Journal of Chromatography, 462 (1989) 13–21
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

CHROM. 20 990

AN INFORMATION THEORY OF CHROMATOGRAPHY

I. EVALUATION OF ANALYTICAL SYSTEMS BY MEANS OF *FUMI*

RIEKO MATSUDA, YUZURU HAYASHI*, MUMIO ISHIBASHI and YASUSHI TAKEDA
Division of Drugs, National Institute of Hygienic Sciences, 18-1, Kamiyoga 1-Chome, Setagaya-ku, Tokyo 158 (Japan)

(First received April 7th, 1988; revised manuscript received September 19th, 1988)

SUMMARY

The application of the *fu*nction of *m*utual *i*nformation (*FUMI*) to the logical evaluation of methods for chromatographic quantitation is described. *FUMI* provides the Shannon mutual information of overlapped peaks with various resolutions. Hence an analytical method can be evaluated logically based on both the amount of information transmitted by overlapped peaks and the observation period of chromatography. As an example of the evaluation, a previously proposed chromatographic analysis consisting in a rapid but incomplete separation of naphthalene and diphenyl, and peak-deconvolution based on the Kalman filter, is considered. The "best" chromatogram that can transmit the maximal mutual information in unit time is given.

INTRODUCTION

A major aim in analytical chemistry is to elaborate or search for a method through which more information can be effectively transmitted from analytes (Ω) of interest. For high-performance liquid chromatography (HPLC), efforts are made towards the development of column structure, elution conditions, etc., and also data processing of signals. The development of methods often depends on trial and error, and may result from chance. Such new methods are evaluated quantitatively, even if empirically, by some chemometric strategies. A more desirable aim, however, is strict evaluation on a theoretical basis, without recourse to experience.

The aid of information theory is desirable for the above-mentioned purpose. The information on analytes Ω is quantitatively described as the Shannon information $I[\Omega]$ (ref. 1). In general, it is through observation and data processing of the raw data Ψ that we can actualy acquire knowledge about the analytes Ω. The total amount of information $I[\Omega]$ involved originally in the samples, however, cannot be obtained from the data Ψ because of inevitable noise contamination, baseline drift, interference, etc., in the measurement process. Of more analytical importance is the available information called the mutual information $I[\Omega; \Psi]$ between the analytes Ω and the data Ψ (ref.

1). Excess peak overlap, for example, which often arises in rapid chromatography, induces a critical loss of the mutual information $I[\Omega; \Psi]$ and makes ambiguous our knowledge about the samples of interest. The amount of mutual information $I[\Omega; \Psi]$, therefore, should underlie the logical evaluation of the whole analytical system.

The simple *fu*nction of *m*utual *i*nformation (*FUMI*), which describes the mutual information of chromatography, has recently been proposed[2]. The derivation is based on information theory and the theory of the Kalman filter[2]. Given the numerically expressed shape of each chromatographic peak and the noise level in the measurement process, *FUMI* can provide mutual information about overlapped chromatographic peaks with various resolutions. The whole analytical system, therefore, can be evaluated by taking into account both the observation period of the chromatography and the information loss caused by the peak overlap. The best method is defined as one with the maximal flow of mutual information under the experinental conditions adopted, *i.e.*, one that can transmit the largest amount of information in unit time.

Our aim is to apply *FUMI* to the logical evaluation of an HPLC method for quantitation. The method adopted here as an example concerns a rapid but incomplete HPLC separation of naphthalene and diphenyl and the mathematical processing of the overlapped peaks by means of the Kalman filter[3]. We examine the problems of whether or not the data processing of the Kalman filter can outweigh the information loss caused by the peak overlap in rapid chromatography and whether the whole system can provide the maximal information flow. Logical optimization of overlapped chromatograms is treated in Part II[4].

THEORETICAL

We give a brief review of *FUMI* and some additions for its rational utilization. An approximate function $P_k^\dagger$ plays an important role in the derivation of *FUMI*[2]. $P_k^\dagger$ is derived from the strict P_k (the error variance involved in the Kalman filter algorithm) by mathematical induction[2]. It has the further theoretical importance that the correlation between the Kalman filter and the linear least-squares method can be elucidated clearly by $P_k^\dagger$ (ref. 2).

