MASS SPECTRAL LIBRARY SEARCH COMPUTER PROGRAMME +

Branko Ruščić, Leo Klasinc

'Rudjer Bošković' Institute, Zagreb

Jože Marsel

'Jožef Stefan' Institute, Ljubljana

The computer programme AMSIS, a mass spectral storage and retrieval system, is described. It operates on the Aldermaston MSDC library of complete mass spectra. The retrieval of data is achieved through library search. The purpose, the conception and the possibilities of this programme are discussed.

The chemist is often faced with the problem of substance identification. There are handy modern spectroscopic techniques available to him today, yet he wastes every day more and more of his precious time trying to elucidate the outcoming spectrum. Due to the world-wide tendency to use an electronic computer for any lengthy, tedious or boring job, several methods of computer-aided spectra interpretation are developed [1,2]. The classical approach of comparing the unknown spectrum with known spectra can be solved through computerized library search [3], thanks to the commercially available spectra libraries.

The computer programme for the Analysis of Mass Spectra by Information System (AMSIS), written in FORTRAN IV, is a mass spectral storage and retrieval system. Its data base consists of the Aldermaston's MSDC library [4] which comprehends 16,902 mass spectra. This library is available on magnetic tape, in a computer intelligible form. Every peak in the spectrum is coded by two binary numbers: the position, and the intensity. Moreover, there are subsidiary information, such as compound name, molecular weight, compound formula, experimental conditions etc. Every spectrum is given by an unpredictable number of records, which makes the processing troublesome.

+Referat održan na Multifunkcionalnoj konferenciji TEHNIČKI I DRUŠTVENI ASPEKTI INFORMACIJA I DOKUMENTACIJA, Zagreb, 1977-10-17/22

The master file has nearly 270,000 sequential records, or expressed in another way, 130 M bits. This is definitely unsuitable for routine processing. One way to overwhelm the problem is to short up the file [5] but this procedure involves considerable loss of information content. We chose the other way, namely, we split up the master tape into 17 separate files, each containing 1,000 spectra. This is done by the first programme of the preprocessing pack (Fig. 1). The second programme generates the so-called pointer records. Every pointer record corresponds to a singular spectrum, and contains: (i) the molecular weight, (ii) formula of the compound, (iii)

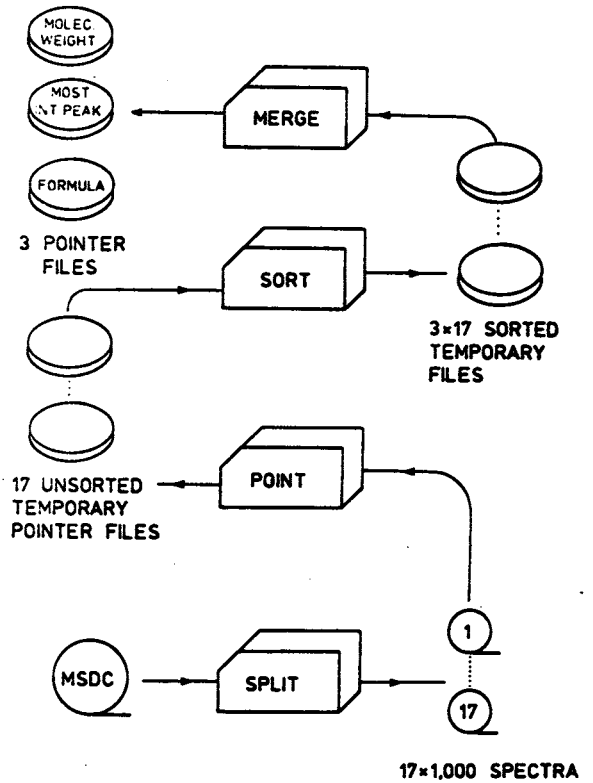

Fig. 1

the position of the most intense peak in the spectrum, and (iv) the ordinal number of the spectrum for identification purpose. The third programme sorts out the records according to each of the keys (i) - (iii). The fourth programme does the MERGE sorting [6], collecting all the pointer records sorted by the same key into one file. After the preprocessing procedure, instead of one huge file, there are 17 reasonably small files and 3 pointer files, thus converting the data base into a handy form.

The programme itself consists of the main programme and a library of subroutines. It can handle several independent logical units, each containing mutually connected tasks. The main programme is of the superviser type, which means that it does not perform any arithmetics at all. It only recognizes the tasks of a very unit and smoothes the way for the execution. The subroutines, which can be divided into 3 groups (i.e. input, output, and processing), are liable for the execution. AMSIS recognizes the tasks by the input keywords. This makes the programme a bit complicated, but for reward, the handling is more comfortable. The main programme decodes the key-words, and memorizes their sequence, It can detect the common user errors, correct them (save for ambiguous cases), and inform the user with the aid of warning messages. After that, the main programme resolves upon which subroutine will be called. The subroutines communicate with the aid of a COMMON zone, with enough core for 2 comlete spectra.

