E. Lytle, L. M. Petrosky, and L. R. Carlson, *Anal. Chim. Acta*, **57**, 239 (1971); (c) E. Bielli, P. M. Gidney, R. D. Gillard, and B. T. Heaton, *J. Chem. Soc., Dalton Trans.*, 2133 (1974).
- (36) D. A. Ucko and S. J. Lippard, unpublished results.
- (37) L. Pauling, "The Nature of the Chemical Bond", 3d ed, Cornell University Press, Ithaca, N.Y., 1960, p 260.
- (38) (a) H. Booth and A. H. Bostock, *Chem. Commun.*, 637 (1967); (b) R. A. Bauman, *J. Org. Chem.*, **32**, 4129 (1967); (c) *Tetrahedron Lett.*, 419 (1971); (d) R. Gallo, A. Liden, C. Roussel, J. Sandström, and J. Metzger, *Tetrahedron Lett.*, 1985 (1975).
- (39) L. E. Erickson, J. E. Sarneski, and C. N. Reilley, *Inorg. Chem.*, **14**, 3007 (1975).
- (40) W. J. Geary, *Coord. Chem. Rev.*, **7**, 81 (1971).
- (41) R. G. Hayter and F. S. Humiec, *Inorg. Chem.*, **2**, 306 (1963).
- (42) See paragraph at end of paper regarding supplementary material.
- (43) E. Goldschmid and N. C. Stephenson, *Acta Crystallogr., Sect. B*, **26**, 1867 (1970).
- (44) F. W. B. Einstein and B. R. Penfold, *Acta Crystallogr.*, **20**, 924 (1966).
- (45) M. R. Caira and L. R. Nassimbeni, *Acta Crystallogr., Sect. B*, **31**, 581 (1975).
- (46) R. Melanson, J. Hubert, and F. D. Rochon, *Can. J. Chem.*, **53**, 1139 (1975).
- (47) L. L. Merritt, Jr., and E. D. Schroeder, *Acta Crystallogr.*, **9**, 801 (1956).
- (48) (a) B. Pullman and A. Pullman, *Prog. Nucleic Acid Res. Mol. Biol.*, **9**, 327 (1969); (b) C. E. Bugg, J. M. Thomas, M. Sundaralingam, and S. T. Rao, *Biopolymers*, **10**, 175 (1971).
- (49) J. S. Miller and A. J. Epstein, *Prog. Inorg. Chem.*, **20**, 1 (1976).
- (50) E. Subramanian, J. Trotter, and C. E. Bugg, *J. Cryst. Mol. Struct.*, **1**, 3 (1971).
- (51) M. Hospital and B. Busetta, *C. R. Acad. Sci., Ser. C*, **268**, 1232 (1969).
- (52) C. Courseille, B. Busetta, and M. Hospital, *C. R. Acad. Sci., Ser. C*, **275**, 95 (1972).
- (53) S. K. Obendorf, H. L. Carrell, and J. P. Glusker, *Acta Crystallogr., Sect. B*, **30**, 1408 (1974).
- (54) S. Neidle and T. A. Jones, *Nature (London)*, **253**, 284 (1975).
- (55) C.-C. Tsai, S. C. Jain, and H. M. Sobell, *Proc. Natl. Acad. Sci. U.S.A.*, **72**, 628 (1975).

## Applications of Artificial Intelligence for Chemical Inference. 22. Automatic Rule Formation in Mass Spectrometry by Means of the Meta-DENDRAL Program<sup>1a</sup>

B. G. Buchanan,\* D. H. Smith, W. C. White, R. J. Gritter,<sup>1b</sup> E. A. Feigenbaum, J. Lederberg, and Carl Djerassi

Contribution from the Departments of Computer Science, Chemistry, and Genetics, Stanford University, Stanford, California 94305. Received January 27, 1976

**Abstract:** The DENDRAL computer program uses established rules of molecular fragmentation to help chemists solve complex structural problems from mass spectral data. This paper describes a computer program called Meta-DENDRAL, that can aid in the discovery of such rules from empirical data on known compounds. The program uses heuristic methods to search for common structural environments around those bonds that are found to fragment and abstracts plausible fragmentation rules. The program has been tested on the well-characterized, low-resolution mass spectra of aliphatic amines and the high-resolution mass spectra of estrogenic steroids. The program has also discovered new fragmentation rules for mono-, di-, and trike-toandrostanes.

The DENDRAL computer program is designed to aid chemists with complex structure elucidation problems. One main part uses established molecular fragmentation rules to help chemists interpret mass spectra;<sup>2</sup> another main part generates lists of isomers that satisfy constraints derived from

a variety of spectroscopic techniques.<sup>3</sup> Because the mass spectrometry rules used by the DENDRAL program have been culled from the literature, the program's growth depends upon manual examination of collections of spectra. But investigating the spectral data of new compound classes to determine frag-

mentation and rearrangement processes is a long and complicated task for chemists. Thus we became interested in the extent to which a computer program, subsequently called Meta-DENDRAL, could suggest rules that explain origins of peaks in mass spectrometric data.

The Meta-DENDRAL program interactively aids chemists in determining the dependence of mass spectrometric fragmentations on substructural features, under the hypothesis that molecular fragmentations are related to topological graph structural features of molecules. Our goal is to have the program suggest qualitative explanations of the characteristic fragmentations and rearrangements among a set of molecules. We do not now attempt to rationalize all peaks nor find quantitative assessments of the extent to which various processes contribute to peak intensities.

The program emulates many of the reasoning aspects of manual approaches to rule discovery. It reasons symbolically, using a modest amount of chemical knowledge. It decides which data points are important and looks for fragmentation processes that will explain them. Then, as a chemist does, the program tests and modifies the rules.<sup>4</sup>

This paper can be read as two distinct parts: Method and Results. The first discusses the organization of the computer program and might be skimmed by mass spectrometrists mainly interested in new results. The second part discusses some mass spectrometry results produced by the program. Correspondingly, the latter might be skimmed by persons mainly interested in uses of computers in chemistry. The conclusions section is intended as a conclusion for both parts.

#### Method: Meta-DENDRAL Program

The Meta-DENDRAL program is organized as three subprograms called INTSUM, RULEGEN, and RULEMOD, as shown in Figure 1.

**Explaining the Data in Terms of Bond Cleavages: The INTSUM Program.** The INTSUM program (named for data interpretation and summary) interprets spectral data of known compounds in terms of possible bond cleavages. Since this first step has already been described in detail elsewhere,<sup>5</sup> we will recapitulate very briefly those key features that are relevant to the present study.

For each molecule in a given set, INTSUM first produces the plausible bond-cleavage processes which might occur, i.e., breaks and combinations of breaks, with and without the transfer of hydrogens and other neutral species.<sup>6</sup> These processes are associated with specific bonds in a portion of the molecular structure, or skeleton, that is chosen because it is common to the molecules in a given set. Then INTSUM examines the spectra of the molecules looking for evidence (spectral peaks<sup>7</sup>) for each process. INTSUM, however, does not recognize that different cleavages (of the skeleton or substituents) may represent fragmentation processes which are similar in that the bonds cleaved have similar substructural environments. This is a particular limitation for classes of molecules, such as the aliphatic amines, where the common skeleton is a single atom, a nitrogen atom.

INTSUM gives explanations of spectral peaks for each molecule and then produces a summary showing the total evidence associated with each possible process. The summary reports the following: (1) a set of molecules—their names and structural descriptions; (2) a set of processes—their names, a description of the bonds to be broken, an indication of charge placement, and any neutral transfers occurring in conjunction with the fragmentation; (3) a set of supporting evidence associated with each process—the mass spectral peaks that provide evidence for the occurrence of the process for each molecule.

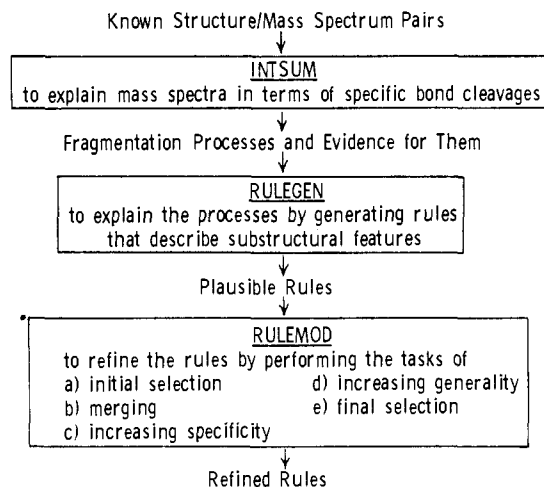


Figure 1. Organization of the Meta-DENDRAL program.

Meta-DENDRAL next attempts to correlate the fragmentations, as reported by INTSUM, with substructural features of molecules: in our terms we say that it classifies the structural environments ("bond environments") around the bonds that are cleaved.

