STANFORD ARTIFICIAL INTELLIGENCE LABORATORY
MEMO AIM - 215

STAN-CS-73-387

ANALYSIS OF BEHAVIOR OF CHEMICAL MOLECULES: RULE FORMATION ON NON-HOMOGENEOUS CLASSES OF OBJECTS

BY

Bruce G. Buchanan
N. S. Sridharan

SUPPORTED BY

ADVANCED RESEARCH PROJECTS AGENCY
ARPA ORDER NO. 457
& NATIONAL INSTITUTES OF HEALTH
August 1973

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

AI MEMO 215
STAN-(X-73-387

ANALYSIS OF BEHAVIOR OF CHEMICAL MOLECULES: RULE FORMATION ON NON-HOMOGENEOUS CLASSES OF OBJECTS *

Bruce G. Buchanan
N.S. Sridharan

Abstract:

An information processing model of some important aspects of inductive reasoning is presented within the context of one scientific discipline. Given a collection of experimental (mass spectrometry) data from several chemical molecules the computer program described here separates the molecules into "well-behaved" subclasses and selects from the space of all explanatory processes the "characteristic" processes for each subclass. The definitions of "well-behaved" and "characteristic" embody several heuristics which are discussed. Some results of the program are discussed which have been useful to chemists and which lend credibility to this approach.

*This research was supported by the Advanced Research Projects Agency (SD-1831 and the National Institutes of Health (RR-612).

INTRODUCTION

Induction in science has been understood to encompass many different levels of tasks, from theory construction as performed by Einstein to everyday non-deductive inferences as made by scientists looking for explanations of routine data. For the most part, it is not well defined however one understands it (a notable exception being statistical inference). Although general statements can be made about non-deductive inference, it is unlikely that there exists one general "inductive method" that guides scientific inference at all levels. Nor does it seem likely that a method of scientific inference at any one level can succeed without recourse to task-specific information, that is, information specific to the particular science. Within these assumptions we are exploring an information processing model of scientific inference in one discipline.

A unifying theme in our explorations is that induction is efficient' selectfon from the domain of all possible answers. Previous papers on the Heuristic DENDRAL Program (1) have advanced this theme with respect to hypothesis formation in routfne scientific work. Recently, we have been exploring this theme wtth respect to the higher-order task of finding general rules to explain large collections of data (2). This paper extends the previous work to the task of finding rules for subclasses of objects, given empirical data for the objects but without prior knowledge of the number of subclasses or the features that characterize them.

THE TASK AREA

For reasons discussed previously (2), the task area is mass spectromet ry, a branch of organic chemistry. The rule formation task is to find rules that characterize the behavior of classes of molecules in the mass spectrometer, given the mass spectrometric data from several known molecules,

The chemical structure of each molecule is known. The data for each molecule are a) the masses of various molecular fragments produced from the electron bombardment of the molecule in the instrument and b) the relative abundances 0f fragments at each mass. The data for each molecule are arranged in a fragment-mass table (FMT), or mass spectrum. Typically, there are SO-100 data points in one FMT. The task is to characterize the experimental behavior of the whole class of molecules.

Rules which characterize the behavior of the molecules are represented as conditional sentences in our system. The antecedent of a simple conditional rule is a predicate which is true or false of a molecule (or class of molecules); the consequent is a description of a mass spectrometric action (henceforth "process") which IS thought to occur when that molecule is in the experimental context. We have termed these rules "situat ion-act ton rules" (or "S-A rules?. The rule syntax has been described previously (3) and is not critical to an understanding of the present paper.

An example of a *rule*, rewritten in English, is: "IF the graph of the molecule contains the estrogen skeleton, THEN break the bonds between nodes labeled X3-17 and 14-15." This process (the consequent of thts rule,) is named BRK10L in Table I. The graph of the estrogen skeleton mentioned in the antecedent is shown with the conventional node numbering in Figure 3.

