Report 77-11
Stanford -- KSL

Scientific DataLink

Structure Elucidation Based on Computer Analysis of High and Low Resolution Mass Spectral Data. Dennis H. Smith, Raymond E. Carhart, 1978

Structure Elucidation Based on Computer Analysis of High and Low Resolution Mass Spectral Data

DENNIS H. SMITH and RAYMOND E. CARHART

Departments of Chemistry, Genetics, and Computer Science, Stanford University, Stanford, CA 94305

A tremendous effort has been directed toward development of advanced instrumentation for mass spectrometric analysis. Advancements include ever-increasing sensitivities and resolving powers, new ionization techniques, metastable ion probes of ion decomposition and structure and computer systems for rapid acquisition and reduction of data. We sometimes lose sight of the fact that these developments are designed to provide information about chemical and biochemical structures at greater depth and in greater detail than previously available. The ultimate goal in most research in mass spectrometry is to provide powerful tools for molecular structure elucidation, either directly, by exploitation of existing techniques, or indirectly by development of new techniques.

Concurrently, several computer-based techniques designed to assist chemists in the analysis and interpretation of mass spectral data have been developed. Analytical procedures for treatment of combined gas chromatographic/mass spectrometric data obtained at low resolving powers (1) (gc/lrms)

© 0-8412-0422-5/78/47-070-325$10.00/0

Reprinted with permission from Smith, Dennis H. and Carhart, Raymond E. in "High Performance Mass Spectrometry: Chemical Applications," Michael Gross, Ed., in ACS SYMPOSIUM SERIES, No. 70; American Chemical Society: Washington, D.C., 1978, pp.325-347.

provide mass spectra of high quality for subsequent examination by manual or computer methods. Library search procedures (2) and their extensions (3) or pattern recognition programs (4) may provide clues to the identity of the structure or be used to determine the structure uniquely. A computer program for analysis of spectra based on class-specific fragmentation rules, is available (5). These techniques have obvious limitations (3,4,5); in fact, the ability to interpret mass spectral data in terms of molecular structure lags far behind the capabilities of modern spectrometers to produce high quality data. There are several reasons for this lag: (1.) There is no formal theory relating molecular structures to their respective mass spectra which has predictive power of use to the structural chemist. (2.) (a corollary of 1) Mass spectrometry rarely provides detailed substructural information to assist in elucidating a structure except in known chemical contexts where previously developed rules may be applied retrospectively. (3.) Current methods do not make adequate use of other knowledge about a particular compound, and; (4.) The ~~the~~ combinatorial complexity of dealing with the actual information content of a mass spectrum has not, until now, been addressed. Points (3) and (4) will be discussed in some detail subsequently.

In our laboratories we have been trying to bring newly-developed computational tools to bear on general approaches to assisting structural chemists in interpretation of mass spectra. In this paper we will discuss the strengths and limitations of these new tools while assuming that the requisite mass spectral data are available. We are engaged in research which involves the gc/ms analysis of complex mixtures together with subsequent analysis of these data to extract spectra of individual components (1b) and search for the spectra in libraries of mass spectral data. The gc/lrms analyses provide an important pre-screening of mixtures. Combined gc/ms data obtained at high mass spectrometer resolving powers (gc/hrms) (6) yield elemental composition data for novel components. In any case the computer programs described in subsequent sections accept either low or high resolving power data (nominal masses or elemental compositions, respectively).

Computer-Assisted Structure Elucidation Based

Primarily on Mass Spectral Data.

Assume that an important unknown component has been observed in data from a gc/ms analysis of a mixture (e.g., Figure 1). Assume further that the spectrum of the component (Figure 2) was not found in existing libraries. This problem becomes a classic problem of structure elucidation. One, for example, might attempt to isolate larger quantities and obtain additional spectral data. For many problems this is time-consuming and difficult. Realizing that high resolving power data are less ambiguous than data provided by the low resolution spectrum (Figure 2), one might obtain a gc/hrms spectrum and determine elemental compositions for the observed ions. The data obtained in the example are presented in Table I for major ions in the spectrum. These data may be the only data one can

TABLE I.

Elemental Compositions for Significant Fragment Ions in the Mass Spectrum Given in Figure 2.

m/e	Composition
293	$C_{15}H_{19}NO_5$
261	$C_{14}H_{15}NO_4$
234	$C_{13}H_{16}NO_3$
202	$C_8H_{12}NO_5$
174	$C_7H_{12}NO_4$
142	$C_6H_8NO_3$
116	$C_5H_{10}NO_2$
91	C_7H_7

easily obtain to determine structural information about the unknown. Many current programs operate under the assumption that these are the only data. But, of course, this is not true. In almost every instance a great deal of additional information about the unknown is available. Some of this information is factual, for example, the physical and chemical properties and the source of the materials,

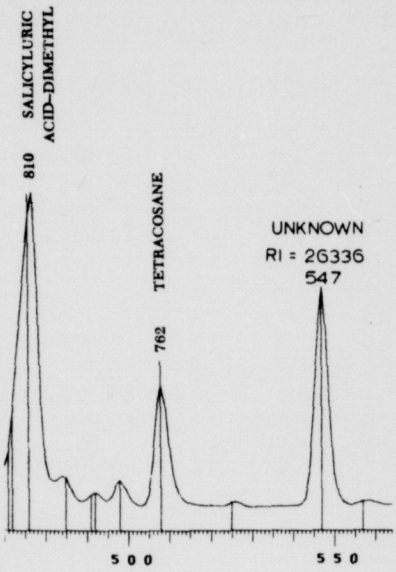


Figure 1. Total ion current plot, scans 470–565, of the GC/lrms analysis of the ether/ethyl acetate extracts of an acidified human urine subsequent to a 1 hr 1N NaOH hydrolysis. Numbers and names associated with component spectra detected by the CLEANUP program (1b) refer to the match scores and names of close library matches. Tetracosane, scan 508, is an internal standard for relative retention index calculations.