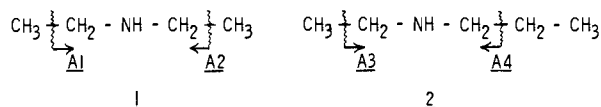
**Explaining Cleavages in Terms of Bond Environments: The RULEGEN Program.** Rules that explain the INTSUM processes in more general terms are produced in a predefined format called a "situation-action" form. Each rule is a conditional statement which is interpreted as saying that if any molecule satisfies the subgraph (situation), then the corresponding fragmentation process (action) will occur in the mass spectrometer.

The program creates rules by selecting "important" features of the molecular structure around the site of the fragmentations proposed by INTSUM. Essentially, it searches through a space of subgraph descriptions, looking for successive subgraphs that are supported by successively "better" sets of evidence.

**Rule Search.** The results of INTSUM are first rewritten by RULEGEN as a set of detailed bond-environment descriptions for the cleaved bonds together with associated spectral evidence. This is illustrated below with some secondary amines. Currently we describe bond environments in terms of a topological, or connectivity, model of structure. We specify atom type ("type"), degree of substitution ("nbrs"), numbers of hydrogens ("nhs"), or number of multiple bonds ("dots") at any atom place. Other features of atoms can also be used if they are computable from the connectivity graph model of a molecule, e.g., ring size or chain length. For example, the  $\alpha$ -cleavage processes (A1-A4) in the secondary amines **1** and **2** shown in Figure 2 may be considered to be four different processes by INTSUM (A1-A4). The program rewrites this information in terms of detailed bond environments, out to a predetermined distance from the cleavage site.<sup>8</sup> Then it collects A1-A3 and evidence for them into the same environment (B1, Figure 2), while cleavage A4 is described differently (B2, Figure 2). Figure 2 also shows B1 and B2 in terms of the features described for each atom out to the specified distance.

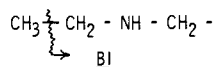
We use heuristic search<sup>4</sup> to examine possible generalizations of the environments, where each generalization can be interpreted as a possible rule when coupled with information about the bond cleavages and transfers of hydrogens or other neutral species. Conceptually, the program begins with the most general subgraph description R\*R (where R is any unspecified atom and the asterisk is used to indicate the bond cleaved, with the charged fragment written to the left of the asterisk). Then it generates refined descriptions by successively specifying one additional feature in all possible ways. The most useful rules lie somewhere between the overly general environment R\*R

## Fragmentations Observed:

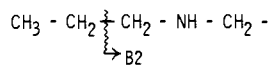


## Bond Environments:

A1, A2, A3 are rewritten as the bond environment:



A4 is rewritten as the bond environment:



## Features described in bond environments B1 and B2:

B1	type=C nhs=3 nbrs=1 dots=0	type=C nhs=2 nbrs=2 dots=0	type=N nhs=1 nbrs=2 dots=0	type=C nhs=2 nbrs=2 dots=0
	↑	↑	↑	↑

B2	type=C nhs=3 nbrs=1 dots=0	type=C nhs=2 nbrs=2 dots=0	type=C nhs=2 nbrs=2 dots=0	type=N nhs=1 nbrs=2 dots=0	type=C nhs=2 nbrs=2 dots=0
	↑	↑	↑	↑	↑

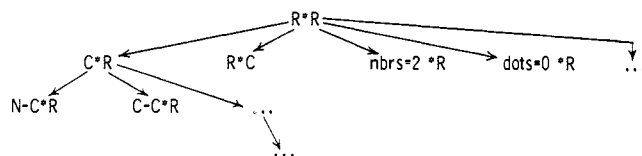
**Figure 2.** Bond environment descriptions for the  $\alpha$ -cleavages A1 and A2 in the mass spectral fragmentation of diethylamine (1) and A3 and A4 in the fragmentation of ethyl-*n*-propylamine (2).

and the overly specific complete bond-environment descriptions such as B1 and B2, Figure 2.

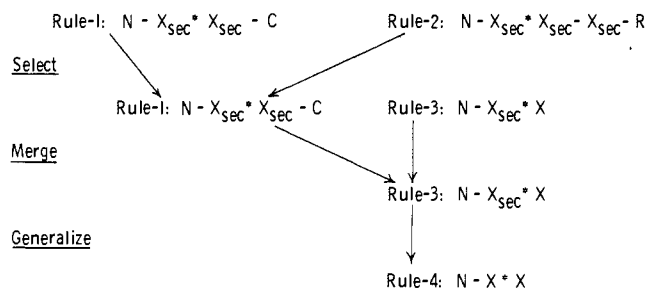
The program adds feature specifications to subgraphs one at a time, always making a "parent" subgraph more specific. From the parent  $R^*R$ , there are several ways to specify exactly one feature. In Figure 3, two "daughters" of  $R^*R$  are shown that specify the type ("type = carbon") of atom adjacent to the cleaved bond (other daughters will have specified the neighbors of a single atom, dots of a single atom, and so forth). Any single feature at any single atom place could have been specified in the first generation of daughters, e.g., by specifying a value of type, nbrs, nhs, or dots at the first atom place on either side of the bond. Each of the daughters of the parent  $R^*R$  is checked to see if the supporting evidence is "better" (see below) than the evidence for the parent. Those which satisfy the test become new parents for a next level of daughters with one more feature specified.<sup>9</sup> In our example, the bond environment descriptions B1 and B2 (Figure 2) are both included, with their associated evidence, under the rules in the path from  $R^*R$  to  $N-C^*R$  (Figure 3). Further specifications of  $N-C^*R$  show no improvement so the program stops here on this path (but continues to explore other paths). Thus  $N-C^*R$  becomes a rule candidate together with associated transfers of hydrogen atoms or other neutral species.

The program continues to make rules more specific until it finds a daughter rule that is (a) specific enough to focus on one type of process, such as  $\alpha$  cleavage (and to avoid many counterexamples) and (b) general enough to account for more than a few special cases.<sup>10</sup> The information for deciding whether a daughter rule is better than its parent is obtained from the record of detailed bond environment descriptions and their associated evidence. We avoid using ion intensity information at this stage in order to ensure finding rules that are applicable to most molecules irrespective of intensities. The next stage of the program (see RULEMOD below) considers intensities in order to focus on rules that explain the most intense peaks.

**Refining the Explanations: The RULEMOD Program.** The last phase of the program (RULEMOD) evaluates the plausible rules generated by RULEGEN and modifies them by making



**Figure 3.** Example of successive specification of subgraph features by RULEGEN in its search for plausible, general rules.



**Figure 4.** Example of successive refinement of plausible rules by RULEMOD. X represents any atom type other than hydrogen. The subscript "sec" means the atom is secondary, which is equivalent to "nbrs = 2".

them more general or more specific. Its task is to analyze the validity of predictions<sup>11</sup> made by the rules on the original set of molecules, modify the subgraph descriptions of the rules to improve the accuracy of their predictions, merge similar rules, and finally select a subset of the modified rules. RULEMOD will typically turn out a set of 8–12 rules covering substantially the same data points as an original set of approximately 25–100 rules, but with fewer incorrect predictions. RULEMOD is written as a set of five tasks (corresponding to the five subsections below) which we feel are closely analogous to this aspect of human problem solving.

**Selecting a Subset of Important Rules.** As a first step, the selection procedure is applied to the whole set of rule candidates produced by RULEGEN. The local evaluation in RULEGEN has not discovered that different RULEGEN pathways (Figure 3) may yield rules which are different but explain many of the same data points. Thus there is often a high degree of overlap in those rules, and rules may have many counterexamples. For example, rule 1 (Figure 4) is selected by the procedure outlined below because it explains all the same  $\alpha$ -cleavage peaks in the amine spectra as rule 2 (Figure 4) and some others in addition. Rule 1 applies to propyl and higher alkyl amines, while rule 2 does not accommodate propyl amines. When RULEMOD computes scores by eq 1 below, it discovers that rule 2 is totally redundant and eliminates it because it has no independent evidence of its own.

To select rules, scores are calculated by eq 1, the rule with the best score is selected, and the evidence peaks supporting that rule are removed from the supporting evidence for other rules. Then the whole process is repeated until either (i) all scores are below a selected threshold or (ii) all evidence has been explained. Equation 1 captures the following intuitions.

(a) The score should reflect the strength of evidence, i.e., it should be proportional to average peak intensity.

(b) Data points (peaks) that are uniquely explained by a rule should count more than peaks that can be explained by two or more rules.

(c) Negative evidence (peaks predicted and not found) should count heavily against a rule.

