Journal of Chromatography, 318 (1985) 247–253
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

CHROM. 17,251

USE OF HISTOGRAMS IN COMPUTER-AIDED COMPARISON OF CHROMATOGRAMS

GABOR VARHEGYI* and GABOR ALEXANDER
Hungarian Academy of Sciences, Research Laboratory for Inorganic Chemistry, Budaörsi ut 45, H-1112 Budapest (Hungary)
(First received June 6th, 1984: revised manuscript received September 20th, 1984)

SUMMARY

The computer evaluation of chromatograms measured on any large series of samples is discussed. The computer processing of the empirical frequency distribution of the peak retention times has proved useful for obtaining an easy survey and optimal tabulation of huge data sets. The method can also be used to pre-process chromatographic data before factor analysis or pattern recognition techniques.

INTRODUCTION

It is a frequent task to evaluate the chromatograms of a series of samples in order to classify them or to find regularities in the data set. If the number of the samples is high (*e.g.*, higher than 40) and a large number of peaks have to be taken into account, computer data processing is required. From a computational point of view, the type of the sample series is irrelevant: it may be a series of drinking water extracts, coal pyrograms, human blood, etc.

The simplest data processing is the tabulation of the occurrence of the various peaks on the chromatograms. An equally simple approach is the redrawing of the chromatograms into forms easy to survey by the human eye[1]. More advanced techniques include factor analysis and non-linear mapping[2,3] and also pattern recognition techniques[4–9]. For the application of the above-listed methods, one should decide which peaks may correspond to the same chemical components in the different chromatograms. In other words, one has to decide which peaks may occur in the same column of a table or which peaks belong to the same feature in a pattern recognition process. Usually, this classification of the peaks is made manually, through visual inspection of the chromatograms. If the chromatograms are too complex or the number of the data to be processed is too high, the manual approach may become tiresome and not sufficiently reliable. In such instances, computer pre-processing is required.

A limited number of chromatograms can be surveyed directly by computer if advanced computer graphics are available. Excellent graphic reports were presented by Janssens and Beernaert[10], showing eight peaks on nine chromatograms. This

0021-9673/85/$03.30

graphical approach, however, is not applicable if there are 40 or more chromatograms with hundreds of irregularly spaced peaks. A simple but questionable peak classification method was applied by Küllik *et al.*[8] in the pattern recognitions of pyrolysis gas chromatograms. They divided the retention time domain into 39 zones and chose the highest peak in each zone of each pyrogram. The width of the zones, however, was far wider than the random scattering of the retention times, so there was no guarantee of the chemical identity of the peaks chosen from a given zone of the different chromatograms. A more reliable algorithm was presented by Mayfield and Bertsch[2]. They generated a "reference list" by processing the chromatograms one by one and putting into the reference list the value of any retention time of a peak differing by at least a certain limit from the retention times already in the reference list. By this recursive method all discernible retention time values are collected. After the generation of the reference list they select the retention times that occur in the data set frequently enough for the further data processing. This method is correct and elegant and its only drawback seems to be that it is too mechanistic. In our opinion, it would be better to consider directly the distribution of the values of retention times in the full data set so that the selection mechanism would be based on physically meaningful steps easy to survey (and modify if necessary) by the chromatographer. An algorithm of this type is the subject of this present paper.

EXPERIMENTAL

In the calculations a PDP-11/34 computer was used, with a 64 kbyte memory and two single-density floppy disks under operation system RT-11. The program was written in Fortran IV. The special inpout/outpout facilities permitted by the PDP-11 Fortran were used extensively. A copy of the program can be obtained free of charge from the authors.

The data used in the calculations were obtained from the complex multi-peak chromatograms of drinking water extracts, received as part of an environment protection project. The aim of the chromatographic measurements was the study of the organic impurities in drinking water samples from various sources.

The gas chromatographic measurements were performed on a Hewlett-Packard 5880 A Level Four gas chromatograph. The instrumental and analytical parameters were as follows: column, 20 m × 0.25 mm I.D. glass capillary column coated with OV-1 methylsiloxane gum, phase ratio 300; oven, 80°C for 2 min, programmed at 4°C/min up to 220°C, then isothermal; injection, split injection, splitting ratio 1:15; detector, flame-ionization; integrator, built-in, with adjustable area reject (smallest peak area) value. The output of the instrument was a report table consisting of retention time–peak area data pairs. These data formed the input of the calculations. The data transfer to the computer was manual.