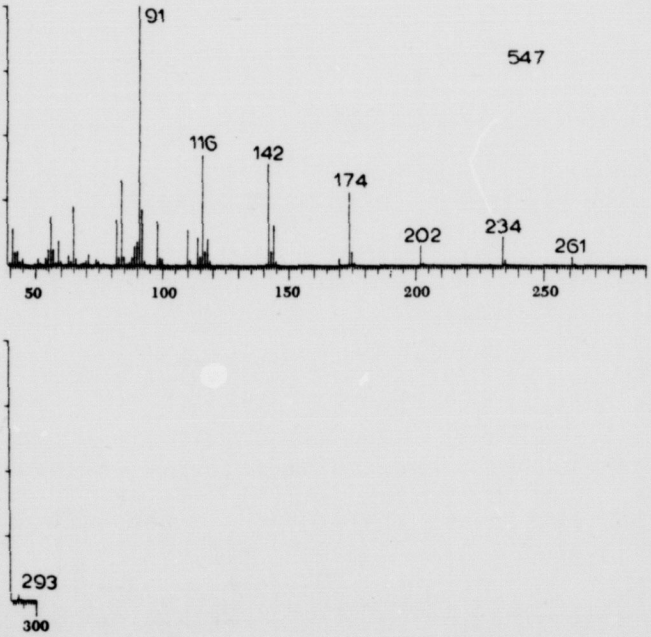


Figure 2. 70 eV low resolution mass spectrum obtained for the component eluting at scan 547 (see total ion current plot, Figure 1)

and isolation and derivatization procedures. Other information is judgmental; for example, knowledge of other compounds present in the same mixture, plausibility of chemical or biochemical processes which may have yielded the compound, and good intuitions. If this information can be brought to bear on the problem, it should be easier to solve.

We have developed the CONGEN program for computer-assisted structure elucidation. (7) This program has a flexible mechanism for expression and use of constraints on the plausibility of certain chemical features (substructures, ring systems). (8) Although designed for the general problem of incorporating structural inferences from different spectroscopic techniques or other sources of chemical information, we are introducing capabilities for more thorough use of mass spectral data. The capabilities were developed initially in earlier DENDRAL work. (5,9) The importance of CONGEN in problems such as that outlined above is that it gives the chemist a mechanism for exploring structural possibilities under constraints expressing his/her factual or judgmental knowledge. This knowledge can be defined, applied to a current problem, and saved for future use in related problems, as illustrated below.

Returning to the example; it is foolhardy to consider structural possibilites in the light of the presumed molecular ion, $C_{15}H_{19}NO_5$. Without constraints, the number of possible structures is huge. However, knowledge that the compound is a component of a mixture of organic compounds isolated from human urine, and that the urine was subjected to basic hydrolysis prior to extraction provides additional information which can to some extent be expressed as structural constraints (see below). More specific structural information is available from the fact that the fraction consists of ether/ethyl acetate extractable organic acids, which were subsequently esterified using diazomethane prior to gc/ms analysis.

This "history" of the sample provides a tremendous reduction in the scope of possible structures. Chemists use this reasoning automatically in manual examination of structural possibilities. To be truly effective, a program must somehow provide the same capabilities when confronted with a mass spectrum. To assist chemists in making use of such reasoning, we are extending CONGEN to allow exploration of structural possibilities for an unknown

within constraints provided by the mass spectrum and by factual and judgmental knowledge supplied by the chemist. We are developing two general approaches to the use of mass spectral data. These two approaches, MSPRUNE and Mass Distribution Graphs (MDG's), mirror the classic interplay between maximum use of information in retrospective testing vs. prospective guidance (planning) toward hypothetical solutions in the problem-solving paradigm of heuristic search. (5,7) In principle, the approches are complementary. They will yield the same answers by working on a problem from two different directions. In practice we have made more progress to date on the former (see below). We will illustrate both with examples.

I. MSPRUNE - Retrospective Testing of Structural Candidates.

If it is possible to arrive at a set of structural candidates for an unknown based on constraints derived from chemical considerations, other spectroscopic data and/or characteristic ions in a mass spectrum, it should be possible to test each candidate in some detail to determine if it is capable of producing the observed spectrum. MSPRUNE makes such tests and will "prune", or reject those structures which could not have yielded observed ions.

In many problems, including the example (Figure 2, Table I), it is possible to arrive at a reasonable set of candidate structures for an unknown using available data and the following general procedure.

A. Determine the molecular weight and formula. In the example, a candidate molecular ion is found at m/e 293, of composition $C_{15}H_{19}NO_5$. This ion is selected by MOLION (10), our molecular ion determin-

TABLE II.
Molecular Ion Candidates and Rankings for the Spectrum Given in Figure 2.