For a single peak, *FUMI* represents the mutual information that we can collect through the filtering of the raw data ranging from a data point $i = 1$ to k ($k = 1, \ldots, N$)[2]:

$$FUMI = -\frac{1}{2}\log\left(P_k^\dagger\right) \quad (1)$$

$$= \frac{1}{2}\left[\log\left(\sum_{i=1}^{k} F_i^2\right) - \log\left(\tilde{W}_c\right)\right] \quad (2)$$

where F_i denotes the signal intensity of a peak at a data point k and $\tilde{W}_c$ is the variance of the contaminating noise (= constant). We see that larger peaks provide more information and that *FUMI* increases as a new signal F_{k+1} appears, but never decreases. The mutual information for q peaks partially overlapped is given at the last point N (the observation period of the chromatogram) and is simply described as the sum of the individual peak information[2]:

$$FUMI = \frac{1}{2}\left\{\sum_{j=1}^{q} \log\left[\sum_{i=k_c(j)+1}^{k_f(j)} F_i(j)^2\right]\right\} - \frac{1}{2} q \log(\tilde{W}_c) \tag{3}$$

where $[i = k_c(j) + 1, k_f(j)]$ denotes the region where the signals $F_i(j)$ of the jth peak contribute to $FUMI$. The cutoff point $k_c(j)$ is specified to be the point where the signal $F_i(j)$ first gains predominance over the noise $\tilde{W}_c$. The filtering-off point $k_f(j)$ is defined as the cutoff point $k_c(j + 1)$ of the following peak $j + 1$. Without peak overlap, the overall shape of every peak contributes to $FUMI$. If two peaks weakly overlap, then the virtual peak lacking the tailing edge after $k_f(1)$ $[= k_c(2)]$ is input in $FUMI$.

The efficiency of the chromatograms is given as the mutual information in unit time:

$$I_E[1, N] = \frac{FUMI}{N} \tag{4}$$

This function denotes the averaged flow of mutual information through the chromatography and the filtering of the whole data sequence F_1–F_N $[N = k_f(g)]$.

For convenience to chromatographers, we shall describe the mutual information in terms of the relative standard deviation (R.S.D.) of the filtering error $P_k^\dagger$. The error variance $P_k^\dagger$ is derived from the assumption that the amount (or concentration) indicated by the signals F_k of the peak is unity. Hence the R.S.D. of the error at a point k is given by

$$\text{R.S.D.}_k = (P_k^\dagger)^{1/2} \cdot 100 \tag{5}$$

R.S.D.$_k$ can be described by the minimum error $P_{\min}^\dagger$. Let $P_k^\dagger$ be $\alpha^2 P_{\min}^\dagger$:

$$\text{R.S.D.}_k = \alpha(P_{\min}^\dagger)^{1/2} \cdot 100 \tag{6}$$

The coefficient α denotes the ratio

$$\alpha = \frac{\text{R.S.D.}_k}{\text{R.S.D.}_{\min}} \tag{7}$$

The minimum error R.S.D.$_{\min}$ or $P_{\min}^\dagger$ corresponds to the maximum information $I_{\max}$ (see eqn. 1). Usually, the observation period N is specified to be wide compared with the peak region and taken as a point that can give the maximum information ($I_{\max} = I_N$). The information loss δI is defined as

$$\delta I = I_{\max} - I_k \tag{8}$$

$$= -\frac{1}{2}\log(P_{\min}^\dagger) + \frac{1}{2}\log(P_{\min}^\dagger \alpha^2) \tag{9}$$

$$= \log \alpha \tag{10}$$

where I_k denotes *FUMI* at k. The R.S.D. ratio α is given by the information loss:

$$\alpha = \exp(\delta I) \tag{11}$$

The filtering error R.S.D.$_{,k}$ is described in the intelligible form

$$\text{R.S.D.}_{,k} = \exp(\delta I) \cdot \exp(-I_{\max}) \cdot 100 \tag{12}$$

(see eqns. 1 and 6). R.S.D.$_{,k}$ and δI denote the two measures of the excess error or the lost information when the data processing is stopped at k.

For multi-component chromatograms, let us consider the information loss or the filtering error, which depends on the degree of peak overlap, the whole data sequence is analysed. We give a convenient method covering the worst condition that the loss is concentrated on a peak: only two peaks overlap. For simplicity, it is assumed that all the peaks move along the time scale (k) of the chromatogram without any changes in shape. The maximal information $I_{\max}$ will be picked up from a chromatogram with peaks sufficiently or even well separated from each other. The error in the worst case is easily estimated by the information loss. The mean of $I_{\max}$ is defined as

$$\langle I_{\max} \rangle = I_{\max}/q \tag{13}$$

According to eqn. 12, the worst error is given by the information loss δI:

$$\langle \text{R.S.D.} \rangle = \exp(\delta I) \cdot \exp(-\langle I_{\max} \rangle) \cdot 100 \tag{14}$$

We can easily obtain, using eqn. 14, an estimate of the error ⟨R.S.D.⟩ concentrated on a peak. If $\delta I = 0$, then the error of the filtering is equal to the minimum error and may be completely negligible, indicating that the selected experimental conditions are optimal. If $\delta I = 1$, then the filtering error is e-fold more than the minimum error. It should be noted that R.S.D.$_{,k}$, ⟨R.S.D.⟩, $P_k^\dagger$, etc., are not referred to the total HPLC error involving the elution process, detection process, etc., but depend only on the error in the Kalman filtering[2].