(d) Since the number of molecules in the set remains the same during rule formation and we insure that every rule applies to a minimum number of molecules (in our case half the molecules), the score for purposes of selection need not

explicitly factor in the sample size. When we want to compare sets of rules formed from different sets of molecules, however, it will be necessary to weight the scores of rules by the number of molecules considered.

$$\text{Score} = I(P + U - 2N) \quad (1)$$

$I$  (average intensity expressed as %  $\Sigma$  total ion current) = the sum of intensities of the peaks counted as positive evidence divided by the positive evidence count.  $P$  (positive evidence count) = the number of times there is any evidence (spectral peaks) for a rule. The positive evidence count is incremented by one if any of the predicted ions (including predicted transfers of neutral species) is found.  $U$  (unique evidence count) = the number of times there are peaks predicted *only* by the given rule. The unique evidence count is incremented if any or all transfers associated with a rule produce uniquely explained peaks in the spectrum.  $N$  (negative evidence count) = the number of times a rule applies to a molecule but no predicted peaks are observed in the spectrum.

**Merging Rules.** Although most of the redundant rules have been deleted in the first step of RULEMOD, there may still remain sets of rules that explain many of the same data points. For any such set of rules, the program attempts to find a slightly more general description that (a) includes all the evidence covered by the overlapping rules and (b) does not bring in extra negative evidence. If it can find such a description, the overlapping rules are merged into the more general expression. For example, two different forms of  $\alpha$ -cleavage rules remain after rule selection, rule 1 and rule 3, Figure 4.<sup>12</sup> These two rules are now merged into a form that is equivalent to rule 3 because conditions a and b above are satisfied. In general, however, the form of the merged rule will be different from any of the component rules.

**Deleting Negative Evidence by Making Rules More Specific.** RULEMOD tries to add additional specifications to each rule in order to delete some counterexamples, or negative evidence, while keeping all of the positive evidence. In our  $\alpha$ -cleavage example, the merged rule (rule 3, Figure 4) could not be made more specific in a way that would get rid of the two pieces of negative evidence and still keep all of the positive evidence (two spectra of ethyl-substituted amines did not show loss of methyl above the 1%  $\Sigma$  threshold used to process the data).

**Making Rules More General.** RULEGEN often forms rules that are more specific than they need to be. At this point we have a choice whether to leave the rules as they are or to seek a more general form that covers the same (and perhaps new) data points without introducing new counterexamples. Rule 3 (Figure 4), for example, may be made more general by removing the specification "NBRS = 2" (i.e., "secondary") from the atom adjacent to the nitrogen.<sup>13</sup> This yields the final form for the  $\alpha$ -cleavage rule, rule 4 (Figure 4).

**Selecting the Final Rule Set.** The selection procedure described above is applied again at the very end of RULEMOD in order to weed out redundancies that might have been introduced and to select the "best" of the rules. For example, for the aliphatic amines the program selected just five rules for the final set (see Results section below). These were derived from 27 candidates produced by RULEGEN which were reduced to ten by RULEMOD's initial selection.

**Evaluating the Rules.** One way of evaluating rules is measuring how well they explain, or "cover", the given spectra (which we call the "explanatory power" of rules). We also want to be able to estimate how well they can be used for selecting the most plausible structures from a list of candidate explanations of an unknown spectrum (from a known class), which we call "discriminatory power". Two other related measures that we present in the Results section below are positive evidence count, which is the number of ions predicted by rules and

found in the actual spectrum, and negative evidence count, which is the number of ions predicted and not found.

**Explanatory Power.** An objective measure of explanatory power is the percent total ionization explained by the rules, i.e., the sum of intensities of spectral peaks predicted and found. This number typically is between 25 and 50%  $\Sigma$ , but is not necessarily significant because the rules have not been chosen on the basis of their ability to rationalize the whole spectrum. Another measure is just the number of peaks predicted and found. The rules typically predict six to eight ions per molecule, some with satellite ions due to transfer of neutral species. A less objective facet of explanatory power is the significance of the ions that are explained, measured in terms other than peak intensities, in particular, mass and heteroatom content. For this study we have associated greater significance to ions of higher mass. If peaks judged to be significant are left unexplained, either the rules are incomplete or the compound exhibits fragmentation processes that are not general to the class (and thus not covered by the general rules).

**Discriminatory Power.** Another measure of a set of rules is their ability to differentiate among hypotheses. Our measure of discriminatory power determines how well a set of rules explains a structure's actual mass spectrum as compared with how well the rules applied to all other candidates explain the given spectrum. This is particularly powerful when coupled with a generator of all possible molecules in a class,<sup>3</sup> for then we can ask how well a set of rules allows us to discriminate the correct structure from among all possible members of the class.

Since we have directed RULEGEN toward finding general rules, there is no guarantee that the rules will discriminate well among alternative structures. The discriminatory power of the rules will be proportional to the number of unique ions predicted. That is, if a set of rules predicts ions for a given compound which are not predicted by their application to another compound then it is easy to discriminate between the two.

The score for each comparison was computed by eq 2. The scoring function used in discrimination penalizes a candidate structure if its predicted spectrum shows significant peaks that are not in the actual data. The other kind of mismatch, failure to predict peaks that appear in the actual spectrum, does not penalize a candidate. The reason for this asymmetry is that the rules are selected for their generality and thus should make correct predictions, but they are not expected to explain all ions in a spectrum. When comparing a predicted mass spectrum from a candidate structure against an actual spectrum, the peaks that arise from nongeneral fragmentation processes will not be predicted.

$$\text{Comparison Score} = \sum_R \text{sig} [\text{numpkts found or} \\ -1 \text{ if numpkts found is zero}] \quad (2)$$

$R$  = rules; sig = significance of a predicted ion = mass of ion/ $M^+$ ; numpkts = number of peaks predicted by rule  $R$ .

Intuitively we are computing a weighted sum of positive evidence (number of peaks predicted and found in the spectrum) minus negative evidence ( $-1$  if none of the predicted peaks is found). For example, assume a rule predicts ions of composition  $C_{11}H_{15}O$  ( $m/e$  163, no hydrogen transfer) and  $C_{11}H_{14}O$  ( $m/e$  162, loss of a hydrogen atom) for a compound of molecular weight 300. The occurrence of both ions in the spectrum would contribute  $163/300 \times 2$  to the score ( $= 1.09$ ) computed by eq 2 for this rule.

**Limitations of the Method.** The major limitation of the heuristic search method in any domain is the necessity of finding (or inventing) a generator of possible solutions. In the rule formation domain that means that we have to invent a program that generates possible rules. That, in turn, requires a strict definition of the allowable forms of the rules and a

definition of the allowable primitive terms that add content to the form. The representation we have found for expressing rules is fixed for any one run, but can, at least, be modified or extended manually between runs.

A second major limitation on heuristic search is the necessity of finding heuristics, rules of thumb, that guide the generator. For rule generation it is necessary to find heuristics that steer the generator toward the small number of interesting rules and away from the very large number of uninteresting rules. The problem is that it is difficult to find these guiding principles. In addition, putting confidence in the heuristics requires an act of faith. Once that step is made, however, there is often the temptation to put *too much* faith in the heuristics and forget that the solutions (rules) were found in the context of a large number of assumptions. For example, one might tend to forget the criteria for data filtering, or the range of allowable hydrogen transfers, or the restrictions on how complex the rules were allowed to become, or the criteria for filtering the rules. All of the heuristics together define the range of rules considered and thus should temper our judgments about the generality of the rules.

Another limitation on the use of heuristic search is that the computer programs are often slow, not because they are inefficient as much as they must do a lot of computation. The Meta-DENDRAL program is also inefficient now because it is still in the development stage.

There are also limitations imposed by the domain of chemistry, which have been mentioned elsewhere. To reiterate, the program now works only on sets of molecules that share a common substructural skeleton represented topologically. The program depends on a chemist's judgment for the program's chemical heuristics. (This is also a strength as well as a limitation.) Finally, the program's rules say almost nothing about spectral peak intensities because they are meant to be more qualitative than quantitative. That is, they predict fragmentations and assign the resulting peaks an average expected intensity, but they do not predict intensities accurately.

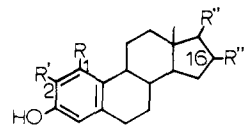
## Results

**Known Test Cases.** We have tested these programs on two widely different classes of compounds (aliphatic amines and estrogenic steroids) for which fragmentation processes had been well characterized in the literature. We selected these tests on the basis of available mass spectra and characteristics of the molecules for testing different parts of the program.