FUNDAMENTAL CONSIDERATIONS

The algorithm to be discussed is based on the following plausible considerations. The basic problem is to find the peaks of identical substances on the different chromatograms in spite of the random scatter of the retention time values. This random scatter can be studied using the empirical frequency distribution function of

the values of the retention times of the peaks. In other words, a histogram is constructed where the abscissa is the retention time scale divided into small intervals and the ordinate indicates how many chromatograms contain peaks in each of these retention time intervals. If a certain group of retention times in the data set belong to the same chemical component, their differences are caused only by random errors, so a regular peak (probably a peak with a roughly Gaussian shape) will appear on the histogram. By the computer determination and processing of these histogram peaks, we can collect those retention times which may be regarded as identical. The histogram also reveals the occurrence of the chemical components with retention times close to each other (here the term "close to" means differences comparable to the deviation of the random scattering). These chemical components are represented on the histogram by double peaks or by peaks with strongly distorted shapes.

THE STEPS OF THE ALGORITHM

A schematic flow chart of the whole procedure is shown in Fig. 1. The description of the individual steps is as follows.

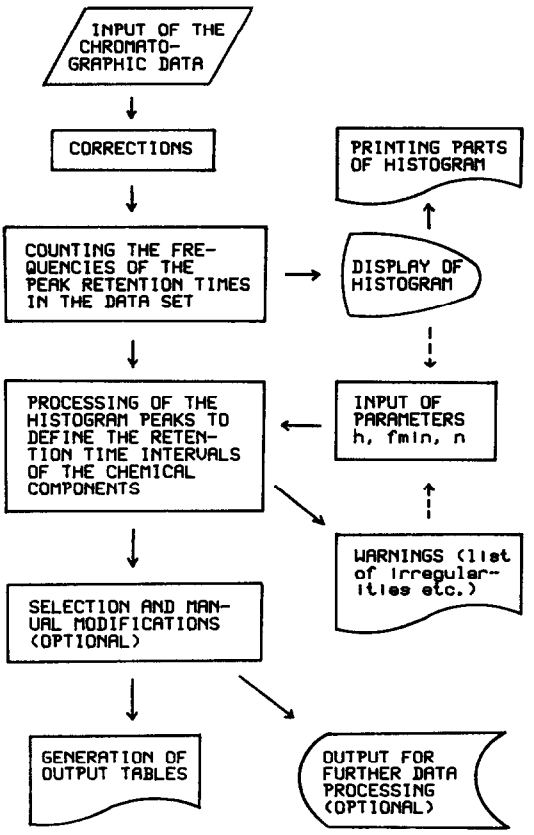

Fig. 1. Schematic flow chart of the procedure. The broken arrows indicate user interactions; depending on the display and the warnings, the user defines or redefines parameters.

Input and corrections

The first step is the input of the retention times and peak areas of the chromatograms. An adjustable area reject parameter, which may have values differing from the area reject parameter of the gas chromatographic procedure, is used to neglect non-significant small peaks. The retention times are given with a precision of (at least) 0.01 min (in the calculations, they are stored in 16-bit words with a precision of 0.005 min). The systematic errors of the retention time data may be compensated by "shift terms". For the determination of these shift terms an internal standard with a known retention time has to be measured on every chromatogram. If more than one internal standard is measured on the chromatograms, one of the methods listed by Mayfield and Bertsch[2] should be included in the program. The chromatographic data are stored on an input file. A given input file can be processed several times, if necessary, to obtain more or less compact tables, etc. The parameters defining the data processing are given on the terminal at those points of the processing where they are actually needed.

Histogram generation

In the second step, the retention time domain will be divided into small intervals with a length of 0.01 min and the number of the chromatograms that contain peaks in the individual intervals will be counted. The histogram obtained in this way is stored in an array of 10,000 16-bit words during the computation.

Display

It is worth displaying on the videoterminal or printing on the line printer the significant parts of the histogram (here the term "significant parts" refers to the vicinity of every retention time with a higher frequency than an input parameter f'_{min}). In this way an overview can be obtained of the data set. After the data processing (see below) it may be useful to recall the histogram peaks again and to check whether the peak retention times have a regular scatter around their means or more than one chemical components have fitted into a single feature. Note that no graphical capabilities are needed for the display of histograms. Details of a computer-printed histogram are shown in Fig. 2.