The rules will he used in the Heuristic DENDRAL performance program to determine the structure of compounds, reasoning from the mass spectrometric data of each. They are also of use to chemists interested in extending the theory of mass spectrometry.

OVERVIEW OF METHOD

The rule formation program contains three major sub-programs, which are described below under the headings Data Interpretation, Process Selection, and Molecule Selection. The control structure for the overall program is described after the discussions of the three major sub-programs. A brief overview of the whole program will he given first, however, in order to set the context.

The purpose of the program is to find the charactertstic processes which determine separable subclasses of molecules given the experimental data and molecular structure of each molecule. The overall flow of the program, as described below, is shown in Figure 1. The three major steps are to reinterpret the experimental Rata as molecular

Processes, find the characteristic processes for the given molecules, and select the set of molecules that are "well-behaved" with regard to the characteristic processes. The reinterpretation of the data is done once for each molecule in the whole set, and the results are summarized once. The second and third sub-programs are called successively until they isolate a well-behaved subclass of molecules and determine the processes which characterize their behavior. The monitor then subtracts the well-behaved subclass from the starting class of molecules, and repeats the successive calls to the second and third subprograms. The whole program stops when there are N or fewer molecules not yet in some well-behaved subclass. (For now, N=3.)

The data interpretation program has been described previously with some aspects of the process selection program (3). The molecule selectfon program and class refinement loop in the control sequence are new additions.

DATA INTERPRETATION

As mentioned above, the purpose of the Rata interpretation and summary program (INTSUM) is to reinterpret the experimentally determined data, the FMT, for each molecule and summarize the results. Because the program has been described previously (3), details will he omitted here. It should he noted that the successful application of this program to a sub-class of estrogens has already been reported in the chemical 1 iterature (4). The INTSUM program is general in that it

will work on FMT's for any class of molecules with a common skeletal graph and it is flexible in that the knowledge used by the program is easily changed and there are numerous options controlling the operation of the program.

The INTSUM program is called with the initial set of molecules and their FMT's. It is also given the graph structure of the skeleton common to all molecules in the initial set. The first step is to search the space of all possible processes which could explain data points in the FMT of any molecule with the given skeleton, The space of explanatory processes is camhtnatortal; simple processes that cut the graph into two fragments are generated first, followed by pairs of simple processes, triples, and so on. The heuristics listed below constrafn the search:

Simplicity (Occam's Razor)

If two or more processes explain the same data point, prefer the simpler one, i.e., the process involving fewer simple steps.

Chemical Constraints

(a) Break no more than NB bonds in any process, whether simple or multi-step (NB=5 in our current version); (b) Do not allow any process to break two bonds to the same carbon atom; (c) Do not allow a fragment to contain fewer than NA atoms (NA=5 currently); (d) Do not allow any

process to contain more than NP simple processes (NP=2 currently); (e) Break only single bonds (no double or triple bonds).

The heuristic search produces a list of plausible processes without reference to the data. The second step of the INTSUM program is to determine for each process and each FMT whether there is evidence for the process in the FMT. If so, then that process can explain the data point and the strength of the evidence is saved. The final step is to summarize for each process and all molecules the frequency, total strength of evidence and number of alternative explanations. (Frequency for a given process is the percentage of all molecules that have evidence for the process.) These statistics are passed to the process select ion program.

PROCESS SELECTION

The process selection program chooses the most characteristic processes for the given class of molecules from the list of a priori plausible processes that are output by the INTSUM program. It assumes that the molecules given to it are all in one well-behaved class. Thus, it can merely filter the list of processes to find those which satisfy the criteria for characteristic processes.

A process mentioned in a rule statement must satfsfy several criteria in order to be counted as a characteristic process for the molecules under consideration. The INTSUM program provides a summary of

statistics for the plausible processes it has chosen from the space of all processes. The process selection program applies heuristic criteria to sort out the most likely processes and to distinguish among alternative explanations, when alternatives remain. It uses the information from the data for filtering, in contrast to the a priori filtering in the INTSUM program. For example, an a priori simplicity criterion filters out processes that break too many bonds. The criteria for "most likely processes" -- frequency, strength of evidence, and degree of uniqueness -- are discussed below. To a large extent the choice of these criteria and particularly the choice of parameter settings are arbitrary. However, the following discussion provides some rationale for our choices.