**Aliphatic Amines.** Five rules were produced by the program to explain the low-resolution spectra of 11 aliphatic amines ranging in size from  $C_4$  to  $C_{14}$ : (1)  $\alpha$  cleavage was described; (2) the two-step process of  $\alpha$  cleavage with concomitant  $\beta$  hydrogen transfer and C-N cleavage was included; (3,4) two rules describing  $\gamma$  cleavage with concomitant C-N cleavage were produced; and (5) a rule was produced describing cleavage of any two bonds that will produce a nitrogen-containing fragment. These rules correlate well with reported fragmentations of amines.<sup>14</sup> These rules explain 84%  $\Sigma$  of the total ion current in these eleven spectra, with the molecular ion contributing another 3.5%  $\Sigma$  on the average. These rules (arrived at independently) and their discriminatory power have been thoroughly described by others.<sup>14</sup>

**Estrogenic Steroids.** The program produced eight rules to explain the high-resolution spectra of ten estrogenic steroids. These included all five of the rules discussed previously.<sup>2</sup> The additional three rules describe cleavages through rings B and C, which are plausible explanations of peaks observed in the spectra. These eight rules account for over 40%  $\Sigma$  of the total ion current in these spectra, with the molecular ion accounting for an additional 23%  $\Sigma$ , on the average.

We know that these rules have considerable discriminatory power because of previous work, where they were used for structure elucidation.<sup>2</sup> We verified this in a few cases by showing that the program can distinguish 2-hydroxyestradiol (3) and estriol (4) from each other and from all other possible estriols (with the exception that 1-hydroxy- (5) and 2-hydroxyestradiol (3) are indistinguishable). Also the program discriminates 2-hydroxy- and 16-hydroxyestrone (6, 7) from each other and from all other possible hydroxyestrone (with the exception that 1-hydroxy- (8) and 2-hydroxyestrone (6) are indistinguishable).



- 3,  $R' = R'' = OH$ ;  $R = R''' = H$   
 4,  $R'' = R''' = OH$ ;  $R = R' = H$   
 5,  $R = R'' = OH$ ;  $R' = R''' = H$   
 6,  $R' = OH$ ;  $R'' = \text{keto}$ ;  $R = R''' = H$   
 7,  $R''' = OH$ ;  $R'' = \text{keto}$ ;  $R = R' = H$   
 8,  $R = OH$ ;  $R'' = \text{keto}$ ;  $R' = R''' = H$

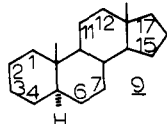
**Ketoandrostanes.** The high-resolution mass spectra of several mono-, di-, and triketoandrostanes provide an interesting case study in the use of Meta-DENDRAL for analyzing the dependence of fragmentation processes on the structural features of molecules. The monoketoandrostanes have been extensively utilized as model systems for the study of mass spectral fragmentation processes in this and related classes of steroids.<sup>15,16</sup> Low-resolution spectra, extensive deuterium labeling, and some high-resolution measurements have clarified the origins of several important ions in these compounds.<sup>15,16</sup> Brief comparisons among the spectra of the monoketones reveal a remarkable lack of similarity in the fragmentations. Indeed, these compounds were studied individually in our laboratory<sup>15,17</sup> to answer other questions and little attempt has been made to identify common fragmentation modes (if any). The spectra depend on the placement of the keto group, but with little obvious consistency. For example, in some cases,  $\alpha$  cleavage adjacent to the carbonyl group is important; in others,  $\beta$  cleavage; in still others both, or neither.

More recently, transformation of steroidal alcohols to their respective keto analogues has been recommended as a general method for localization of hydroxyl functionalities.<sup>18</sup> This technique has been utilized in a study of some specific structures.<sup>19</sup> Again, published results tend to focus primarily on individual compounds, presumably because there are few obvious correlations among several members of the class. Although such studies have unquestionable value, the sobering fact is that there are 55 possible diketoandrostanes and 165 possible triketoandrostanes (structures listed by CONGEN<sup>3</sup>) excluding functionalization of the C-18 and C-19 methyl groups. This suggests that a more detailed investigation which might uncover systematic behavior would be useful because it might be possible to generalize such findings to a study of unknown compounds, for example, those isolated in conjunction with microbial transformations.<sup>20</sup>

We examined the spectra of the ketoandrostanes with two important questions in mind. First, are there consistent fragmentation processes which depend on local substructures within the molecules? It is not obvious that there are processes which are general to the androstane skeleton (9) itself<sup>21</sup> (see INTSUM results, below). We did not expect this approach to rationalize the sometimes intense ions which appear to be due to processes occurring in single (or small numbers of) structures; the INTSUM output alone provides explanations for such processes. But there are many other ions in the spectra which might have considerable structural significance. A second

question concerns our continuing interest in interactions among functional groups,<sup>22</sup> in this case interactions of keto groups within the rigid framework of the steroid system. Thus we examined the extent to which the fragmentation modes of the di- and triketones reflect the fragmentation modes of the respective monoketones. The following sections summarize our findings.

**Monoketoandrostanes.** We had complete high-resolution spectra of ten of the 11 possible monoketo-5 $\alpha$ -androstanes and two 5 $\beta$  isomers (see Table I). Androstane (**9**) is included as a



point of reference for the skeletal fragmentation processes.

The INTSUM program was used to interpret and summarize the spectral data of **9–21**.<sup>23</sup> The fragmentation processes proposed by INTSUM are sufficient to explain on the average of 57%  $\Sigma$  of the total ion current for **9–21** (range of 45–77%  $\Sigma$ ). In nearly every case the residual ion current is due to hydrocarbon ions, predominantly at low mass ( $m/e$  40–100). Manual examination of these hydrocarbon ions (including those at higher masses) indicated that their degree of unsaturation is higher than that expected from simple cleavages. These ions arise from loss of the oxygen substituent (as CO or H<sub>2</sub>O or as part of larger fragments) followed by complex fragmentations, yielding ions of little diagnostic significance (e.g., C<sub>7</sub>H<sub>9</sub>, C<sub>7</sub>H<sub>7</sub>, and C<sub>6</sub>H<sub>7</sub>). There are, of course, occasional exceptions. For example, the intense ion of mass 98 (C<sub>6</sub>H<sub>10</sub>O) in the spectrum of androstan-4-one (**12**) has a more complex origin<sup>24</sup> and thus is not explained by INTSUM under the given constraints.<sup>23</sup>

Careful review of the spectrum-by-spectrum output from INTSUM has revealed general consistency for those processes and ions discussed in earlier studies.<sup>15,16,25</sup> The summary results of INTSUM provide quantitative measures of our intuitions that **9–21** do not behave very homogeneously with respect to skeletal cleavages. There are several important processes which occur in only a few molecules. For example, cleavage of the C-5,6 and C-9,10 bonds yields an abundant ion ( $m/e$  178) in the spectrum of the 7-ketone **14** (30%  $\Sigma$ ). Four other molecules display ions which can be ascribed to this process; androstane (**9**) itself, the 4-ketone **12**, and the isomeric 11-keto compounds **15** and **21**. No evidence for this cleavage above the 0.5%  $\Sigma$  threshold is observed for any of the other compounds. Similarly, ring D loss, with or without loss of additional hydrogen atoms, is observed in only six of the 13 compounds.

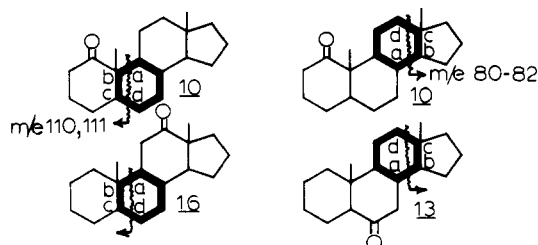
We summarize in Table II the fragmentation rules which emerge from RULEGEN and RULEMOD analysis from such seemingly inconsistent data. We present the rules exactly as they emerged (changing only the format for legibility, but not the content). Some of the most important interpretations implicit in the rules are mentioned in Table II.

The subgraph for a rule (second column, Table II) may "fit" (or match) several places, or none at all, in a given molecule. The indicated bond cleavage and resulting ion is predicted<sup>11</sup> to occur in every place where the subgraph fits. Consider for example rule M-2. In most monoketoandrostanes the subgraph fits twice, as indicated in bold face for androstan-1-one (**10**). In the case of **10**, ions of composition C<sub>7</sub>H<sub>11</sub>O and C<sub>7</sub>H<sub>10</sub>O are predicted and observed ( $m/e$  110 and 111) as are ions of composition C<sub>6</sub>H<sub>8</sub>–C<sub>6</sub>H<sub>10</sub> ( $m/e$  80–82). However, rule M-2 applies only once for androstan-6-one (**13**) because the rule requires that atom d (corresponding to C-6) be secondary, preventing a second match as in the case of androstan-1-one (**10**). Similarly, rule M-2 applies only once to androstan-12-one

Table I. Monoketoandrostanes Analyzed by Meta-DENDRAL

5 $\alpha$ -Androstane ( <b>9</b> )	5 $\alpha$ -Androstan-12-one ( <b>16</b> )
5 $\alpha$ -Androstan-1-one ( <b>10</b> )	5 $\alpha$ -Androstan-15-one ( <b>17</b> )
5 $\alpha$ -Androstan-3-one ( <b>11</b> )	5 $\alpha$ -Androstan-16-one ( <b>18</b> )
5 $\alpha$ -Androstan-4-one ( <b>12</b> )	5 $\alpha$ -Androstan-17-one ( <b>19</b> )
5 $\alpha$ -Androstan-6-one ( <b>13</b> )	5 $\beta$ -Androstan-3-one ( <b>20</b> )
5 $\alpha$ -Androstan-7-one ( <b>14</b> )	5 $\beta$ -Androstan-11-one ( <b>21</b> )
5 $\alpha$ -Androstan-11-one ( <b>15</b> )	

(**16**), as shown. In some cases a given rule may not apply to certain molecules at all. For example, rule M-5 applies only to those molecules with an unsubstituted ring A, because of the requirement that atoms a–d be secondary (Table II).



