Reprinted from

Advances in

Mass Spectrometry

Volume 7

The Institute of Petroleum

The Application of Artificial Intelligence in the Interpretation of Low-Resolution Mass Spectra

By ARMAND BUCHS, ALLAN B. DELFINO, CARL DJERASSI, A. M. DUFFIELD, B. G. BUCHANAN, E. A. FEIGENBAUM, J. LEDERBERG, GUSTAV SCHROLL,* and G. L. SUTHERLAND

(*Departments of Chemistry, Computer Science, and Genetics, Stanford University, Stanford, California* 94305, *U.S.A.*)

INTRODUCTION

THE application of high-speed digital computers to organic mass spectrometry has become important to many laboratories, especially in connection with problems of data acquisition in high resolution mass spectrometry. The use of computers for the structural identification of organic compounds from their mass spectra by data retrieval is valuable when the solution space is limited to a series of compounds with known mass spectra. In general, this is not the case; hence attempts to use digital computers in the interpretation of unknown mass spectra are of vital importance for any future automatic reduction of experimental data.

We wish to describe one approach to a general computer interpretation of low-resolution mass spectra of organic compounds and to present the results obtained with one special class, aliphatic amines.

THE GENERAL APPROACH

The molecular composition of an unknown compound defines the size of the solution space as the complete set of isomers with that empirical formula. The size of the search space increases drastically with the number of atoms. Whereas C_3H_9N yields four isomers, there exist 14,715,813 isomers with the composition $C_{20}H_{43}N$. The first step in an identification process will therefore be to reduce the solution space to a size where the generation of candidate structures is a feasible process in view of the cost involved in computer time.

The programme (called Heuristic DENDRAL) takes the molecular composition and the mass spectrum as input and returns a list of acceptable candidate structures. Optionally, other physical data (*e.g.* NMR) can be introduced and this will further truncate the list of candidate structures.

Heuristic DENDRAL originally contained five sub-routines,† Preliminary

* Present address: Chemical Laboratory II, University of Copenhagen, The H. C. Ørsted Institute, DK-2100, Copenhagen, Denmark.

† Programme modules are written in upper case.

Inference Maker, Structure Generator, Predictor, Consistency Check, and Scoring Function. For the problem under consideration, aliphatic amines, only the first two sub-routines were used, since all the mass spectrometry theory was placed in the Preliminary Inference Maker and no further heuristics existed for use in other phases of the original programme. The present programme was deliberately designed to achieve maximum truncation of the search space within the Preliminary Inference Maker, since saturated amines yield many more isomers than aliphatic ethers of equal carbon content and it was desirable (in view of the operational time factor) to reduce to a minimum the number of possible solutions presented by the Preliminary Inference Maker.

In order to unambiguously define all possible amine sub-graphs (*i.e.* structural units containing the hetero-atom and having at least one free valence) the symbols T (tertiary), S (secondary), P (primary), and M (methyl) are used to delineate the degree of substitution on the α-carbon atoms of any saturated amine. Thus P refers to

$$-CH_2-NH_2;\ \text{SM to}\ -CH(-)-NHCH_3\ \text{and TS to}\ -\overset{|}{C}H-NH-\overset{|}{\underset{|}{C}}-.$$

The canonical order of the symbols is T > S > P > M and, using this convention, there exist 31 possible amine super-atoms.*

Initially, all 31 possible amine super-atoms are placed on GOODLIST and each is removed when it fails a heuristic decision. All the programme's knowledge of the theory of mass spectrometry and nuclear magnetic resonance spectroscopy is stored within the Preliminary Inference Maker. It should be noted that an NMR spectrum, if available, can be successfully used, but the programme performs in an efficient manner using only mass spectrometric input.

If the empirical composition of an unknown's molecular ion corresponds to $C_nH_{2n+3}N$, the programme locates the amine rules (Fig 1), places the 31 possible superatoms on GOODLIST, then checks to see whether sufficient carbon atoms are contained in the composition to build each of the separate superatoms (Fig 1, decision SIZE).

An NMR spectrum, if available, serves as the next input datum for the programme to scrutinize. Heuristic DENDRAL determines the total number of carbon-bound methyl groups, the number of N-methyl groups, and, if this latter parameter is zero and an integral curve exists, the number of protons attached to the α-carbon atoms of the amine. Those superatoms which are incompatible with the NMR spectrum are then deleted from GOODLIST.