EXPERIMENTAL

All the programs were written in BASIC. A PC-9801 VX desk-top computer (NEC) equipped with an Intel-80286 compatible CPU (8 MHz), a 640-kbyte RAM and two 5-in. floppy disk drives was used. The cutoff point $k_c(i)$ was specified to be the point of the fronting edge with 0.5% signal of the peak maximum.

The chromatographic experiments were performed on a Model 655A-11 liquid chromatograph system (Hitachi) with an Inertsil ODS column (250 mm × 4.6 mm I.D.) (Gasukuro Kogyo). The details have been described previously[3].

RESULTS AND DISCUSSION

A chromatogram of naphthalene and diphenyl is shown in Fig. 1. The overlapped peaks have an apparent ratio of the height of the valley to that of the mean

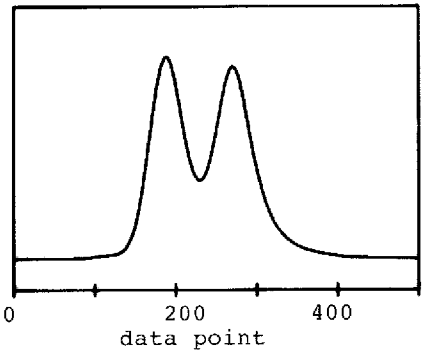

Fig. 1. Chromatogram of a mixture of naphthalene (leading) and diphenyl (trailing). The abscissa denotes the number of the data acquired by an analogue-to-digital converter at 200-ms intervals, 420 s after injection. For details, see ref. 3.

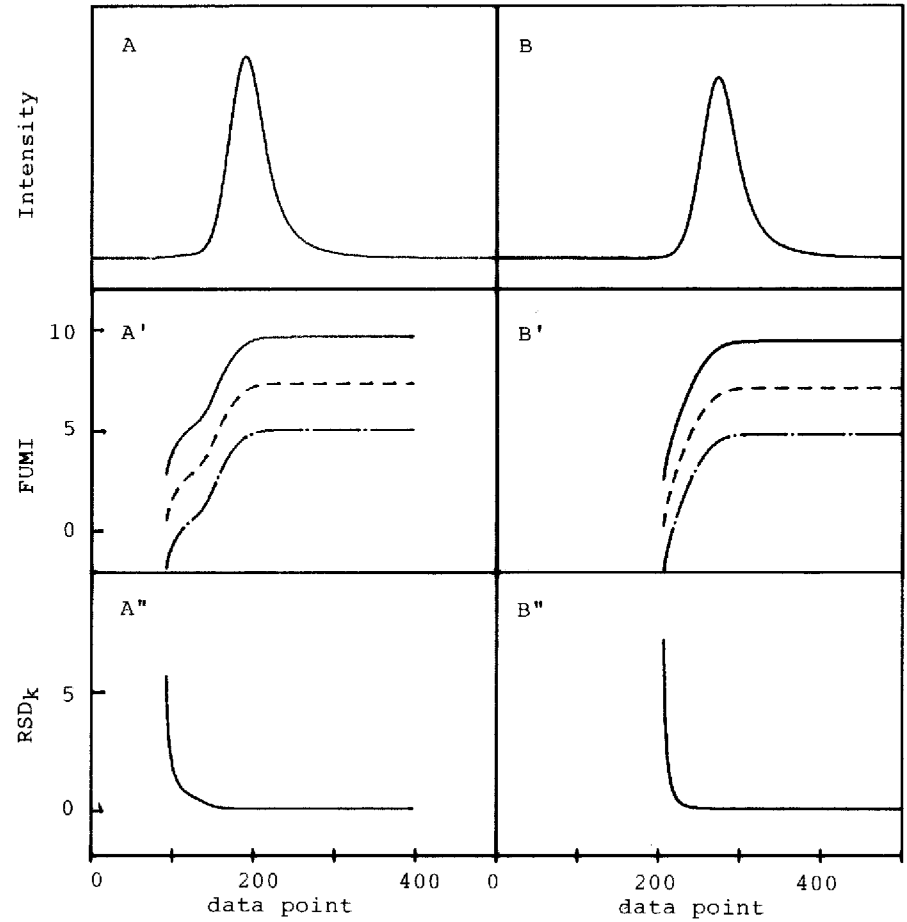

Fig. 2. Chromatograms of (A) naphthalene and (B) diphenyl, (A′ and B′) the time (k) courses of the mutual information and (A″ and B″) the filtering error. (A) I_{max} = 9.68; (B) I_{max} = 9.46. ——, X_s = 1; – – – –, X_s = 10; –·–·–, X_s = 100. Experimental conditions as in Fig. 1.

peak maximum of *ca.* 40% and were analysed by the Kalman filter[3]. The analytical system was concluded to be satisfactory: the observed total error, obtained from five experiments, was less than 0.7% (R.S.D.) for both peaks[3]. The above system was originally evaluated quantitatively but empirically[3]; it has now been evaluated again but deductively based on *FUMI*.