Candidate	Ranking
m/e 293	100
294	71
325	50
352	46
308	43

ation program, and makes sense chemically. The five best candidates and their rankings are given in Table II.

B. Derive superatoms (7) and constraints from available data. In the example, knowledge of the source of the sample tells us that it is an organic acid from human urine and was esterified to form methyl esters prior to gc/ms analysis. This fraction contains aromatic and aliphatic acids in addition to conjugates of these acids with basic nitrogens, such as in amino acids. We can define a set of superatoms, or building blocks, which can be used to construct structures and can be saved on a computer file for future use in related problems. Such a set of superatoms with their associated names is shown in Figure 3, where the bonds to unspecified atoms are free valences which will subsequently be bonded to other atoms including hydrogen. In the example, the abundant m/e 91 ion suggests the superatom BZ (Figure 3), with no other substituents attached to the aromatic ring. The number of oxygen atoms and degree of unsaturation suggest two methyl ester functionalities (EST, Figure 3). The single nitrogen suggests at least the part structure AMI, arising from an acid conjugate with a basic nitrogen. There are perhaps other ways to phrase this problem, and alternative assumptions, but these assumptions will suffice for illustrative purposes.

C. Generate structures under appropriate constraints from the composition of superatoms and remaining atoms. In our example, the composition is $BZ_1EST_2AIM_1C_3H_5$. Without constraints there are 78 structural possibilities. This list contains many implausible structures. For example, if we assume that the compound is an amino acid conjugate, then a part structure similar to ACI must be present, in fact $-NHCH_{(1-2)}-COOCH_3$. Implementation of this constraint leaves 16 structures.

D. Use MSPRUNE to test the remaining structural candidates to determine which could yield key ions in the observed spectrum. MSPRUNE is an extension to CONGEN which allows interaction with the program to carry out the tests. MSPRUNE operates using the following sequence of steps:

D.1 Obtain fragmentation rules: A series of questions to the user of MSPRUNE/CONGEN elicits the

mass spectrometric fragmentation rules to be used in interpretation of the data. We are currently restricted to rules used previously in the Meta-DENDRAL program INTSUM (9). These rules include constraints on cleavage of aromatic rings, multiple bonds, more than one bond to the same atom, number of steps in a fragmentation process, hydrogen transfers and loss (or transfer) of other neutral species such as water or carbon monoxide. For the example, the constraints summarized in Table III were used. This set of constraints is particularly restrictive. The only danger in this level of restriction is that an incorrect structure may yield a simpler explanation of the spectrum than the correct structure. The correct structure in such a case may be missed.

D.2 Input mass spectral ions to be explained. The user is then asked to input the ions he/she wishes to be explained. These ions may be entered

TABLE III.
Fragmentation Process Constraints Used in the Analysis of the Mass Spectrum Given in Figure 2.

Constraints	User Response
Allow Adjacent Breaks:[a]	No
Allow Aromatic Breaks:	No[b]
Allow Breaks of Double or Triple Bonds:	No
Max Bonds to Break in a Single Step:[c]	1
Max Steps Per Process:	2
Max Bonds to Break in a Process	2
Allowed H Transfers:	-2 -1 0 1 2
Allowed Neutral Transfer:	-

[a] Cleavage of more than one, non-hydrogen bond to the same atom.

[b] I.e., do not cleave the aromatic ring.

[c] There are no aromatic rings or multiple bonds allowed to cleave; any number >1 is meaningless because there are no other degrees of unsaturation (rings); thus every cleavage yields a fragment ion. (9)

as either nominal masses or elemental compositions. Obviously the latter form is much more effective than nominal masses of possible compositional ambiguity. The method of selecting the ions to be used in the analysis is up to the user. In the example, we chose ions on the basis of a) high mass (intuitively of greater structural utility), and, b) high abundance. Ions of low mass and low abundance have a greater chance of resulting from either several different places in the molecule or from complex processes beyond the ability of the simple rules (Table III) to explain.

D.3 Test each candidate structure to determine if it could yield ions input. All possible fragmentations allowed by the constraints are determined for each structure, using an algorithm similar to that developed for INTSUM (9). For each fragmentation the mass and composition of the resulting ion is determined, including allowed hydrogen and/or neutral transfers. A simple comparison of these ions with the ions input reveals whether or not the structure could yield all ions input. If not, the structure is rejected. If so, the structure is retained. This experimental version of MSPRUNE takes no cognizance of ion abundances in this comparison. It makes a simple existence test only. Given the ions of Table I and the constraints of Table II, the list of 16 best candidates is trimmed to five 1 -5, (see Figure 4). If the original set of 78 possibilities is tested under the above conditions for MSPRUNE, 15 structures remain; the spectrum itself is a powerful constraint on possible structures. If only nominal masses are input, the set of 16 structures is not reduced by MSPRUNE; 16 structures remain. This kind of comparison can quantitate the information content of low vs. high resolution spectra.

E. Evaluate Remaining Structures. The structures which result can be examined with the help of CONGEN to determine additional constraints or to design experiments to differentiate among the possibilities. With knowledge of human metabolic processes and the chemistry of the isolation procedure, it is easy to assign 1, (see Figure 4) phenylacetylglutamic acid dimethyl ester (6), as the correct structure. Phenylacetic acid is normally conjugated with glutamine and excreted as phenylacetylglutamine. The base catalyzed hydrolysis converted the primary amide functionality into the observed carboxylic acid (6, see Figure 4): the

dimethyl ester was formed on subsequent derivatization.