Frequency

If nature presented clear and unambiguous data to us we could expect all and only characteristic processes for a class of molecules to occur 100% of the time. This is what we would like to mean by 'characteristic' process. Yowever, the data contain noise and, more Importantly, we are Forced to Tnterpret the data in terms of processes that we construct. Thus, in the literature one finds discussions of exceptions to rules together with presentation of the rules. A low frequency threshold (60%) is used as a criterion for plausible process Instead of a high one because the marginal processes which are included at one steo can be excluded at a later reffnement step if they prove to be uncharacteristic of a class of molecules.

Strength of Evidence

The program considers the strength of evidence found for each process, besides the frequency of molecules that show the process. Associated with each fragment mass in the experfmental data is a measure of the percent of total ions (or ion current) contributed by fragments of that mass. (The evidence from mass spectrometry is not merely binary, i.e., yes/no, although we have considered it that way in the past,) The total ion current for any molecule can be visualized as the sum of all y-values in a bar graph in which- the x-values represent fragment masses. The strength of evidence for a process, then, is the percent of the total of all ion currents (for all molecules) that can be explained by the process. The present value of this parameter is 0.005, i.e., 0.5% of the data must be explained by any process that will be said to be characteristic of the given molecules.

There may be much information in the weaker data points, but until we can interpret the strong signals, we do not want to start looking critically at the weak ones. This is why we have a strength of evidence threshold (although in our trials we have kept it fairly low).

Degree of Uniqueness

The program will discard processes that cannot uniquely explain at least n data points for each molecule. The rationale behind thfs criterion is that processes that are always (or often) redundant with

other processes have no explanatory power of their own.' In spite of the intuitive appeal of this criterion, it was not used for the trials reported here in which molecule selection is coupled with process selection. For process selection alone, it is a useful filter.

These three criteria filter the processes to provide the characteristic processes for the molecules given to the program. However, the processes may still overlap in the data points that they explain. If two (or more!) processes are ambiguous, i.e., they explain most of the same data points, the program tries to resolve the ambiguity in favor of a single explanation. This is not easy, for the competing explanations have all passed the tests for "most likely processes" just discussed. Thus, they all appear good enough to be rules on their own.

The resolution of ambiguities among processes is made according to relative values of the criteria used to judge them likely in the first place. That is, the values of frequency, strength of evidence and degree of uniqueness are compared - in any order - to determine which process is preferred, if any.

MOLECULE SELECTION

Molecule selection, by itself, is a simple program whose purpose is to find a subclass of molecules that are "well-behaved" with respect to a set of processes. Its inputs are (a) a class of molecules and (b)

a set of processes that are characteristic of those molecules (output of the process selection program just described).

The processes that are chosen as roughly characteristic of a class of molecules are used by the molecule selection program to refine the extension of the class. Several processes will each have a few except ions - the number permitted depending on the frequency threshold used by the program. But if the same molecules appear as exceptions over and over again (for several processes) then they probably do not belong in the same subclass with the molecules whose behavior is characterized by those processes.

A molecule is said to be well-behaved with respect to a set of processes (or well-behaved) if it shows evidence for at least MP of the processes. The current value of MP is 85% of the number of processes in the set. Currently this is the only criterion used to identify members of the subclass, although other features of the molecules could also be used for clustering. For example, the structural features of chemical molecules could also help classify molecules which "belong" together. The reason descriptive features such as these are not used during molecule selection is that they constitute a good check (by chemists) on the adequacy of the results of the molecule separation procedure,

CONTROL STRUCTURE OF THE RULE FORMATION MONITOR

The overall flow of control has been briefly described and diagrammed

in Figure 1, and the three major components of the whole program have been discussed. The interaction between process selection and molecule selection Is the last important detail in the description of the program. It is shown schematically in Figure 2 and selected portions of intermediate output are shown in Table II.