Histogram processing

The program looks for the peaks of the histogram. To reduce the errors arising from the discrete nature of the histogram, the positions of the peak maxima are refined by the centroid method[11]. The results are rounded to a precision of 0.005 min (higher precisions do not seem meaningful). The retention times obtained in this way will be regarded as the centres around which the retention times of the individual chemical compounds scatter. The scatter is supposed to be in a certain interval of $\pm h$, where h is an input parameter provided by the user. A suitable value of h can simply be found by the visual inspection of the well separated peaks on the displayed histogram. Problems arise when two or more histogram peaks strongly overlap. This indicates that the retention times of different chemical compounds are too close to each other. When the program finds such a situation, it gives a warning. If one of the overlapping peaks is considerably higher than the other(s), only this high peak will be selected.

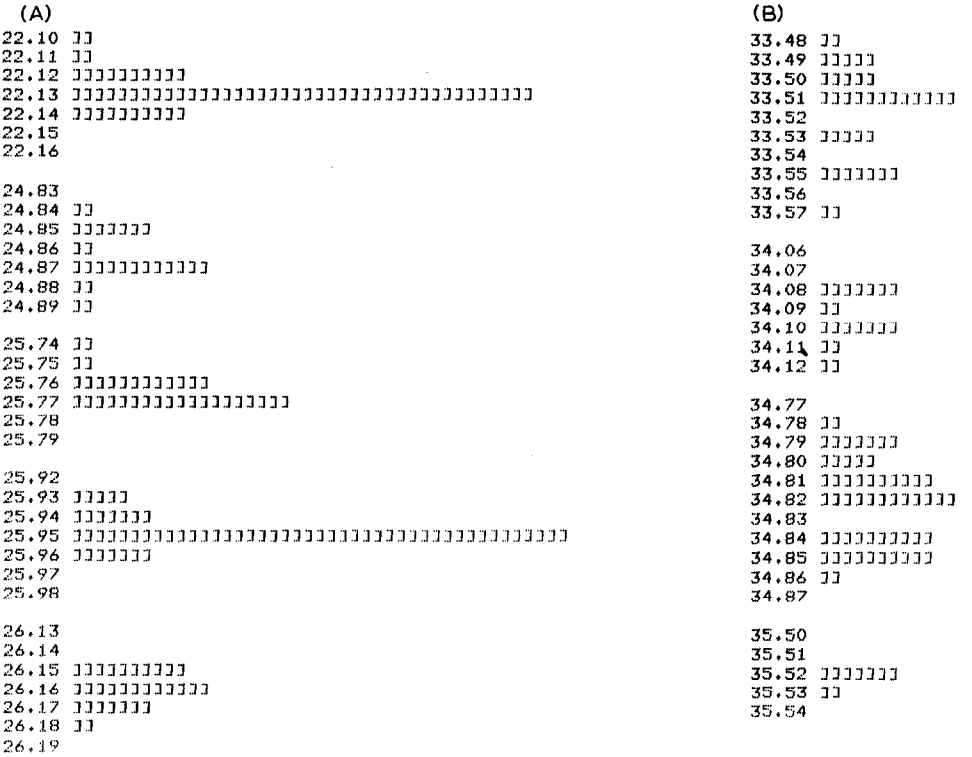

Fig. 2. Details of a histogram showing the frequencies of the peak retention times in a series of chromatograms. The numbers printed on the figure are retention times in minutes. (A) Typical details with regular histogram peaks; (B) contains less regular peaks also. The computer-assigned peak positions in B are listed in the headline of Table I.

If the overlapping peaks have roughly the same magnitude, the program merges them and assigns a single interval containing both overlapping peaks. In this instance chemically inhomogeneous features are obtained. However, there is no other choice in such situations and it is better to use chemically inhomogeneous features than to lose information. On a histogram, "plateaux" and "hillsides" may also occur. If they are high (*i.e.*, if many histograms contain peaks in those regions), it is worth selecting interval centres there also, to avoid a loss of information. The computer processing of the histogram is considerably speeded up if the parts containing only frequencies lower than a certain limit f_{min} are skipped (f_{min} is an input parameter).

Selection

The number of the defined features may be too high. To reduce the size of the output, the next step is an optional selection of the n most populous intervals (n is an input parameter). There is the possibility in the program of manual deletion, modification and input of intervals, also. In this way the user can override the decisions of the program.