In larger problems, it is possible to use other features of CONGEN to test intuitions on possible structures. For example, in this problem where an amino acid conjugate was suspected, it was possible to test automatically every structure for the presence of one of the known amino acid skeletons. Of the five final structures, (1-5, see Figure 4) only one, 1, possesses a known amino acid skeleton. Of the 16 assumed conjugates, four formally possess a glycine, two a phenylalanine, one an aspartic and one 1, a glutamic skeleton. Possible origins of important fragmentations in the spectrum of 1 are illustrated in Scheme 1. We have no isotopic labelling data to support these suggestions.

An Application of MSPRUNE.

The procedure outlined above proved extremely helpful in analysis of unknown compounds observed in a gc low resolution ms experiment. A patient exhibiting signs of mental retardation was referred to Stanford. We examined organic compounds in this patient's urine using procedures described above for the example of structure 1. In fact, these compounds were observed in the same organic acid fraction, but this time prior to any alkaline hydrolysis.

The portion of the total ion current vs. scan number plot where the unknowns were observed is shown in Figure 5. The components in question, A-C, were detected by the program CLEANUP (1b) at scans 382, 402, and 406 (Figure 4). Resolved spectra (1b) are shown in Figures 6a and 6b. The spectra bear obvious similarities; in fact the spectra at scans 402 and 406 are nearly superimposable. Gc/hrms (6) analysis of this fraction lent further evidence to support the relationship of the unknown structures; all significant ions common to the spectra possess the same elemental compositions (shown in Figure 6a for the significant, higher mass ions). The MOLION program finds m/e 207, $C_{11}H_{13}NO_3$, the highest ranking molecular ion candidate for all three components.

The unknowns are apparently structural isomers. Thus, they can be investigated by CONGEN and MSPRUNE in a single run. For this application, we assumed the presence of an aromatic ring, a methyl ester functionality and an amide (conjugate)

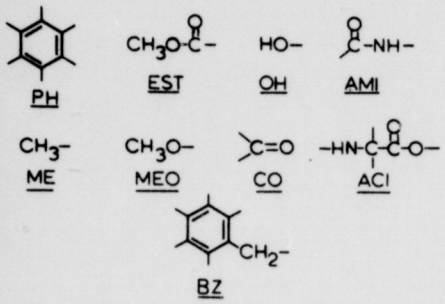


Figure 3. A set of superatoms useful for considering structural candidates in the context of solvent extractable organic acids from human urine (derivatized to methyl esters)

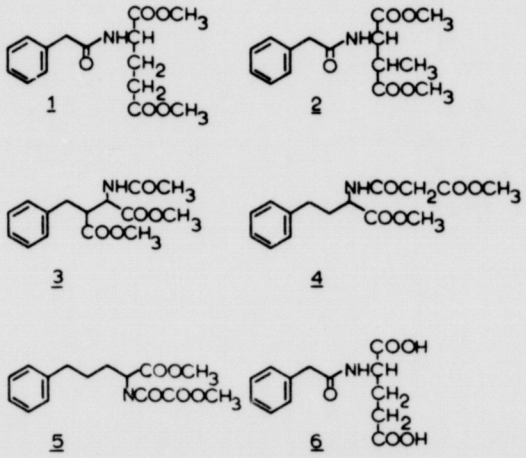
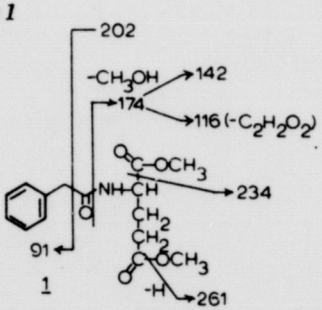


Figure 4. Candidate structures for unknown represented by the mass spectrum in Figure 2

Scheme 1



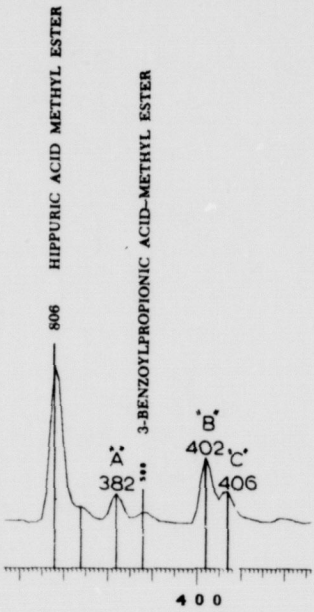


Figure 5. Total ion current plot, scans 357–426, of the solvent extractable organic acids from the urine of a patient exhibiting signs of mental retardation

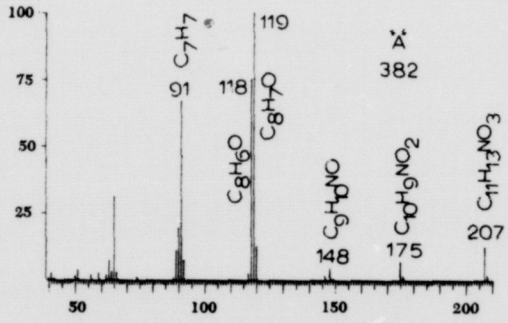


Figure 6a. 70 eV low resolution mass spectrum of component A, scan 382, of the total ion current plot of Figure 5. Elemental compositions were determined by a subsequent GC/hrms experiment (6)

linkage (superatoms PH, EST, and AMI respectively). Constraints forbid direct connection of the amide and ester functions, and alkyl chains of two or more carbons connected to the aromatic ring.