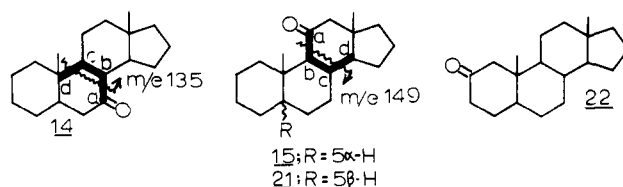
Salient features of rules M1–M8 are as follows.

(a) Most cleavages are adjacent to highly substituted centers.

(b) Only rules M-3 (loss of methyl), M-5 (ring A cleavage with charge retention on C-1,2,3,4), and M-8 (ring D cleavage with charge retention on C-15,16,17) can be considered specific to the androstane skeleton (**9**). The latter two rules are of diminished utility because they (i) predict only low-mass hydrocarbon ions which may arise from other sources; and (ii) have no prediction in the cases of ring A or D oxo substituents, respectively, because there is no evidence in such instances.









(c) The other rules speak of a higher level of generality in the description of fragmentation processes—they are not tied rigidly to cleavage of specific skeletal bonds, but to cleavage of certain bonds when a specific substructure is encountered. These rules usually apply at least twice per molecule.

(d) The near absence of rules that speak of the carbonyl group is notable. This reinforces earlier conclusions<sup>15</sup> concerning the lack of directing influence of the carbonyl group on the fragmentation of these molecules. Certain rules implicitly speak of the carbonyl group, in a negative sense, by having substructural descriptions which prevent the rule from applying where there is a keto group (e.g., rule M-2, see above discussion). Only rule M-6 mentions the keto group explicitly. This rule predicts cleavage  $\alpha$  to the carbonyl group together with cleavage of a  $\gamma$  bond with charge retention on the hydrocarbon moiety. Depending on the location of the carbonyl group, different ions are predicted. For example, rule M-6 predicts the abundant mass 135 ion (C<sub>10</sub>H<sub>15</sub>) in the spectrum of androstan-7-one (**14**), while it predicts an observed ion of mass 149 (C<sub>11</sub>H<sub>17</sub>) in the spectra of the isomeric androstan-11-ones (**15** and **21**).



(e) Rules M-1–M-8 have high explanatory power for compounds **9–21** and a "new" compound (which was not included in the set for rule formation), androstan-2-one (**22**). For compounds **9–21** they explain 42.4%  $\Sigma$  (including the molecular ion) of all spectral data (or 74%  $\Sigma$  of the data explained by INTSUM). Several abundant ions were explained by INT-

Table II. Monoketoandrostane Rules

Name <sup>a</sup>	Subgraph <sup>b</sup>	Other descriptors and interpretations	Score <sup>c</sup>	Positive evidence <sup>d</sup>		Negative evidence	Average intensity % $\sum_{40}$
				Any	Unique		
M-1 (none, +H, -H)		Atom a is not keto substituted	145.8	24	11	1	4.42
M-2 (none, +H, -H, -2 H)			135.3	21	12	3	5.01
M-3 (none)		Loss of methyl	84.5	26	26	0	1.62
M-4 (none, -H, -2 H)			70.5	21	0	1	3.71
M-5 (+H, -H)			39.5	8	4	0	3.29
M-6 (-H)		There must be a keto group on atom a	39.7	9	0	0	4.41
M-7 (-H, -2 H)		Atoms a and d are not keto substituted	23.8	13	13	7	1.99
M-8 (+H, -H)		Atom d is not keto substituted	13.1	3	0	0	4.35

<sup>a</sup> Important transfers of neutral species are indicated below the name of the rule. Absence of transfers is stated explicitly as "none". <sup>b</sup> Specified non-hydrogen substituents are indicated by R (R ≠ H). Other valence positions may be filled with any atoms (including H) except when restricted by other descriptors. <sup>c</sup> Score calculated by eq 1 above. <sup>d</sup> Positive evidence count (any and unique) and negative evidence count are described in the text. The number of positive instances may be greater than the number of molecules because a rule may apply more than once in any molecule. See text for explanation.

Table III. Diketoandrostanes Analyzed by Meta-DENDRAL

5 $\alpha$ -Androstane-2,11-dione (23)	5 $\alpha$ -Androstane-3,12-dione (28)
5 $\alpha$ -Androstane-3,17-dione (24)	5 $\alpha$ -Androstane-6,12-dione (29)
5 $\alpha$ -Androstane-3,6-dione (25)	5 $\alpha$ -Androstane-7,17-dione (30)
5 $\alpha$ -Androstane-3,7-dione (26)	5 $\alpha$ -Androstane-12,15-dione (31)
5 $\alpha$ -Androstane-3,11-dione (27)	

SUM, but not by the final rules, e.g.,  $m/e$  178 in the spectrum of androstan-7-one (**14**).<sup>15</sup> However, processes which yielded these ions are not part of consistent fragmentation behavior based on substructural features. The numbers of ions predicted and found (positive evidence) and ions predicted and not found (negative evidence) are shown in Table II. In every case, the counterexamples are molecules where the keto substituent is remote from the predicted cleavage sites. Because we did not explore bond environments far enough to consider such remote features, and because our assumption is that a rule will apply *wherever* it fits, we encounter some negative evidence.

Rules M-1–M-8 were used to predict fragmentations of a "new" compound (not included in the set for rule formation), androstan-2-one (**22**). Predicted peaks (i.e., peaks that are deductive consequences of the rules) are  $m/e$  41, 43, 55–57, 80–83, 94–96, 108–111, 122–124, 134, 135, 144–151, 162, 163, and 259. These rules do not predict the unique fragmentations of this compound giving rise to ions of mass 231 C<sub>3</sub>H<sub>6</sub>O (M<sup>+</sup> – 43, C<sub>3</sub>H<sub>7</sub>) and 216 (M<sup>+</sup> – 58, acetone). The predicted spectrum for this compound does not explain as many significant ions in the corresponding actual spectrum<sup>26b</sup> as in the case of the di- and triketoandrostanes discussed below.

(f) Rules M-1–M-8 have relatively low discriminatory power. These rules can distinguish the 7- and 11-keto com-

pounds from among all other possible monoketones, but they ranked the other structures anywhere from second to last when comparing the spectrum of the correct structure with predicted spectra for all candidates. This low discriminatory power relative to the estrogens above is due to a combination of three factors: (i) the rules seldom mention the carbonyl group explicitly (by itself this is not necessarily bad); (ii) each rule often predicts the same ions in all the molecules; and (iii) where different ions are predicted they are not unique to the spectrum of the correct compound.

In summary, rules M-1–M-8 indicate consistent fragmentation behavior in a set of molecules whose fragmentations previously seemed unrelated.

**Diketoandrostanes.** The experimental material available to us consisted of the complete high-resolution mass spectra of nine diketoandrostanes (**23–31**) listed in Table III. The low-resolution spectra of a few diketoandrostanes have been discussed previously, including the 1,6-,<sup>27</sup> 3,17- (**24**), and 1,17-<sup>18</sup> diones. The output of INTSUM agrees with previous interpretations<sup>18</sup> of the spectrum of **24**. A review of the INTSUM output for the mono- and diketoandrostanes reveals that (a) the spectra of the diketones do not represent superpositions of the respective monoketones—where any fragmentations characteristic of the monoketones are noted, one of the keto groups usually dominates the pattern; and (b) like the monoketoandrostanes, the diketoandrostanes display fragmentations characteristic of individual molecules with little apparent consistency with respect to skeletal cleavages. Many skeletal processes yield ions in only three to five spectra of the nine diketoandrostanes. The rules which result from analysis of the spectra of **23–31** in terms of substructural features are summarized in Table IV.