After the INTSUM program interprets and summarizes the data for a set of molecules, the process selection program is asked to find a set of processes that characterize those molecules. However, process selection starts with the assumption that the molecules should be characterized al l together, i.e., that the molecules are homogeneous, or belong In one class with respect to mass spectrometry, The purpose of the rule formation monitor, and the molecule selection program In particular, is to remove the necessity of working within this assumption. Because a class of molecules has a common skeleton, there is reason to believe that they are homogeneous (with respect to mass spectrametry processes). Rut this is not necessarily true. Many of the molecules whose structures contain the graph common to estrogens (e.g., the equilenins discussed with Table III In the Results section) fail to exhibit behavior that is characteristic of most estrogens in the mass spectrometer.

The monitor begins with the Null Hypothesis that the initial set M of molecules Is homogeneous with respect to all the relevant processes given as input. With the process selection program it finds plausible processes that roughly characterize the whole class of molecules. It

attempts to confirm the hypothesis by finding the subclass S of molecules that are well-behaved for those processes. If this subclass S is the same as the initial set M, then the assumption of homogeneity is taken to be true. In that case, there is no proper subset to be separated.

When the subclass S is different from the starting class M, however, the program loops back to process selection as shown in Figure 2. This figure shows the procedure for producing one homogeneous subclass of molecules (and the characteristic processes for the subclass); this procedure, rule formation, is itself used repeatedly in the main program as shown in Figure 1.

The inputs to the rule formation procedure are (a) the set RP of relevant processes and statistics for them, viz., the output of INTSUM, and (b) a class M' of molecules, where M' is initially the same as the entire class of molecules, M, given to INTSUM. M' is used to keep track of the best refinement of M so far.

The process selection program selects a set of processes P from RP in the manner described above. P characterizes the class M', insofar as M' can be characterized at all. The criteria for characteristic process can be made more restrictive if the class is known to be homogeneous (e.g., frequency >95%). In this case, however, the loose criteria listed above are used (e.g., frequency >60%) in order to allow many exceptions to the "characteristic" processes.

The molecule selection program selects a subclass of molecules S, from M', that are best characterized by the processes In P. The subclass S includes molecules that show evidence for most (85% or more) of the processes in P, and excludes molecules that are exceptions to many. Thus S is at least as wel 1 behaved as M' with respect to P. And since the two measures of selection are not perfectly complementary, S is likely to be better behaved than M' with respect to P, (If molecule selection uses less restrictive measures than process selection, then S will be less wel 1 behaved than M' and the procedure will fail except when the initial set of molecules is homogeneous.)

One Interest ing part of the procedure Is that after processes are selected, ALL of the molecules are reclassified with regard to the number of times they appear as exceptions to the processes. This shown i n Figure 2 : at step 2 of each level all molecules In the initial set, M (not M' or S), are tested against the processes. Thus, a molecule can be excluded at one level (because it is an exception to too many 0f the processes at that level), but be included again at another level for a slightly different set of processes.

The condition under which we want the program to stop is that the subclass S of molecules after an iteration is the same as the class M' from which the iteratton started (condition 1 in Figure 2). In other words, under this condition the program has found an S and a P such that P characterizes S(S=M') and S is well-behaved with respect to P. The subclass S Is taken to he homogeneous, and the

processes in P can be taken to be mass spectrometry rules for molecules in S.

The refinement level in Figure 2 is the number of times the procedure has been invoked in trying to find one homogeneous subclass of molecules. The second of the stopping conditions tests whether the refinement level is equal to an arbitrary maximum, which is currently 3. This condition is necessary to avoid an infinite loop in the case where the program can find no subclass S that is homogeneous with respect to P. The level 3 has been observed to produce fairly acceptable results: after three iterations through this loop, the subclass S is about as refined as it will get. After more iterations the procedure appears to oscillate in that molecules added to S in one iteration are subtracted from S in a later iteration. Our experience is very limited. Because there is no guarantee that the procedure converges, however, some stopping condition like the maximum refinement level is necessary.