Generation of structures under these constraints yielded 52 possibilities. A number of methods were used to examine these structures. For example, automatic survey of the structural possibilities showed 6 structures which formally possess known amino acid skeletons. Use of MSPRUNE under constraints used for the previous example (Table III), and with the ions shown in Figure 5a was not too helpful; 36 candidates remained after this test. More severe constraints were then used, specifically considering only single rather than two step processes. This effected a dramatic reduction, leaving only four structures, 7 (see Figure 7) and the three isomers represented by 8. A possible source of each ion is shown in Scheme 2 for 8. Literature surveys revealed that 7 has been observed in the dog but never in man. Structure 8 is particularly attractive because there are three substitution isomers possible. These compounds are formally conjugates of toluic acids with glycine. We substantiated this hypothesis and proved the structures by synthesis of the three isomers. The retention indices (Table IV) and mass spectra agree completely. Further investigations indicated that xylenes are excreted by the body by first, oxidation of one of the methyl groups, and second, conjugation with glycine. Further, the relative concentrations of the three compounds closely approximate the relative amount of ortho, meta and para isomers in commercial mixtures of xylene. The patient had somehow been exposed to quantities of xylene.

TABLE IV. Relative Retention Indexes (R.R.I.) for Unknowns A-C and Synthetic Ortho, Meta and Para-toluylglycines, as Determined by CLEANUP (1b).

Unknown	Scan#	R.R.I.	Synthetic Compound	R.R.I.
A	382	2060	ortho-toluylglycine methyl ester	2058
B	402	2128	meta-toluylglycine methyl ester	2128
C	406	2141	para-toluylglycine methyl ester	2143

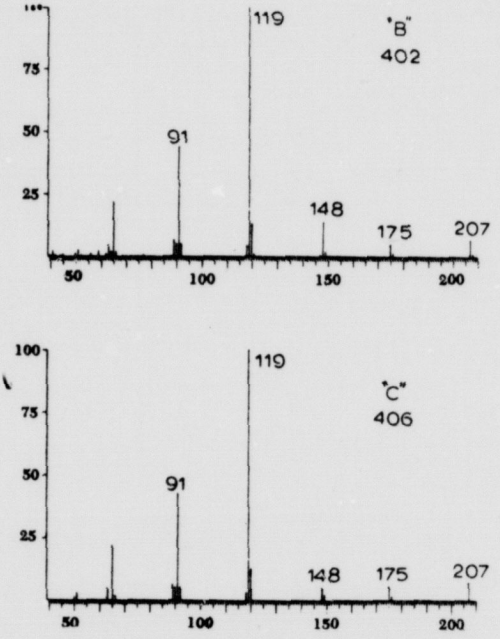
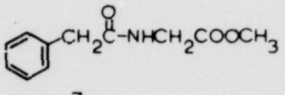
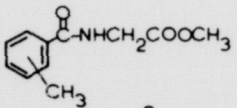


Figure 6b. 70 eV low resolution mass spectra of components B and C (Figure 5). Elemental compositions of major ions are those given in Figure 6a.



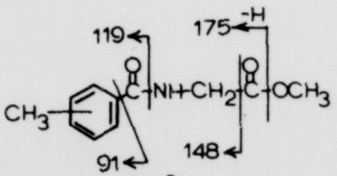
7



8

Figure 7. Candidate structures for unknowns represented by spectra in Figures 6a and 6b

Scheme 2



8

II. Prospective Generation of Structural Possibilities by MDGGEN.

It is a common intuition that mass spectra possess a tremendous amount of structural information, most of which goes unused. Every ion is a piece of the original structure. There are many such ions representing pieces which may overlap to an unknown extent. Ideally, one would like a computational method to perform all possible overlaps of pieces of structure represented by the ions, thereby constructing the set of candidate structures. We are exploring an approach which can in principle carry out this general analysis of mass spectra prospectively, by generation of structural possibilities directly from the mass spectral data. Note that this approach differs from MSPRUNE, which utilizes retrospective testing of previously generated candidate structures.

Our approach introduces the concept of a "mass distribution graph" (an "MDG"). An MDG is a graph whose nodes possess the properties of mass and/or elemental composition. Edges between nodes possess the property of multiplicity, i.e. single, double, etc. MDG's are incompletely specified chemical structures in that nodes of MDG's contain only mass or numbers of atoms without specification of how such mass or atoms are interconnected within each node. In addition, the edges which connect nodes are not precise chemical bonds in that multiple edges may or may not become multiple bonds in complete structures (see below). We have developed a method ("MDGGEN") to construct MDG's, and subsequently, complete structures, directly from the mass spectral data.

We illustrate the use of the concept of MDG's with a simple example, the mass spectrum (Figure 8) and structure of hexanal 9. The spectrum, Figure 8, shows the elemental compositions of ions used in the subsequent discussion. Although MDG's and the section of CONGEN, MDGGEN, which generates them can use nominal masses, results are more precise using elemental compositions, if known (see also MSPRUNE, above).