Table IV. Diketoandrostande Rules<sup>a</sup>

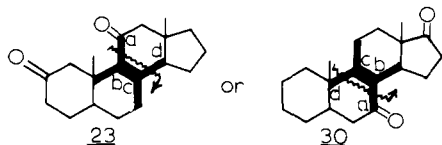
Name	Subgraph	Other descriptors, and interpretation	Score	Positive evidence		Negative evidence	Average intensity, % $\sum_{46}^{\Sigma}$
				Any	Unique		
D-1 (none, +H, -H)		Atom b is not keto substituted	65.2	8	2	1	8.16
D-2 (-H)			36.9	12	10	0	1.68
D-3 (none)		Loss of methyl	30.7	16	16	2	1.10
D-4 (-H, -2 H)		Atoms a, b, and c are not keto substituted	30.6	5	5	2	5.10
D-5 (-H)			29.2	4	4	0	3.65
D-6 (none, -H, -2 H)			28.5	7	3	0	2.85
D-7 (-H)			25.5	5	1	0	4.25
D-8 (none, -H)			25.1	8	8	1	1.79

<sup>a</sup> See footnotes to Table II for an explanation of terms.

The salient features of rules D-1–D-8 are as follows.

(a) All rules except D-3 (loss of methyl) and D-1 (vide infra) refer implicitly to the carbonyl group in a negative sense. Specifications on the substructures (Table IV) require certain atoms to be secondary. Such atoms cannot be carbon atoms of a carbonyl group. In other words, important fragmentations are observed only when the carbonyl group is remote from the cleavage site. As noted above for the monoketoandrostandes, the direct influence of the carbonyl group in directing the fragmentation appears to be minimal.

(b) Only rule D-1 possesses features which refer to the necessity for a carbonyl group. Rule D-1 applies as shown to androstane-2,11-dione (**23**) and androstane-7,17-dione (**30**), for example, and to other diketoandrostandes where atom a of rule D-1 is the carbonyl carbon and d is at least a tertiary center (or vice versa).



(c) There are several similarities between the rules for the mono- and diketoandrostandes. Rule D-1 is another way of expressing rule M-6. Rules D-3 and M-3 are the same. Rule D-2 is a slightly less general form of rule M-1, and rules D-7 and M-8 are the same.

(d) Rules D-1–D-8 do not explain as much of the total ion current (33%  $\Sigma$ ) for the diketoandrostandes as compared to the explanatory power of the monoketoandrostande rules, although this still constitutes 77%  $\Sigma$  of the data explained by INTSUM. The number of ions predicted is smaller, although the number of counterexamples is still small (compare scores and positive and negative evidence of Tables II and IV). This indicates to us that the group of diketoandrostandes is behaving less homogeneously with respect to dependence of fragmentation on

substructures. However, for new compounds, the rules predict important features of the spectrum of androstane-1,17-dione.<sup>18</sup> All ions predicted by rules D-1–D-8 are observed in the spectrum of this compound, including the significant ions at  $m/e$  273, 232, 217, and 124. Also, the program deduces the following ions for androstane-1,17-dione:  $m/e$  41, 109, 110, 122–124, 217, 218, 231, 232, and 273, and for androstane-1,6-dione<sup>27</sup>:  $m/e$  41, 81, 108–111, 217, 218, 258, 259, and 273.

(e) Interestingly, rules D-1–D-8 have much better discriminatory power than rules M-1–M-8 had for the monoketoandrostandes. We compared the predicted spectra for the 55 possible<sup>3</sup> diketoandrostandes (excluding functionalization of the C-18 and C-19 methyl groups) against the actual spectrum for each of **23–31**. Each of compounds **27**, **28**, and **30** were ranked first (i.e., discriminated correctly). Compounds **23–26**, **29**, and **31** were ranked 6th, 3rd, 4th, 16th, 8th, and 9th respectively, out of 55 candidates. The improvement in discriminatory power over the monoketoandrostandes reflects more frequent (implicit) reference to keto groups in the rules and a larger number of unique ions predicted by the rules.

**Triketoandrostandes.** We have recorded the high-resolution mass spectra of the eight triketoandrostandes (**32–39**) summarized in Table V. The fragmentations of **36** and its  $5\beta$  epimer have been discussed previously,<sup>27</sup> as have the fragmentations of **38** and its  $5\beta$  epimer, and **39** and its  $5\beta$  epimer, and  $5\alpha$ -androstane-3,7,17-trione.<sup>18</sup> These studies<sup>27,18</sup> have indicated significant differences in the relative abundances of several ions resulting from fragmentations of triketoandrostandes epimeric at C-5. However, the same fragmentation processes appear to be occurring in both epimers. In general, the output of INTSUM supports previous interpretations<sup>18</sup> of the spectra of **36**, **38**, and **39**. There are, however, alternatives to fragmentations proposed<sup>18</sup> for some ions in these spectra. For example, the abundant ion of mass 191 ( $C_{12}H_{15}O_2$ ) in androstane-3,11,17-trione (**38**) may represent loss of ring A



Table V. Triketoandrostanes Analyzed by Meta-DENDRAL

5 $\alpha$ -Androstane-2,11,17-trione (32)	5 $\alpha$ -Androstane-1,6,17-trione (36)
5 $\alpha$ -Androstane-6,11,16-trione (33)	5 $\alpha$ -Androstane-2,7,17-trione (37)
5 $\alpha$ -Androstane-6,11,17-trione (34)	5 $\alpha$ -Androstane-3,11,17-trione (38)
5 $\alpha$ -Androstane-3,6,16-trione (35)	5 $\alpha$ -Androstane-3,6,17-trione (39)

and C-5,10,19<sup>18</sup> or loss of ring D and C-12,13,14,18 together with an additional hydrogen.

In the INTSUM analysis of this set of molecules we considered H<sub>2</sub>O and CO (up to two of each) as neutral species which could be lost because there are important ions which appear to arise from cleavages together with such losses. The INTSUM results show some consistencies with respect to skeletal fragmentations, even though such fragmentations may be accompanied by losses of one or more H<sub>2</sub>O or CO molecules. The results of RULEGEN and RULEMOD analysis of the INTSUM output are summarized in Table VI.

Triketoandrostane rules T-1-T-10 have the following characteristics.

(a) Rules T-4, T-6, T-9, and T-10 explicitly refer to the requirement for a carbonyl group in the substructure. Atom a or c of rule T-1 can match the carbon atom of a carbonyl group. In most cases predicted cleavages are observed in addition to the same cleavages accompanied by the losses of CO and H<sub>2</sub>O summarized in Table VI.

(b) Considering functional-group interaction, a similar situation to that noted for the diketoandrostanes pertains to the triketones. The spectra are not superpositions of the respective mono- or diketones. What is notable is the almost total lack of influence of the C-17 and C-16 keto groups on the fragmentations of **32-36**. Thus, the fragmentation of androstane-2,11,17-trione (**32**) parallels that of androstane-2,11-dione (**23**); androstane-6,11,16-trione (**33**) displays a spectrum very similar to that of androstane-6,11,17-trione (**34**); and the decompositions of androstane-3,6,16-trione (**35**) and androstane-3,6-dione (**25**) are related, as are the major fragmentations of androstane-1,6,17-trione (**36**) and androstane-1,6-dione.

(c) As indicated by the scores and positive and negative evidence in Table VI, the rules have explanatory power which is considerably better than that for the diketoandrostanes and parallels that for the monoketones. The triketoandrostane rules explain 60.5%  $\Sigma$  of the total ion current (including the molecular ion) or 84%  $\Sigma$  of the data explained by INTSUM. There are very few counterexamples. In addition, rules T-1-T-10 predict many of the salient features of available (low-resolution) spectra of "new" compounds (not used for rule formation), such as androstane-3,12,17-trione<sup>19b</sup> and androstane-3,7,17-trione.<sup>18</sup> Ions deduced from these rules for these two compounds are: for androstane-3,12,17-trione: *m/e* 41, 43, 53, 55, 56, 67-69, 77, 79, 81, 91, 93-95, 110, 119, 121-124, 135, 137, 138, 165, 259, and 287, and for androstane-3,7,17-trione: *m/e* 41, 43, 53, 55, 56, 67-69, 77, 79, 81, 91, 93-96, 105, 107-110, 121, 123, 124, 131, 133, 135, 136, 149, 151, 152, 177, 179, 191, 192, 259, and 287.

(d) The discriminatory power of rules T-1-T-10 is quite high, paralleling the discriminatory power of the diketoandrostane rules. The predicted spectra of the 165 possible triketoandrostanes were compared to the known spectrum of each of the compounds **32-39**. The ranking of the correct structure was 4, 5, 3, 4, 9, 20, 4, and 12, respectively.

## Conclusions

We have shown that the Meta-DENDRAL program is ca-

pable of rationalizing the mass-spectral fragmentations of sets of molecules in terms of substructural features of the molecules. On known test cases, aliphatic amines and estrogenic steroids, the Meta-DENDRAL program rediscovered the well-characterized fragmentation processes reported in the literature. On the three classes of ketoandrostanes for which no general class rules have been reported, the mono-, di-, and triketoandrostanes, the program found general rules describing the mass spectrometric behavior of those classes. The general rules shown in Tables II, IV, and VI explain many of the significant ions for compounds in these classes, while predicting few spurious ions. The program has discovered consistent fragmentation behavior in sets of molecules which have not appeared by manual examination to behave homogeneously in the mass spectrometer.