The last stopping condition shown in Figure 2 tests whether there are enough molecules in the subclass to warrant further refinement. If there *are* fewer than an arbitrary minimum number (=3) of molecules in S, then further refinements will be unreliable. This minimum is not completely arbitrary, since it depends to some extent on the frequency measures used in process and molecule selection. Rut, intuitively, when the number of molecules in S is small there is little value in breaking S up into subclasses anyway.

As shown in the overall flow diagram, Figure 1, after the first major subclass (S) has been defined, all molecules in S are removed from any further consideration by subtracting them from M. The entire procedure is then repeated with the new M. It stops only when there are so few molecules left in M (3 or fewer) that process selection is unreliable and molecule selection appears pointless.

The output of the whole program now is merely the collected set of outputs from all iterations, viz., the collected S,P pairs, as shown in Figure 2. Future work will focus on automatically generalizing the descriptions of the molecules. This is now done by hand, except when the initial class M is homogeneous - then the generalized description is the common graph structure.

RESULTS

The INTSUM program alone has already provided useful new results for chemists, as reported in the chemical literature (4). The process selection program, working with output from INTSUM (but without molecule selection), has successfully found sets of characteristic processes for a well-understood class of molecules (estrogens, Figure 3) and for classes whose behavior is still under investigation (e.g., equilenins, progesterones, amino acids). For 47 estrogens, which were assumed by both an expert and the program to be in one class, rules found by the program agree closely with rules formed by the expert from the same data. (This result is not shown in a table,

but the comparison with the expert's rules looks much like that shown in Table I.) Expert chemists have made suggestions for improvements, but were generally in agreement with the processes selected by the program.

The rule formation program with molecule selection has been tested on several sets of molecules. The results of running the program on a set of 15 estrogens (a subset of the 47 mentioned above) are shown In Table I. The program separated two of the 15 compounds into a second class because they were not as well behaved as the rest - they were exceptions to about 20% of the characteristic processes. However, the chemist thought the separation was reasonable. The processes selected by the program are shown with indications of the discrepancies between the program's choices and the chemist's, The discrepancies mostly arose from the program's applying different criteria to select one process from viable alternatives. Tahle II shows the success of the molecule separation part of the program when rule formation was Rone on data from 19 non-homogeneous estrogenic steroids. The major subclass of chemical interest is the set of 5 equilenins which are identified by common modifications to the skeleton shown in Figure 3. The structural properties were not used by the program although the chemist did classify the compounds by such features. By selecting well-behaved subclasses of molecules the program grouped four or five "equilenins" (molecules #4, 8, 10, 19) and all three "3-acetates" (#3, 11, 18) in the first subclass. The fifth equflenin (#2) was removed from that subclass on the last refinement because it was an

exception to 3 of 9 characteristic processes used to determine the subclass.

In the third iteration shown in Table II, the program grouped three of the chemist's four "3-benzoates" together (molecules #12, 13, lb). In the fourth iteration it grouped together the chemist's two "diacetates" and one "triacetate" (molecules #9, 15, 16). Two iterations produced subclasses with only two members - when put together they encompass two "17-acetates" (#1, 17), one "17-benzoate", and one "gamma-lactone" (#5). The two molecules remaining unclassified at the end of the procedure were the last "equilenin" (molecule #2) and the last "3-benzoate" (#6).

CONCLUSIONS

Building an information processing model of scientific reasoning in mass spectrometry, although not completed, has already led to interesting and useful results. The model incorporates heuristic search in process selection. The procedure for selecting molecules can be thought of as a planning procedure insofar as it reduces the problem of formulating rules for a class of diverse molecules to a number of smaller subproblems, viz., formulating rules for smaller classes of well-behaved molecules. However, the molecule selection procedure is highly dependent on process selection, as described in detail.