The first mass spectrometric condition (ALPHA CLEAVAGE, Fig 1) programmed into Heuristic DENDRAL is related to the well-known propensity of aliphatic amines to fragment by α-cleavage. For those super-atoms with only one free valence

$$\text{(}i.e.\text{ P, (}-CH_2-NH_2\text{), PM, (}-CH_2-\overset{H}{\underset{|}{N}}CH_3\text{) and PMM, (}-CH_2-N(CH_3)_2\text{)}$$

* Three others, M, MM, and MMM, exist but they translate to the special cases of methyl amine, dimethyl amine, and trimethyl amine respectively.

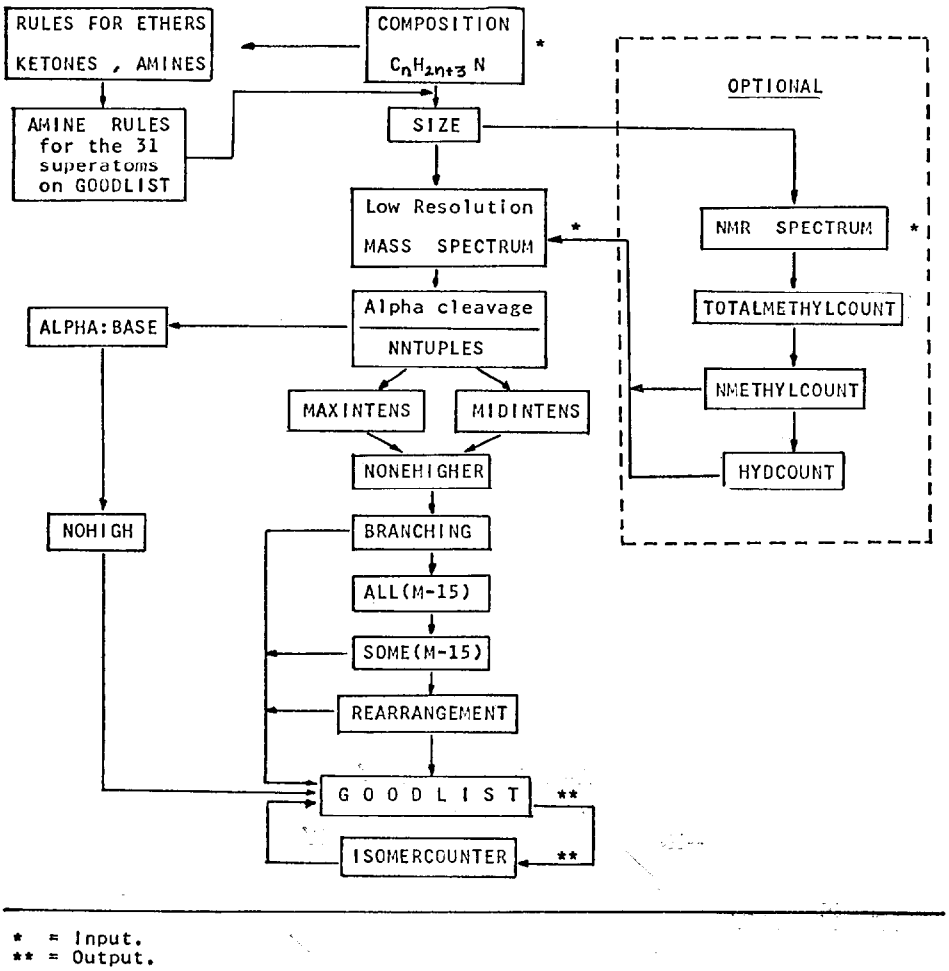


FIG 1

the first condition is that the α-fission peak (m/e 30, 44, and 58, respectively) must be the base peak. A second condition (NOHIGH, Fig 1) states that there should be no other peaks with an intensity higher than 10 per cent relative abundance above the mass value of one half the molecular weight. This latter rule was introduced to take care of special cases which arose when some of the smaller amine molecules were used as examples.