At the most general level, the MDG for a structure is a single node possessing all atoms and degrees of unsaturation (rings plus multiple bonds), or $C_6H_{12}O$ (one degree of unsaturation) for this

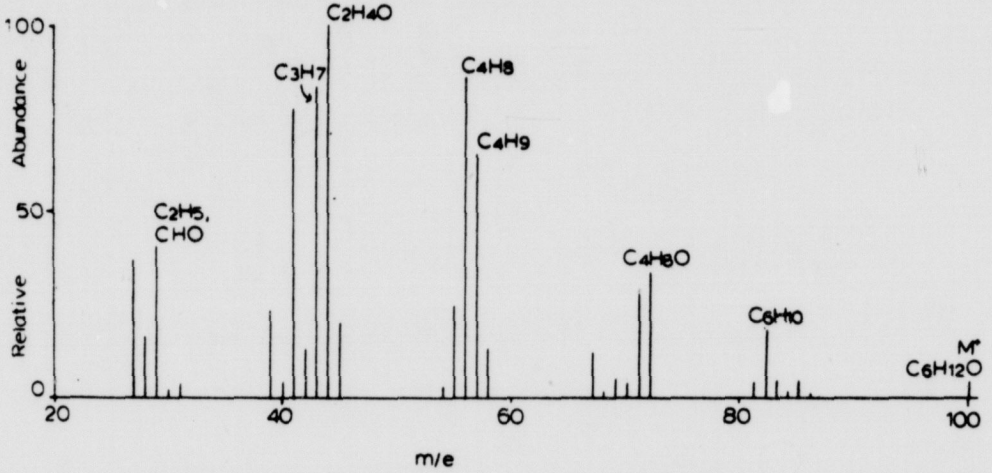


Figure 8. 70 eV low resolution mass spectrum of hexanal. Elemental compositions were determined by accurate mass measurements in a subsequent experiment at high resolving power.

example. MDGGEN requires input similar to that of MSPRUNE. It requires mass spectrometric fragmentation rules. The suite of rules is the same as that mentioned previously (Table III) with an additional parameter for the number of ions which are allowed to go unexplained. It also requires mass spectral data, input by the user, who again selects whatever peaks are deemed relevant.

MDGGEN operates by selecting input ions one at a time. Each ion selected results in an attempt by the program to apportion atoms (or mass) in nodes of the current MDG among nodes of new MDG's such that fragmentation of the new MDG's under input rules would yield the ion. The results for hexanal beginning with the molecular ion MDG ($C_6H_{12}O$) and using the observed ion C_3H_7 as the first selected ion are shown in Figure 9. The number of MDG's resulting after incorporation of a new ion depends on the fragmentation theory in use. Simple bond cleavage of the first MDG (Figure 9) without hydrogen transfer yields C_3H_7 directly. However, if hydrogen transfers (0, 1 or 2) are allowed, the second MDG is also a possibility. Cleavage of the single bond with transfer of two hydrogens into the charged fragment would yield C_3H_7. If two step processes are allowed, together with hydrogen transfers, then there are five other MDG's generated at this level (Figure 9) each of which would yield C_3H_7 with the indicated cleavages and hydrogen transfers.

MDGGEN uses some simple parity considerations to avoid generating nonsense MDG's which have unspecified connections. These considerations are the same as those used by mass spectroscopists. An odd mass ion containing an even number of nitrogens (e.g., 0) can arise from cleavage of an odd number of bonds (e.g., a single bond) accompanied by transfer of an even number of hydrogen atoms (e.g., 0). Or the same ion can arise from cleavage of an even number of bonds (e.g., 2) accompanied by a transfer of an odd number of hydrogens, and so forth. Thus the MDG C_3H_6-C_3H_6O as a single step, one hydrogen transfer process to yield C_3H_7 is nonsense -- only ionic structures or free radicals can be constructed from that MDG. Other constraints are applied to the MDG's if selected by the user, and some examples include forbidding cleavage of more than one bond to the same carbon atom or forbidding cleavage of multiple bonds. Note that the third MDG of Figure 9 cannot be rejected on

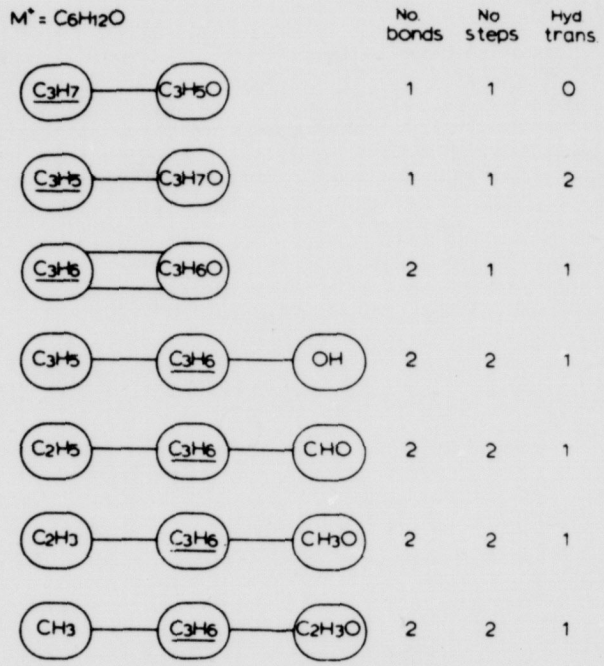


Figure 9. Mass distribution graphs illustrating the various ways of obtaining a C_3H_7 ion from a $C_6H_{12}O$ molecular ion under different assumptions concerning numbers of bonds cleaved, number of steps, and numbers of hydrogen atoms transferred in the fragmentation process

the latter constraint. The two interconnecting bonds may or may not be a double bond in final structures depending on how interconnections are made among atoms in various nodes of the MDG.