The incompleteness of the program as a model of the entire rule formation procedure should be readily apparent. We have not described anything that approximates confrontation of rules with new data, for example. But as the results section indicates, the program can separate subclasses of well-behaved molecules and can find characteristic processes for the subclasses with enough accuracy (on a few examples) to gain preliminary acceptance by an expert in the field.

ACKNOWLEDGEMENTS

We have received invaluable assistance from colleagues. In particular, Mssrs. William White and Stephen Reiss have done almost all of the programming; Dr. Dennis Smith has provided chemical expertise; and Professors E.A. Feigenbaum and J. Lederberg have provided much of the intelligence in the design of the system. Financial support is gratefully acknowledged from the Advanced Research Projects Agency (SD-183) and from the National Institutes of Wealth (RR-612-02).

REFERENCES

(1) B.G. Buchanan and J. Lederberg, "The Heuristic DENDRAL Program for Explaining Empirical Data". In proceedings of the IFIP Congress 71, Ljubljana, Yugoslavia (1971). (Also Stanford Artificial Intelligence Project Memo No. 141).

(2) R.G. Buchanan, E.A. Feigenbaum, and J. Lederberg, "A Heuristic Programming Study of Theory Formation in Science." In proceedings of the Second International Joint Conference on Artificial Intelligence, Imperial College, London (September, 1971). (Also Stanford Artificial Intelligence Project Memo No. 145)

(3) B.G. Buchanan, E. A. Feigenbaum, and N.S. Sridharan, "Heuristic Theory Formation: Data Interpretation and Rule Formation". In Machine Intelligence 7, Edinburgh University Press (1972).

(4) D.H. Smith, B.G. Buchanan, W.C. White, E.A. Feigenbaum, C. Djerassi, and J. Lederberg, "Applications of Artificial Intelligence for Chemical Inference X. INTSUM. A Data Interpretation Program as Applied to the Collected Mass Spectra of Estrogenic Steroids", Tetrahedron (in press).

Figure 1. OVERALL FLOW OF RULE FORMATION PROGRAM

```
                    INPUT:  List of Molecule - Data Pairs
                                    |
                                    |
            PROGRAM:   INTSUM - Data Interpretation and Summary
                                    |
                                    |
------------------------>   List of Molecules, M.
|                   List of Relevant Processes, RP, with
|                   Summary Statistics for Each Process
|                                    |
|                                    |
|                        PROGRAM:  Rule Formation*
|                                    |
|                                    |
|                 Set of Characteristic Processes, P (P<⊆RP).
|                 Class of Well-Behaved Molecules, S(S<⊆M).
|                                    |
|                                    |
|               SUBTRACTION STEP:  Remove all Molecules in S
|                                 from M.
|                                    |
|                                    |
|               STOPPING CONDITION:  M contains 3 or Fewer
|    No                            Molecules.
--------------                      |
                                    |    Yes
                                    |
                                  STOP
                      OUTPUT = All S-P pa? rs found.
```

---

* Details in Figure 2.

Figure 2.

DETAILS OF INTERACTION BETWEEN PROCESS SELECTION AND MOLECULE SELECTION IN THE RULE FORMATION PROGRAM

```
INITIALIZE:   Refinement Level = 0
              M = Original class of molecules.
              M' = M.
              RP = Relevant processes (from INTSUM) including
                   evidence and statistics for the processes.
                   |
                   |
INPUT:        M', RP
                   |
                   |
---> SUB-PROGRAM: Process Selection (using the null hvpothesis that
|                 all molecules can he characterized by the same set
|                 of processes)
|                   |
|                   |
|                 Set of processes, P, that are characteristic of
|                 M' (P⊂RP)
|                   |
|                   |
|   SUB-PROGRAM: Molecule Selection
|                   |
|                   |
|                 Subclass of Molecules, S, selected from M such that
|                 every molecule in S is well-behaved with respect to
|                 the processes in P
|                   |
|                   |
|                 Increment Refinement Level
|                   |
|                   |
|   Test for Stopping Conditions:
|        1. S = M', or
|        2.  Refinement level = 3, or          YES      STOP.
|        3.  Fewer than 3 molecules in S -------->OUTPUT = S, P
|                   |
|                   | NO
|                   |
|                   |
---- SUBCLASS QEFINEMENT:  Reset M' to S (M' = S).
```

Figure 3.