Tabulation

For many tasks, a proper tabulation of the data is sufficient to find the regularities in the data set. For this purpose, tables are printed by the program. The columns of the tables correspond to the selected features, the headline contains the interval centres and the rows belong to the individual samples. The listed data are the peak areas from the individual chromatograms multiplied by suitable scale factors. An example is shown in Table I. It is necessary to print warnings in instances

TABLE I

IMPURITIES IN DRINKING WATER SAMPLES

The retention times in the headline were obtained by the computer processing of the histogram section shown in Fig. 2B. The data in the table correspond to the areas of those chromatographic peaks which have appeared in the vicinities of ±0.015 min of the retention time values given in the headline. The rows of the table belong to individual samples. The first characters in the rows are sample codes. The area data have been converted into concentration units (ppb). This table is a detail of a larger table (see text).

RETENTION:	33.505	33.550	34.080	34.100	34.790	34.815	34.845	35.520
Population:	(8)	(3)	(4)	(5)	(6)	(11)	(9)	(4)
B7				0.8				
B10				0.2				0.4
CSE1						0.7		
CTO1								
CTO2								0.4
DGM					1.3 =	1.3		
D1							6.6	
D3					1.1			
D4					0.3			
D5	3.0					1.0		
D6						2.7		
D7								
D8A	0.5						2.1	
D8B	0.1					1.5		
D9	0.4	0.1				0.8		
D10A								
D10B								
E80A								
E80B	0.3							
FORSZ						0.5		
G2A			3.2				0.4	
G2B			9.6				0.4	
G3								
G5								
K2A							7.0	
K2B				0.1			3.9	
K3A								
K3B								
K4	168.3							
K6							3.3	
K7								
K17A								
K17B			0.1 =	0.1				
SUR	0.4					0.5		1.5
S2A			1.1			0.5		
S2B	2.2	0.8			3.2 =	3.2		
VK0					0.2			
VK1							4.0	
VK2					0.1			
VK3						0.9		
VK53				0.7			14.1	0.4
VK55		138.6						

of overlapping features. In such instances the program indicates the areas of the not uniquely classified peaks in both neighbouring columns and prints a warning character (=) between the identical values.

Output for further data processing

The selected area data of the chromatograms can be output on a disk file in the form of vectors for further data processing by other programs (factor analysis, pattern recognition, etc.)

Final warnings

Any data processing based on the frequencies of the data obviously misses those chromatographic peaks which occur in only a few chromatograms but are still important because of their dominating area. As visual inspection of the chromatograms and the tables does not necessarily reveal the occurrence of this problem, the program lists these peaks to the chromatographer for further, non-mathematical consideration.

APPLICATIONS

The program described here was used successfully for processing a series of gas chromatograms of drinking water extracts in an environmental protection project. The examples shown (Fig. 2A and B and Table I) are taken from this work. Table I is a detail from a larger table consisting of five pages with fifteen columns on each page. The publication of the water analysis itself is outside the scope of this paper.

REFERENCES

1 J. E. Buchanan and T. P. Maher, *J. Chromatogr. Sci.*, 9 (1971) 448.
2 H. T. Mayfield and W. Bertsch, *Comput. Anal. Lab.*, 2 (1983) 130.
3 F. S. Hsu, B. W. Good, M. E. Parrish and T. D. Crews, *J. High Resolut. Chromatogr. Chromatogr. Commun.*, 5 (1982) 648.
4 M. L. McConnell, G. Rhodes, U. Watson and M. Novotný, *J. Chromatogr.*, 162 (1979) 495.
5 A. Zlatkis, K. Y. Lee, C. F. Poole and G. Holzer, *J. Chromatogr.*, 163 (1979) 125.
6 E. Jellum, I. Bjørnson, R. Nesbakken, E. Johannson and S. Wold, *J. Chromatogr.*, 217 (1981) 231.
7 R. E. Lea, R. Bramston-Cook and J. Tschida, *Anal. Chem.*, 55 (1983) 626.
8 E. Küllik, M. Kaljurand and M. Koel, *J. Chromatogr.*, 112 (1975) 297.
9 G. F. Gostecnik and A. Zlatkis, *J. Chromatogr.*, 106 (1975) 73.
10 G. Janssens and H. Beernaert, *J. High Resolut. Chromatogr. Chromatogr. Commun.*, 3 (1980) 326.
11 J. R. Chapman, *Computers in Mass Spectrometry*, Academic Press, London, 1978.