In the case of any other super-atom there is a definite number of α-cleavage fragments which must be located within the unknown mass spectrum (decision NTUPLES, Fig 1). Thus, for every free valence present in a super-atom, the programme has to calculate the mass of an equivalent number of alkyl radicals (referred to as NTUPLES). Certain super-atoms must yield α-cleavage peaks in their mass spectra which exceed an empirically determined value of 70 per cent relative abundance. If more than one set exceeds this limit, the largest intensity

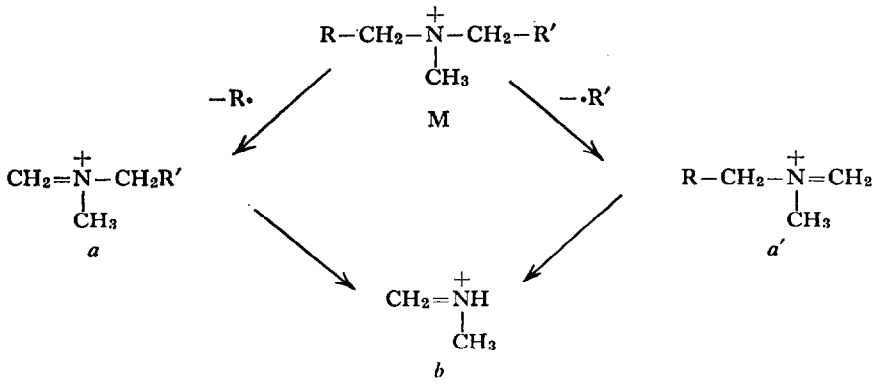
sum is accepted (decision MAXINTENS, Fig 1). Certain super-atoms (those secondary or tertiary α-mono or α-disubstituted amines) must yield an NTUPLE set that exceeds only 30 per cent relative abundance, since these compounds are known to yield very abundant rearrangement ions.

The programme decisions (Fig 1) labelled NONEHIGHER, BRANCHING, ALL (M-15), and SOME (M-15) relate to various facets of the α-cleavage process of aliphatic amines. Super-atoms which either cannot yield rearrangement ions or only ions of moderate intensity are accepted at this stage as viable candidates provided they contain one surviving ntuple.

Those super-atoms which can produce rearrangement ions of major intensity are further tested (decision REARRANGEMENT, Fig 1). They must have at least one intense ion in their mass spectrum originating from the amine rearrangement process (see *a* and *a'*→*b*).

Each candidate super-atom (plus the masses of the alkyl fragments which must be attached to its free valences—termed PARTITIONS in the programme) is sent to a sub-routine (ISOMERCOUNTER) which calculates the number of isomers compatible with that super-atom and its partition list.

Heuristic DENDRAL has been tested with 91 amine mass spectra; in 37 instances these were supplemented by NMR spectra. The programme always included the correct answer in its final output. A tremendous truncation of the hypothesis space (the number of possible isomers) was achieved in most instances. For example, there exist 2,156,010 isomers of tri-*n*-hexylamine and with mass spectrometry alone this figure had been reduced to 240 acceptable candidates. However, with the addition of NMR spectroscopy this list was further reduced to only one entry—the correct answer.



EXPERIMENTAL

The programme described is written in the LISP programming language and runs on the IBM 360/67 computer at the Stanford University Computation Center. Without NMR data, the programme required 4·26 minutes to interpret 91 mass spectra. When NMR data are also used, the process is approximately 30 per cent faster.

ACKNOWLEDGMENTS

Financial assistance from the Advanced Research Projects Agency (Contract SD-183), the National Aeronautics and Space Administration (Grant NGR-05-020-004), the National Institutes of Health (AM 04257), and the allotment of a Fulbright travel award to G.S. is gratefully acknowledged.

Discussion

E. Kendrick (Esso Research Centre, Abingdon, Berkshire, U.K.): You use the programming language "LISP" in the work you have described. Does this language have particular advantages for this type of work and what are these advantages?

B. G. Buchanan: The objects in LISP programmes are the so-called atoms, *i.e.* numbers or strings of letters and numbers. The atoms can be collected in lists and pairs, but the elements of a list or a pair can be lists and pairs as well as atoms. This allows the programmer to operate with complex structures in a much more efficient way than by using the arrays in FORTRAN or ALGOL.

LISP differs in another important way from the ordinary programming languages. In LISP all functions are found in list-structures, and this means that it is possible to write LISP programmes which are writing new LISP programmes. Finally, the possibility of using recursive functions is very helpful in the programming.

G. Schomburg (Max Planck Institute, Mulheim, Germany): Most mass spectra of the components of complex isomer mixtures will be obtained by GC-MS work with high resolution GC (capillary). NMR-spectra are mostly not obtainable from species separated by capillary GC because of low sample load. How do you solve this problem?

B. G. Buchanan: The NMR-spectrum is not required as input of the programme. Hence, there will be no difficulties in applying DENDRAL to GC-MS work. However, when available, the NMR-spectra will lead to a decrease in execution time as well as in the number of candidates.