GRAPH STRUCTURE OF THE ESTROGEN SKELETON

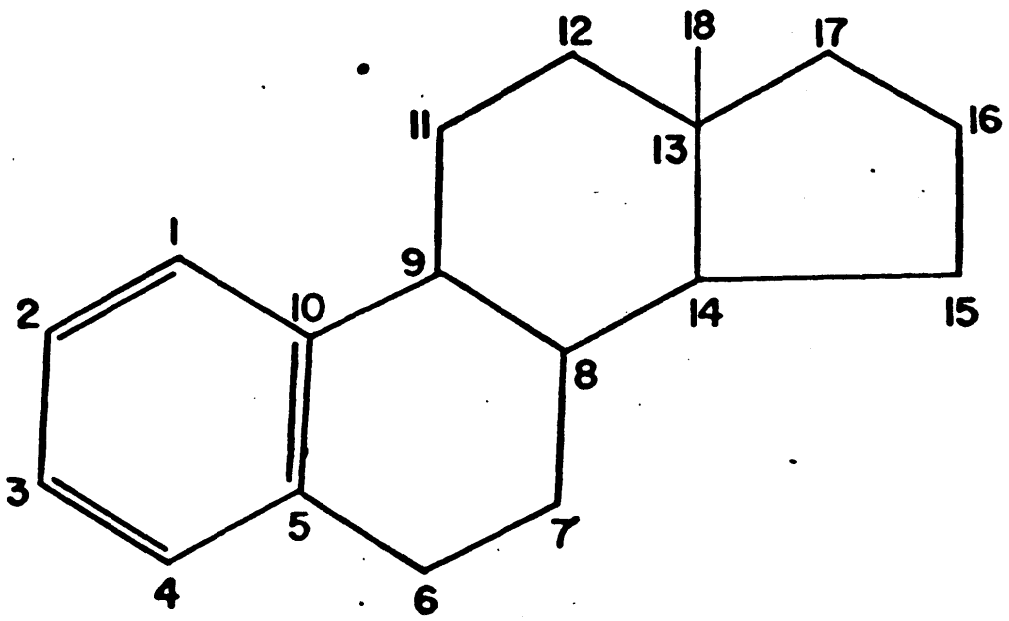
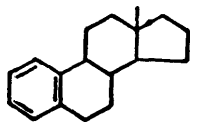
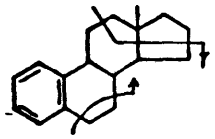
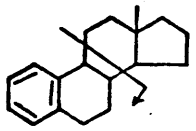
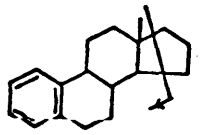
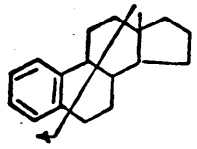
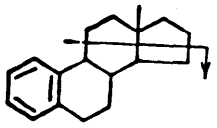
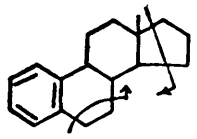
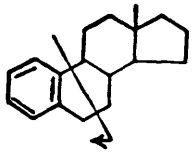
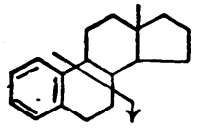
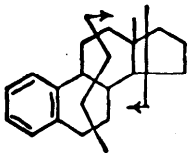


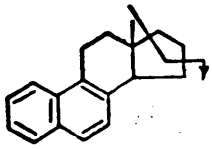
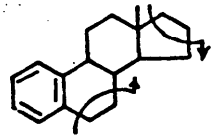
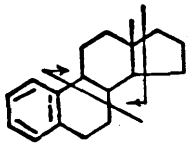
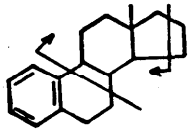
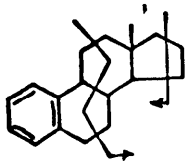
TABLE I.