Programs with knowledge of the scientific domain can provide "smart" assistance to working scientists, as shown by the reasoned suggestions this program makes about extensions to mass spectrometry theory. We are aware that the program is not discovering a new framework for mass spectrometry theory; to the contrary, it comes close to capturing in a computer program all we could discern by observing human problem-solving behavior. It is intended to relieve chemists of the need to exercise their personal heuristics over and over again, and thus we believe it can aid chemists in suggesting more novel extensions to existing theory. It can be argued that the two-dimensional connectivity model of molecules used in this study is not the right model for mass spectrometry; that there are deeper rationalizations of a fragmentation process than subgraph environments. However, this model is commonly used by working chemists and once fragmentations based on this model are defined, chemists can readily provide the remaining "mechanistic" rationalizations or see that further experimental work with labeled compounds is necessary. (Other limitations of the method have been discussed at the end of the Methods section.)

Recent statistical pattern-recognition work<sup>28</sup> addresses some of the points on rule formation and spectrum prediction raised in this paper. We have avoided blind statistical methods for three important reasons. (1) We wish to explore thousands of possible subgraphs with associated features, as we search for those which are in some way important. Current pattern-recognition procedures are restricted to much smaller numbers of manually (or computer-assisted) selected features, adding additional bias to the procedure. (2) We want to know how certain rules were obtained by the program and why certain other rules were rejected or not detected. We can trace the reasoning steps of the Meta-DENDRAL program and determine chemically meaningful answers to such questions in a way that is not possible with purely statistical programs. (3) We wish to constrain the rule formation activity in ways that are natural to a working chemist. For example we may want the program to avoid fragmentations involving aromatic rings or two bonds to the same atom, or, as mentioned above, we may want to look at fragmentations accompanied by loss of CO or other neutral fragments.

Rules can be formulated to explain data in terms that are known to be meaningful to chemists; most importantly, the rule-formation constraints are under the control of the chemist. Also we feel that this approach provides a high level of generality in describing fragmentation processes. Although the rules are developed in the context of a particular set of compounds, they are not tied to that set, but can be applied in other contexts or compared to rules developed from other sets of compounds in a search for common features of the rules. For these reasons, we believe that the Meta-DENDRAL program offers a powerful and useful complement to pattern-recognition programs for finding relationships between structures and spectral data.

Table VI. Triketoandrostane Rules<sup>a</sup>

Name	Subgraph	Other descriptors, and interpretation	Score	Positive evidence		Negative evidence	Average intensity, % $\Sigma_{40}$
				Any	Unique		
T-1 (-H, -H <sub>2</sub> O + H, -H <sub>2</sub> O - H, -CO, -CO + H, -CO -H)			143.6	13	8	0	6.84
T-2 (none, -H, -H <sub>2</sub> O + H, -H <sub>2</sub> O + 2H, -H <sub>2</sub> O - H, -CO, -CO - H)			77.4	16	1	0	4.55
T-3 (-H, -H <sub>2</sub> O + H, -CO + H, -CO -H)			72.6	7	7	0	5.19
T-4 (-H, -CO - H, -2CO - H)			70.5	16	8	0	2.94
T-5 (none, -CO)		Loss of methyl	59.7	14	14	2	2.49
T-6 (none, -H, -H <sub>2</sub> O + H, -H <sub>2</sub> O - H, -CO - H)			53.6	9	3	0	4.46
T-7 (+H, -CO + H, -CO - H, -2CO -H)		Atoms a and d are not keto-substituted	34.4	10	10	1	1.91
T-8 (none, -H, -H <sub>2</sub> O + H, -H <sub>2</sub> O + 2H, -H <sub>2</sub> O - H, -CO - H)			27.7	6	5	0	2.52
T-9 (+H)			14.0	31	31	0	0.23
T-10 (none, -H, -2CO - H)			13.1	5	5	0	2.19

<sup>a</sup> See footnotes to Table II for an explanation of terms.

We are cautiously optimistic about the general applicability of this rule-formation method, although we have demonstrated its utility for only a small number of compound classes and only in the context of mass spectrometry.

### Experimental Section

All high-resolution mass spectra were obtained by electrical recording of a complete spectrum (magnetic scanning) using perfluorokerosene as an internal mass calibration standard. All compounds were introduced into a Varian MAT Model 711 mass spectrometer via the direct insertion probe.

The computer programs are written in Interlisp and run on the DEC 10 SUMEX-AIM computer resource at Stanford University. For additional information on access to these programs, contact the authors.

**Acknowledgment.** This research was supported in part by the National Institutes of Health (Grants RR-612-06, GM-29662, and AM-04257) and the Advanced Research Projects Agency (Contract DAHC 15 73 C 0435). Computing support was furnished by the SUMEX-AIM computer resource at Stanford University, funded by the Biotechnology Resources Branch of the National Institutes of Health (Grant RR-00785).

We thank Sir Ewart R. H. Jones (Oxford University) for providing us with samples of some of the di- and triketoandrostanes and Annemarie Wegmann for performing the mass spectral analyses. We are indebted to Tom Mitchell, who programmed the RULEMOD system, to Steen Hammerum (H.

C. Ørsted Institute), who assisted in our analysis of the INT-SUM output, and to Jim McDonald and Paul Oppenheimer, who assisted greatly in producing the program's results shown here.

### References and Notes

- (1) (a) For Part 21, see C. J. Cheer, D. H. Smith, C. Djerassi, B. Tursch, J. C. Braekman, and D. Dalozé, *Tetrahedron*, in press; (b) Visiting scientist in 1974 from the IBM Research Laboratory, San Jose, California 95193.
- (2) (a) D. H. Smith, B. G. Buchanan, R. S. Engelmores, A. M. Duffield, A. Yeo, E. A. Feigenbaum, J. Lederberg, and C. Djerassi, *J. Am. Chem. Soc.*, **94**, 5962 (1972); (b) D. H. Smith, B. G. Buchanan, R. S. Engelmores, H. Adlercreutz, and C. Djerassi, *J. Am. Chem. Soc.*, **95**, 6078 (1973).
- (3) R. E. Carhart, D. H. Smith, H. Brown, and C. Djerassi, *J. Am. Chem. Soc.*, **97**, 5755 (1975).
- (4) We view this problem in a way that was suggested by previous work:<sup>2,3,5,30</sup> finding explanations of data can be considered as a problem of efficiently searching a (very large) space of possible explanations for the ones that, in some sense, seem best.
- (5) D. H. Smith, B. G. Buchanan, W. C. White, E. A. Feigenbaum, C. Djerassi, and J. Lederberg, *Tetrahedron*, **30**, 3117, (1973).
- (6) An important extension of the earlier work has been to broaden the definition of a process to allow the transfer of any specified neutral composition to or from the charged fragment with hydrogen transfers simply being one instance of the more general mechanism. The practical effect of this extension is to allow the chemist to consider molecular fragmentations occurring in conjunction with the loss of some neutral species such as H<sub>2</sub>O or CO.
- (7) We used an intensity threshold for peak selection (1%  $\Sigma$  for the low resolution spectra of the amines, 0.5%  $\Sigma$  for the remaining examples); alternatively one can use peak clusters,<sup>29</sup> reject low-mass hydrocarbon ions, or select only ions above a certain mass or possessing greater than a minimum number of carbon atoms. Based on our experience, peaks above this intensity are easily distinguished from noise in routine mass spectral analysis.
- (8) We presently specify this distance as two bonds (three atoms) away from