Each MDG resulting from application of the first ion is then further elaborated by selecting another ion and asking how, under the existing fragmentation theory, the next ion could be obtained from each MDG. This generally results in expansion of the MDG's into more complex graphs, accompanied however by a greater specificity of each node. This expansion is under the control of an operator which determines legal overlaps of existing MDG's with new MDG's implied by the next ion. An example is shown in Figure 10. The first MDG represents one way of obtaining the observed ion (Figure 8) C_4H_8 from hexanal, by cleavage of two bonds with no accompanying hydrogen transfer. Assume that the next ion applied was C_2H_5, which can be obtained from the MDG C_2H_5-C_4H_7O. There is only one way to perform the overlap (the operator is designated ⊕, Figure 10), yielding the MDG C_2H_5-C_2H_3=C_2H_4O. The elaborated MDG can yield both C_4H_8 and C_2H_5. However, it is structurally more specific because the assembly of atoms C_4H_8 is apportioned into more precise units, C_2H_5-C_2H_3. Assume that the next ion chosen was C_3H_7. One of the possible MDG's for obtaining C_3H_7, cleavage of the two bonds interconnecting C_3H_6=C_3H_6O with one hydrogen transfer, was shown in Figure 9. The overlap operator expands the previous MDG into five more elaborate, but more specific MDG's (Figure 10), from which all ions to that point (C_4H_8, C_2H_5 and C_3H_7) can be obtained. Note that MDG;s 1, 2, 4 and 5 at this level would be eliminated at this point given the constraint forbidding cleavage of more than one bond to the same carbon atom; there is no other way of obtaining the ion C_3H_7 from these MDG's given the assumed origin of C_3H_7 (above).

This procedure (under the constraints of single step processes, forbidding cleavage of multiple bonds, forbidding cleavage of more than one bond to the same carbon atom, transfer of 0 or 1 hydrogen atoms and allowing at most one ion to go unexplained) results in two possible MDG's for hexanal, using the ions labelled with elemental compositions in Figure 8. These are shown in Scheme 3. The first MDG is structurally very specific and yields only the correct structure of hexanal 9. The second MDG indicates a more common

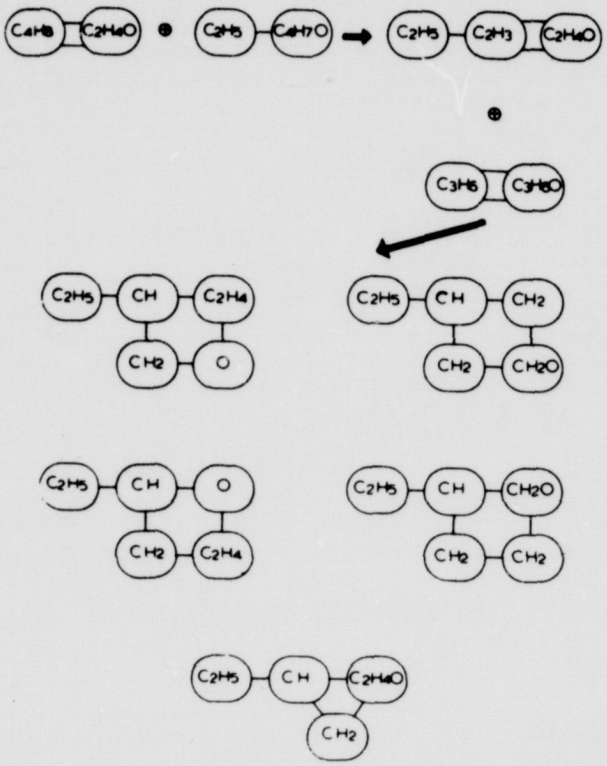
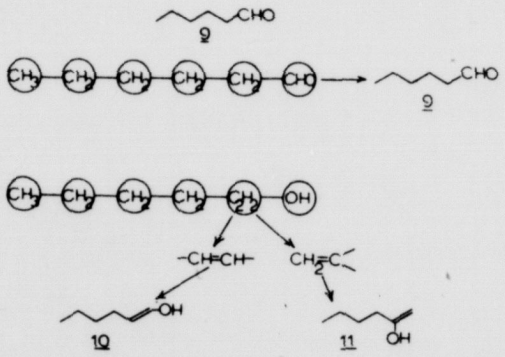


Figure 10. A sequence illustrating the elaboration of MDG's as explanations are determined for new data points (ions of specific elemental composition; see text for details)

Scheme 3



occurrence, where one or more nodes of the MDG can yield different part structures. In this example, the node $-C_2H_2-$ yields either the 1,1- or 1,2-disubstituted C=C, thus producing two alternative structures for the second MDG, 10 and 11, Scheme 3, both enols.