PROCESSES SELECTED FOR 15 ESTROGENS
BELIEVED TO BE IN ONE WELL-BEHAVED CLASS

PROCESS LABEL*	PICTORIAL DESCRIPTION	% OF ALL DATA POINTS EXPLAINED
1. BRK0		22%
2. BRK2L/19L (preferred over BRK7L and BRK2L/18L)		14%
3. BRK6L or BRK2L/17L		11%
4. BRK10L		8%
5. BRK14L or BRK15L		6%
6. BRK17L		5%
7. BRK2L/10L (preferred over BRK18L)		4%
8. BRK4L		3%
9. BRK5L or BRK13L		2%
10. BRK10L/15H *or* BRK5H/20L or BRK4H/19L		2%

* The underlined processes are those selected by an expert chemist on the basis of data from 47 well-behaved estrogens, including these 15.

TABLE I, Page 2

PROCESS LABEL*	PICTORIAL DESCRIPTION	% OF ALL DATA POINTS EXPLAINED
11. BRK11L		2%
12. BRK2L/11L (preferred over BRK20L)		2%
13. BRK5H/10L		2%
14. BRK5H/12L		1%
15. BRK12L/15H or BRK12L/14H		1%
	TOTAL PERCENT OF DATA EXPLAINED	84%

* The underlined processes are those selected by an expert chemist on the basis of data from 47 well-behaved estrogens, including these 15.

TABLE I I

SUMMARY OF STEPS IN THE RULE FORMATION PROCEDURE WITH 19 ESTROGENIC STEROIDS

	Molecules		Processes

ITERATION #1			
Initial Set:	(1,2,3,...,19)	------->	BRK0 BRK10L BRK11L BRK20L BRK2L/19L BRKSUB3L/3L BRKSUB3L/12L
First Refinement:	(2,3,4,5,8,10,11,19)	-*-m-w->	RRKO BRK10L BRK11L BRK20L BRKOOC3*1L BRKSUB3L/2L BRKSUB3L/23L BRKSUB18L/11L
Second Refinement:	(2,3,4,8,10,11,18,19)	------->	RRKO BRK10L BRK11L BRK20L BRKOC3*1L/11L BRKOOC3*1L BRKSUB3L/2L BRKSUB18L/11L BRKSUB3L/23L
Third Refinement: = Subclass 1	(3,4,8,10,11,18,19)	------->	same
-----			
ITERATION #2			
Initial Set (- Subclass 1)	(1,2,5,6,7,9,12,13, 14,15,16,17)	------->	BRK0 BRK16L BRK2L/19L BRKSUB3L/3L
.			
.			
.			
Third Refinement = Subclass 2	(5,17)	------->	BRK0 BRK2L/19L BRKOC3*1L/8L BRKOC3*1L/17L BRKOOC17*1L

ITERATION #3

.
.

Third Refinement (11,12,13,14) -------> BRK0
= Subclass 3 BRKBT3*1H
BRKBT3*1L/3L
BRKSUB3L/3L

ITERATION #4

.
.
.

Last Refinement: (9,15,16) -------> BRK0
= Subclass 4 BRK00C3*1L
BRK00C3*1L/6L
BRK00C3*1L/7L
BRK00C3*1L/8L
BRK00C3*1L/16L
BRK00C3*1L/17L
BRK00C17*1L

ITERATION #5

.
.
.

Last Refinement: (1,7) -------> BPKO
= Subclass 5 BRK6L
BRK7L
BRK8L
BRK10L
BRK11L
BRK14L
BRK15L
BRK16L
BRK17L
BRK2L/17L
BRK2L/19L
BRK00C17*1L
BRKSUB17L
BRKSUB17L/1L

UNCLASSIFIED MOLECULES (2,6)