Journal of Chromatography, 585 (1991) 160–163
Elsevier Science Publishers B.V., Amsterdam

CHROM. 23 570

Short Communication

Data compression in computerized signal processing for isocratic chromatography

J. C. Reijenga

Laboratory of Instrumental Analysis, University of Technology, 5600 MB Eindhoven (Netherlands)

(First received February 22nd, 1991; revised manuscript received June 17th, 1991)

ABSTRACT

A data compression technique for isocratic liquid or isothermal gas chromatography is introduced that effectively stores the information contained in the chromatogram in a reduced-size array. The resulting compressed peaks have an approximately uniform width, which is an obvious advantage for the peak detection and integration algorithm. For identification, the apparent retention times in the compressed array are used, the reproducibility of which does not deteriorate with respect to the original signal. Some improvement of the signal-to-noise ratio is seen.

INTRODUCTION

The computerized signal acquisition of Gaussian-shaped signals originating from chromatographic detection generally requires choosing an optimum sampling rate (number of data points per second). The number of data points per σ of the Gaussian peak shows such an optimum rate of approximately 5, assuring an adequate description of the Gaussian profile. For non-Gaussian or tailing peaks, an adequate description of the peak requires a larger number of data points [1,2].

When too large a sampling frequency is chosen, over-sampling occurs. The disadvantages of over-sampling are: the inefficient use of computer memory, increased sensitivity to high-frequency disturbances (resulting in spikes) and possibly a loss of speed in the integration algorithm. Under-sampling leads to the loss of information with respect to identification (retention time) and quantitation (peak area). Most data acquisition hardware uses a constant sampling rate, matched by the rule of thumb outlined above. In isocratic liquid or isothermal gas chromatography, however, where the peaks become broader as retention increases, this rapidly leads to an over-sampling of the signal at increasing retention times.

This paper describes a method of data compression for sampled chromatographic signals that adapts the apparent sampling rate to minimize the amount of memory needed to store the information contained in the chromatogram.

DATA COMPRESSION

Consider a sampled isocratic chromatogram contained in an array of n points. The initial sampling rate is adjusted so that the standard deviation of the earliest peak, σ_1, is sampled by at least p points. The σ values of subsequent peaks are larger than σ_1 and can be calculated from the plate number, N, of those peaks.

The plate number is first assumed constant over the k' range considered and is equal to:

$$N = (i/\sigma_i)^2 \quad (1)$$

in which σ_i is the standard deviation at retention time i, both in dimensionless units of data points. Data compression now compresses each data point by a reduction factor k_i. This reduction factor k_i is unity at the beginning of the chromatogram and gradually increases with increasing retention time. The standard deviation σ should be sampled p times after data compression, so that:

$$k_i = \sigma_i/p \quad (2)$$

with the restriction that k is not less than unity. With this restriction, the combination of eqns. 1 and 2 yields:

$$k_i = 1 + i/pN^{1/2} \quad (3)$$

In the compressed array, several data points are combined, with $1/k_i$ as a weighing factor, into one new point. If n is the number of points in the original array, then the number of points in the reduced array, m, can be calculated by summation of $1/k_i$ over all i values:

$$m = \sum_{i=1}^{n} 1/k_i = \sum_{i=1}^{n} pN^{1/2}/(pN^{1/2} + i) \quad (4)$$

The implications of eqn. 4 are that much is to be gained from data compression at low plate numbers, where peaks become broader at relatively short retention times. Not much is to be gained at very high plate numbers. The break-even value of n/m is unity.

An important factor is the number of data points per σ required. A number of rules of thumb are in use, ranging from 1 to 10 [3–5], depending on the aim of data acquisition (*e.g.* determination of retention, peak area, peak moments, exact reconstruction). For normal integration purposes $p = 5$ is acceptable.

In Fig. 1, m is plotted *versus* n on log scales for different plate number N and $p = 5$ points per σ. The values were calculated by summation of eqn. 4. From this figure it can be concluded that, for a given value of p, the net gain depends on the plate number (N) and on the size of the original data array (n). Peak-to-peak differences in plate numbers, especially in high-performance liquid chromatography, are due to a number of effects which are beyond the scope of this paper [1,6]. The effect of these fluctuations is limited by the square-root term in eqns. 3 and 4.