As a developing, experimental procedure, MDGGEN has several limitations. The foremost is the inherent combinatorial complexity of even simple problems. Even with severe constraints there are usually several MDG's resulting from the application of a single ion to each MDG at the previous step. The task of determining all the overlaps for several ions in the spectrum of a molecule of significant size is time-consuming. This is due in large part to our current inability to add chemical intelligence to the procedure. As we discussed above, any problem is simplified if there is a way of representing and utilizing constraints based on knowledge of the problem. We are pursuing this problem in MDGGEN by devising ways to incorporate known superatoms in the procedure. As in the problem of structure generation, problems are simplified when collections of atoms are aggregated into known substructures. Another limitation is the extent of duplication inherent in the procedure. The same structure in many cases can be constructed from two or more MDG's. For example, the structure of hexanal can be obtained from MDG's 1, 5 and 7, Figure 9. This results because MDG's are in a sense only explanations of ions. A given structure may account for a given ion in several different ways. We obtain a separate MDG for each of these explanations, each of which will yield the same structure. Another limitation is in the fragmentation theory itself (see Conclusions). If ions arise from processes more complex than those allowed by the theory, then the correct structure may well be missed or no structures obtained at all. For this reason, we currently allow some number of ions to go unexplained. In the hexanal example, this number was one. Under the theory used, structure 9 cannot lose the elements of H_2O observed in the spectrum, and 10 and 11 cannot yield HCO^+, also observed in the spectrum. But with one ion allowed to go unexplained, 9-11 result.

Conclusions.

We have presented some preliminary results of our efforts toward programs to assist in the general analysis of mass spectra. We have pointed out the need to incorporate as much factual and judgmental knowledge as is reasonable. Only in this way can mass spectral data, in conjunction with a generator of chemical structures, provide significant constraints on structural possibilities. Indeed, we feel that successful systems of the future must make use of such knowledge. Chemists use it in their reasoning about molecular structures and mass spectra; program assistants must do the same.

We are progressing along several lines to make the MSPRUNE and MDGGEN extensions to CONGEN more useful. We are currently investigating ways to make much more detailed statements of fragmentation theory to CONGEN, for use either by MSPRUNE or MDGGEN. We are using a subgraph representation of rules for input by a user (the representation used in Meta-DENDRAL [11]). Such rules may refer to alpha-cleavages, allylic cleavages, etc., or be detailed, class-specific rules in instances where that information about a structure is available. We are extending MSPRUNE to allow investigation of a complete spectrum and a ranking of candidate structures based on the agreement of predicted vs. observed spectra. These are all features of earlier DENDRAL work (5) now brought to bear on chemical problems in the general framework of CONGEN to provide a powerful tool for computer-assisted structure elucidation.

Acknowledgment.

We wish to thank the National Aeronautics and Space Administration (NGR-05-020-004), and the National Institutes of Health (RR 00612 and GM 20832) for their support of this research and for the NIH support of the SUMEX computer facility (RR 00785) on which the CONGEN program is developed, maintained and made available to a nationwide community of users.

Literature Cited.

1. a) Biller, J.E. and Biemann, K., Anal. Lett., (1974), 7, 515; b) Dromey, R.G., Stefik, M.J., Rindfleisch, T.C. and Duffield, A.M., Anal. Chem., (1976), 48, 1368.
2. See review by R.G. Ridley in "Biochemical Applications of Mass Spectrometry", G.R.

Waller, Ed., Wiley-Interscience, New York, NY, (1972), p. 177.
3. Kwok, K.-S., Venkataraghavan, R. and McLafferty, F.W., J. Amer. Chem. Soc., (1973), 95, 4185.
4. Jurs, P.C. and Isenhour, T.L., "Chemical Applications of Pattern Recognition", Wiley-Interscience, New York, NY (1975).
5. Smith, D.H., Buchanan, B.G., Engelmore, R.S., Duffield, A.M., Yeo, A., Feigenbaum, E.A., Lederberg, J. and Djerassi, C., J. Amer. Chem. Soc., (1972), 94, 5962.
6. Gc/hrms data were obtained using a Varian-Mat 711 mass spectrometer connected to a Digital Equipment Corp. PDP 11/45 computer system of our own design. 70 ev mass spectra were obtained at a resolving power of 5000 (extended by software doublet resolution routines) at 8 or 10 sec/decade scans.
7. Carhart, R.E., Smith, D.H., Brown, H. and Djerassi, C., J. Amer. Chem. Soc., (1975), 97, 5755.
8. Carhart, R.E., and Smith, D.H., Computers in Chemistry, 1, 79 (1976).
9. Smith, D.H., Buchanan, B.G., White, W.C., Feigenbaum, E.A., Lederberg, J. and Djerassi, C., Tetrahedron, (1973), 29, 3117.
10. Dromey, R.G., Buchanan, B.G., Smith, D.H., Lederbert, J. and Djerassi, C., J. Org. Chem., (1975), 40, 770.
11. Buchanan, B.G., Smith, D.H., White, W.C., Gritter, R.J., Feigenbaum, E.A., Lederbert, J., and Djerassi, C., J. Amer. Chem. Soc., (1976), 98, 6168.

RECEIVED December 30, 1977