- the cleavage site. However, we have made this flexible in order to consider extending the distance to cover the entire molecule.
- (9) Because the space of rules is very large (much larger than in the simplified illustration in Figure 3), the rule generator is guided in its search more closely than is suggested by the simple incremental specification picture presented. It uses an abstract description of rule classes and it avoids generating rules that have no empirical support. The program avoids considering whole classes of rules by noting that specifying any value for type, nbrs, nhs, and/or dots at an atom place in a subgraph will give no improvement in the score for the emerging rule. These heuristics have been discussed previously.<sup>31</sup>
  - (10) The precise stopping conditions for RULEGEN compare a daughter rule with its parent and indicate that the program should continue generating daughter rules as long as the daughters are "improvements" over their parents. It should be noted that the following definition of improvement is under the chemist's control so that the degree of specificity of rules generated by the program can be changed to suit an individual's problem. The program judges a rule to be an improvement over its parent if three conditions hold: (a) the daughter rule predicts fewer ions per molecule than its parent (i.e., the daughter is more specific); (b) it predicts fragmentations for at least half of all the molecules (i.e., the daughter is not too specific); and (c) either the daughter rule predicts fragmentations for as many molecules as its parent or the parent rule was "too general" in the following sense: the parent predicts more than two ions in some single molecule or, on the average, it predicts more than 1.5 ions per molecule.
  - (11) When we speak of ions predicted by a set of rules we mean those ions that are deductive consequences of applying the rules to a molecule. This does not imply that the predicted ions are previously unreported.
  - (12) Rule 3 totally subsumes rule 1, but both are left in the set of candidates after initial selection because rule selection is order dependent and rule 1 was selected on its own merits before rule 3 was examined.
  - (13) All of the amines considered were unbranched at the  $\alpha$  carbon. Thus in choosing to generalize beyond the data we increase the range of applicability of the rule at some risk of being too general. We can minimize the risk either by having a sufficiently varied set of data over which the program generalizes or by not generalizing beyond the scope of the data.
  - (14) H. Budzikiewicz, C. Djerassi, and D. H. Williams, "Mass Spectrometry of Organic Compounds," Holden-Day, San Francisco, Calif., 1967.
  - (15) H. Budzikiewicz, C. Djerassi, and D. H. Williams, "Structure Elucidation of Natural Products by Mass Spectrometry", Vol. II, Holden-Day, San Francisco, Calif., 1964, p 64, and references cited therein.
  - (16) M. Spiteller-Friedmann and G. Spiteller, *Fortschr. Chem. Forsch.*, **12**, 440 (1969), and references cited therein.
  - (17) G. Jones and C. Djerassi, *Steroids*, **10**, 653 (1967), and references cited therein.
  - (18) H. Obermann, M. Spiteller-Friedmann, and G. Spiteller, *Chem. Ber.*, **103**, 1497 (1970).
  - (19) (a) F. J. Hammerschmidt and G. Spiteller, *Tetrahedron*, **38**, 3995 (1973); (b) E. Zeitz and G. Spiteller, *ibid.*, **39**, 597 (1974).
  - (20) Sir E. R. H. Jones, G. D. Meakins, J. O. Miners, J. M. Pragnell, and A. L. Wilkins, *J. Chem. Soc., Perkin Trans. 1*, 1552 (1975), and earlier references.
  - (21) L. Tokes and C. Djerassi, *J. Am. Chem. Soc.*, **91**, 5017 (1969).
  - (22) See (a) W. L. Fitch and C. Djerassi, *J. Am. Chem. Soc.*, **96**, 4917 (1974); (b) J. R. Dias and C. Djerassi, *Org. Mass Spectrom.*, **14**, 385 (1972), and references cited therein.
  - (23) The following constraints were employed to define possible fragmentation processes: cleavage of more than one (nonhydrogen) bond to the same carbon atom was forbidden; only single-step processes were considered (i.e., bond cleavages that break the molecule into two parts); only processes which cleaved a maximum of two bonds were considered; only hydrogen atoms (from loss of two to a gain of two hydrogen atoms) were considered as neutral transfers;<sup>6</sup> all ions above 0.5%  $\Sigma$  total ionization were considered. These constraints were used because past studies<sup>15,16</sup> have not indicated substantial contributions to diagnostic ions from more complex processes.
  - (24) J. Gutzwiller and C. Djerassi, *Helv. Chim. Acta*, **49**, 2108 (1966).
  - (25) Exceptions—the abundant ion at  $m/e$  135 in the spectrum of androstan-6-one (**13**) is predominantly  $C_{10}H_{15}$  (75%  $\Sigma$ ), not  $C_9H_{11}O$  (25%  $\Sigma$ ), and thus arises largely from a process different from that described previously<sup>16</sup> (INTSUM suggests cleaving the C-7,8 and C-9,10 bonds, retaining the charge on the hydrocarbon moiety in concert with loss of a hydrogen atom); the ion at  $m/e$  217 in the spectrum of androstan-16-one<sup>26</sup> is predominantly of composition  $C_{15}H_{21}O$  (88%  $\Sigma$ ), not  $C_{16}H_{25}$  (12%  $\Sigma$ ),<sup>16</sup> so cannot be due to ring D loss. INTSUM suggests loss of ring A together with a hydrogen atom as a potential source of the  $C_{15}H_{21}O$  ion, a fragmentation which parallels the decomposition of androstane (**9**).<sup>21</sup>
  - (26) (a) H. Budzikiewicz and C. Djerassi, *J. Am. Chem. Soc.*, **84**, 1430 (1962); (b) J. E. Gurst and C. Djerassi *ibid.*, **86**, 5542 (1964).
  - (27) R. T. Aplin and P. C. Cherry, *Chem. Commun.*, 628 (1966).
  - (28) G. S. Zander and P. C. Jurs, *Anal. Chem.*, **47**, 1562 (1975).
  - (29) R. G. Dromey, B. G. Buchanan, D. H. Smith, J. Lederberg, and C. Djerassi, *J. Org. Chem.*, **40**, 770 (1975).
  - (30) A. M. Duffield, A. V. Robertson, C. Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg, *J. Am. Chem. Soc.*, **91**, 2977 (1969).
  - (31) B. G. Buchanan, Proceedings of the NATO Advanced Study Institute on Computer Oriented Learning Processes, Bonas, France, 1974.

## Organometallic Chemistry of the Carbon-Nitrogen Double Bond. 1. Nickel Complexes Prepared from Iminium Cations and the X-Ray Structure of $\{[(C_6H_5)_3P]Ni[CH_2N(CH_3)_2]Cl\}$

Dennis J. Sepelak,<sup>1a</sup> Cortlandt G. Pierpont,<sup>\*1b,2a</sup> E. Kent Barefield,<sup>\*1a,2b</sup> Jerome T. Budz,<sup>1a</sup> and Craig A. Poffenberger<sup>1a</sup>

Contribution from the W. A. Noyes Laboratory, School of Chemical Sciences, University of Illinois, Urbana, Illinois 61801, and the Department of Chemistry, West Virginia University, Morgantown, West Virginia 26506. Received July 1, 1975

**Abstract:** Reactions of iminium salts with nickel(0) complexes,  $(R_3P)_2NiC_2H_4$  and  $(R_3P)_4Ni$  ( $R_3P$  is triphenylphosphine or tri-*p*-tolylphosphine), are reported. Complexes of stoichiometry  $\{[(R_3P)Ni(X)L]\}$  ( $X$  is Cl, Br, or I;  $L$  is iminium cation  $H_2C=N(CH_3)_2^+$  or  $H_2C=N(CH_3)(CH_2CH_3)^+$ ) and  $\{[(C_6H_5)_3P]_2Ni[CH_2N(CH_3)_2]ClO_4\}$  were isolated and characterized by analysis and infrared and  $^1H$  NMR spectroscopy. A single-crystal x-ray structure determination was performed on  $\{[(C_6H_5)_3P]Ni[CH_2N(CH_3)_2]Cl\}$ . The orange complex crystallizes in the monoclinic space group  $P2_1/c$  with  $a = 9.695$  (3) Å,  $b = 14.749$  (3) Å,  $c = 14.276$  (3) Å,  $\beta = 101.42$  (4)°,  $Z = 4$ ,  $\rho_{\text{exptl}} = 1.378$  (5) g/cm<sup>3</sup>, and  $\rho_{\text{calcd}} = 1.375$  g/cm<sup>3</sup>. The structure was solved using 2243 reflections with intensity greater than  $2\sigma$ . The positions of all hydrogen atoms were located and inclusion of these atoms in the structure refinement gave final discrepancy indices of  $R_1 = 0.041$  and  $R_2 = 0.045$ . The complex can be considered as a trigonally coordinated molecule with the iminium cation bonded in a  $\pi$ -fashion to the nickel atom; the carbon atom of the iminium cation is trans to the chlorine atom and the nitrogen atom trans to the phosphorus atom. The dihedral angle between the C-N bond and the Cl-Ni-P plane is 3.8 (2)°. The C-N bond length is 1.392 (6) Å. Ni to C, N, Cl, and P bond distances are 1.884 (5), 1.920 (4), 2.213 (2), and 2.136 (2) Å, respectively. Bonding in the complex is considered in terms of a  $\pi$ -alkene model. Reactions of this complex are described including one with sodium cyclopentadienide that yielded  $\{(\eta^5-C_5H_5)Ni[CH_2N(CH_3)_2]P(C_6H_5)_3\}$ , which contains a dimethylaminomethyl group  $\sigma$ -bonded to nickel.

This paper represents the first of what we anticipate will be a series on the organometallic chemistry of small unsaturated species that contain nitrogen atoms and that are iso-

structural and/or isoelectronic with olefins. Examples of such species include  $R_2N=CR_2^+$  (iminium cations), which are both isoelectronic and isostructural with the analogous  $C=C$