When compressing such a chromatogram, there are two options: (1) to be safe, a larger value of p can be used; and (2) a better strategy is to take a plate number slightly higher than the highest value

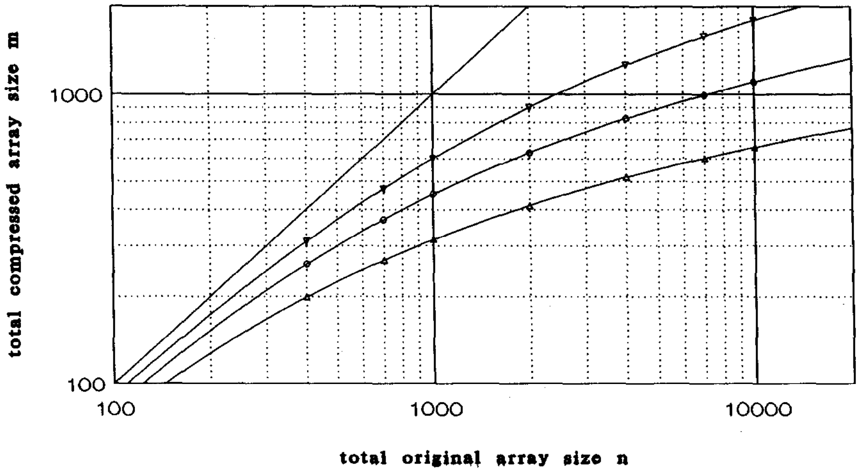

Fig. 1. Size op the reduced data array (m) plotted *versus* the size of the original data array (n) at different plate numbers, using eqn. 4. Peaks are compressed to $p = 5$ data points per σ. The staight line represents the break-even point in terms of array size, the other curves are fourth-order polynomal fits to the points indicated: (△) $N = 1000$; (○) $N = 4000$; (▽) $N = 16\,000$.

expected from different peaks. In this instance the signal can never be over-compressed.

Real-time or post-run compression

The compression algorithm is executed fairly fast compared with the corresponding analysis time. Consequently, a choice can be made between real-time and post-run compression. If the data acquisition rate is sufficiently low not to be limited by the conversion rate of the hardware used, real-time compression can be considered. The advantage of this is that full use can be made of the reduction of the memory required.

Retention data after compression

When comparing the chromatograms, compressed in an identical manner, the reproducibility of the apparent retention times is not deteriorated with respect to the original retention times. This can be explained by the non-linear relationship between the original and compressed time axes (Fig. 1).

If, in addition, values for k' or real retention times are required, the additional information on the compression factor, stored in a separate array, is necessary. To obtain a real retention time, values in this array are accumulated up to the corresponding apparent retention time (in units of data points). However, exact reconstruction of the original raw data is not possible because of the averaging performed by the compression algorithm.

Signal amplitude after compression

An additional advantage of this data compression technique is that the signal-to-noise ratio is improved when $k_i > 1$.

Compression decreases the amplitude of the random white noise. A net gain of $k_i^{1/2}$ for the signal-to-noise results [2,7], which is especially of importance at longer retention times.

Temperature-programmed or gradient elution chromatography

In temperature-programmed gas chromatography or gradient elution liquid chromatography, a dynamic change in retention behaviour generally results in chromatograms in which the peak width is less dependent on the retention time [1]. Most often, the highest plate numbers are calculated from peaks with long retention times.

If the sampling rate of the initial data acquisition is adjusted so that the first peak is sampled by, for instance, 5 points per σ (and N is calculated from the last peak), compression offers few advantages, but the application of the compression algorithm does not lead to a loss of information.

RESULTS AND DISCUSSION

Computer simulations were carried out using an ordinary XT-type personal computer with an 8087 co-processor. Any high-level programming language, supporting the use of a co-processor and the available graphics display, can be applied. Quickbasic 4.5 (Microsoft) was used in this work; the source of the algorithm is available from the author on request.