There are 4 different ways to submit the spectrum on punched cards and paper tape, but it can be read also from the library (Fig. 2). The output can be punched on cards or paper tape in 3 different ways, or else it can be printed out. If it is a new compound, then the library can be updated. The main programme recognizes the key-word for plotting, although the subroutine itself is not available yet.

Besides these, there are subroutines for processing the data, namely, a routine for normalizing the intensities to 100 and searching the 10 most intense peaks, a routine for adding and subtracting the spectra, and a set of library search routines. As it is not advisable to compare the unknown spectrum with the whole library, the routines first search through the pointer files, and find out the ordinal numbers of the spectra that confirm with the imposed restrictions, as the molecular weight, the compound formula, the most intense peak in the spectrum, or any combination of them. Only after that, the routines actually spot down these spectra in the

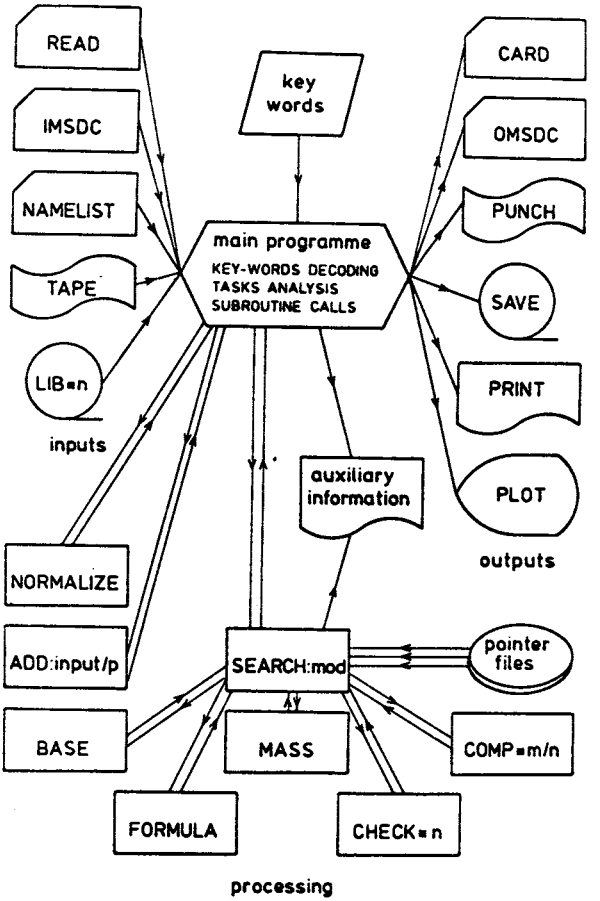
READ
IMSDC
NAMELIST
TAPE
LIB=n
inputs
key words
main programme
KEY-WORDS DECODING
TASKS ANALYSIS
SUBROUTINE CALLS
auxiliary information
NORMALIZE
ADD:input/p
BASE
SEARCH:mod
MASS
FORMULA
CHECK=n
pointer files
COMP=m/n
processing
CARD
OMSDC
PUNCH
SAVE
PRINT
PLOT
outputs

Fig. 2

data files, and proceed on narrowing the proposed compounds list. The routines can check the existence of a feature, by testing if there are any peaks at some predetermined positions in the spectrum. At the end, the routine may compare the abbreviated spectra [5]. It means, that the routine selects the m most intense peaks every n mass units, both in the unknown and in the proposed spectrum, and looks for a match. In the case when positive integer parameters m and n are chosen as 1, the comparison of the complete, unabridged spectra will be done.

An actual test example performed with a data base of 1,000 spectra should help to explain how the programme works. One spectrum was randomly chosen from the first file, and punched out. In another run, this spectrum was submitted as an unknown.The imposed restrictions were: molecular weight 122, formula $C_7H_6O_2$ the most intense peak 43, and after that an abbreviated spectra (1 peak every 14 mass units) comparison was proposed. The programme searched through the pointer file and found 21 compounds with the correct molecular weight. The list was narrowed to 3 compounds taking into account the compound formula. There was only one compound left when the programme applied the next restriction, and, finally, checking the abbreviated spectra confirmed that it was 2-furaneacrolein.

The programme design allows for change and growth required for the future. Although the programme is concieved for batch processing, the conversion to demand processing is a simple task.

REFERENCES

1. A.B. Delfino and A. Buchs, Computers in Chemistry, Topic in Current Chemistry 39 (1973) 109.
2. P.C. Jurs and T.L. Isenhour, Chemical Applications of Pattern Recognition, Wiley, New York, 1975.
3. F.A. Mellon, Mass Spectrometry 3 (1975) 117.
4. MSDC Full Mass Spectra Collection, Mass Spectrometry Data Centre at AWRE, Aldermaston, Reading, U.K.
5. S.R. Heller, Anal. Chem. 44 (1972) 1951.
6. T.F. Fry, Computer Appreciation, Butterworth and Co., London, 1972.