Fig. 2a shows a simulated chromatogram of 5000 data points with 2500 plates. Because of the limitation of the resolution of the graphics screen (EGA or CGA), the representation is limited to each 16th point. After compression with p = 5 points per σ,

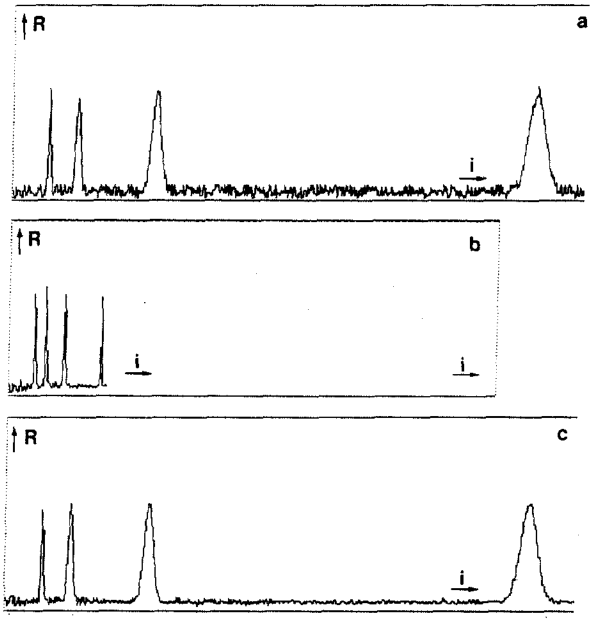

Fig. 2. Simulated chromatogram of 5000 data points (a) reduced to 782 points when compressed to p = 5 data points per σ with a plate number of 2500 (b). After decompression, some data reduction remains (c). See text for further details. The axes in this figure represent arbitrary response (R) and time (i).

782 data points are sufficient (Fig. 2b). This number is reduced with the same factor 16 to illustrate the factor 6.4 gain.

The plate number, needed for compression, is not critical. Column deterioration, leading to the loss of plate numbers, will therefore not be a problem. Column deterioration will usually also lead to change of retention time. The sensitivity of the compression algorithm for slight changes in retention times was checked by a series of chromatograms, produced by a gas chromatography simulator [8]. The standard deviation of the apparent retention times in the compressed chromatogram was always smaller than that in the original signal.

CONCLUSIONS

The data compression algorithm effectively stores the information contained in the chromatogram. The resulting chromatogram with peaks of uniform width is treated as a normal chromatogram. The standard deviation of retention does not deteriorate with respect to the original signal. Some noise reduction has taken place, while at the same time the effect of baseline drift is apparently amplified. The area of the peaks is decreased by a factor k but the signal-to-noise ratio is increased. When the plate number is not the same for all peaks, compression with the highest plate number expected is advised. This will not lead to a loss of information, only to a slight variation of the peak widths in the compressed signal.

When restoring the original chromatogram using the array with the compressions factors, some reduction of data is inevitable, as seen by the lower noise level due to averaging. The original retention data is now obtained (Fig. 2c).

REFERENCES

1 E. Heftmann (Editor) *Chromatography, Part A* (*J. Chromatogr. Library*, Vol. 22A), Elsevier, Amsterdam, 1983.
2 P. A. Leclercq, in J. Novak, *Quantitative Analysis by Gas Chromatography* (*Chromatographic Science Series*, Vol. 41), Marcel Dekker, New York, 1988, p. 163.
3 K. L. Rowlen, K. A. Duell, J. P. Avery and J. W. Birks, *Anal. Chem.*, 61 (1989) 2624.
4 G. Guiochon and M. J. Sepaniak, *Anal. Chem.*, 63 (1991) 73.
5 K. A. Duell, J. P. Avery, K. L. Rowlen and J. W. Birks, *Anal. Chem.*, 63 (1991) 73.
6 P. A. Bristow, *LC in Practice*, HETP, Handforth, 1976.
7 D. L. Massart, B. G. M. Vandeginste, S. N. Deming, Y. Michotte and L. Kaufman, *Chemometrics, A Textbook*, Elsevier, Amsterdam, 1988.
8 J. C. Reijenga, *J. Chromatogr.*, in press.