Fig. 2 shows the individual peaks of (A) naphthalene and (B) diphenyl in Fig. 1, (A′ and B′) the time (k) course of *FUMI* and the (A″ and B″) R.S.D.$_{\cdot k}$ of the filtering error. As the observation proceeds, the mutual information increases and reaches a maximum around the peak maximum. The filtering error R.S.D.$_{\cdot k}$ displays the opposite behaviour to *FUMI*. In other words, our knowledge about the samples increases greatly or becomes more precise through observation and analysis until the vicinity of the peak maxima. The slight waving in the early region of *FUMI* for the

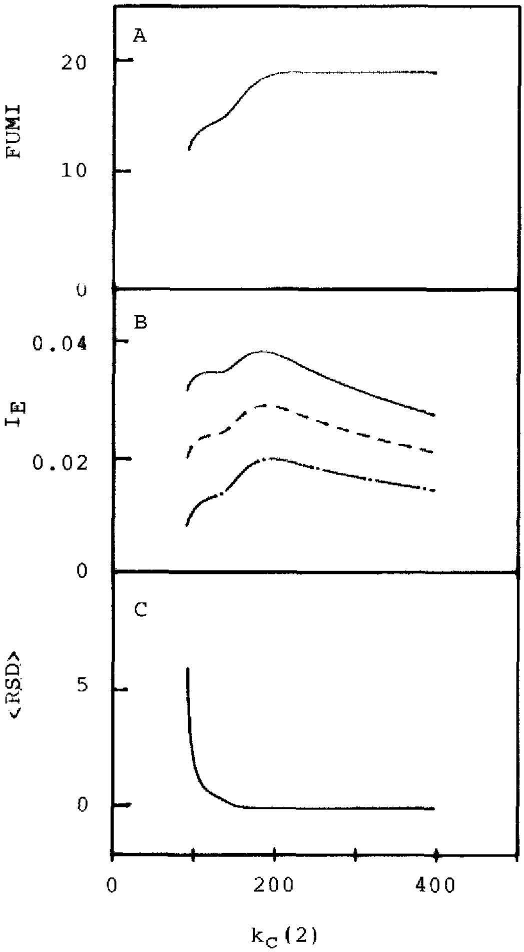

Fig. 3. Influence of the peak overlap on (A) *FUMI*, (B) $I_E[1, N]$ and (C) $\langle$R.S.D.$_{\cdot k}\rangle$ for a mixture of naphthalene and diphenyl. The abscissa denotes the cutoff point $k_c(2)$ of the diphenyl peak. I_{max} = 19.14 (= I_{max} of naphthalene + I_{max} of diphenyl). ——, $X_s = 1$; – – –, $X_s = 10$; – · – · –, $X_s = 100$. $N = k_c(2) + 282$.

naphthalene peak seems to come from the conspicuous fronting of the peak[3]. We can see that after the peak maxima, no further appreciable amount of information can be obtained from the Kalman filtering of the chromatograms. This suggests that the most efficient chromatogram must consist of overlapped peaks, and not only baseline separation.

Let us consider the relationship between the mutual information and the degree of peak overlap. In Fig. 3, (A) *FUMI*, (B) the efficiency $I_E[1, N]$ and (C) $\langle$R.S.D.$\rangle$ for the overlapped peaks are plotted as a function of the cutoff point $k_c(2)$ of the trailing peak. It is assumed that the trailing peak moves the time scale k without any changes in shape; the position of the leading peak is fixed. As the peaks are increasingly separated, the efficiency increases, reaches a maximum and decreases hyperbolically. If the peaks overlap completely $[k_c(1) = k_c(2)]$, then no information can be obtained and the efficiency takes the least value. The hyperbolic decrease in $I_E[1, N]$ is due to the saturation of the mutual information for both peaks. It should be noted that the last peak gives the saturated or maximal information according to our definition.

The most efficient chromatogram with the maximum of $I_E[1, N]$ is simulated in Fig. 4C. The input signals F_i^* $(= F_i/X_s)$ is 100-fold smaller than the actual signals F_i. The peak suppression ($X_s = 100$) means that $I_E[1, N]$ allows for the "unexpected" small peaks (for theoretical meanings, see below). This chromatogram is the "best" in that it can transmit the mutual information in the most efficient way through observation and filtering. In other words, the flow of information through the chromatogram is maximal. Fortuitously, it corresponds to the actual chromatogram shown in Fig. 1. The total errors of the small peaks (6% concentration) of naphthalene and diphenyl

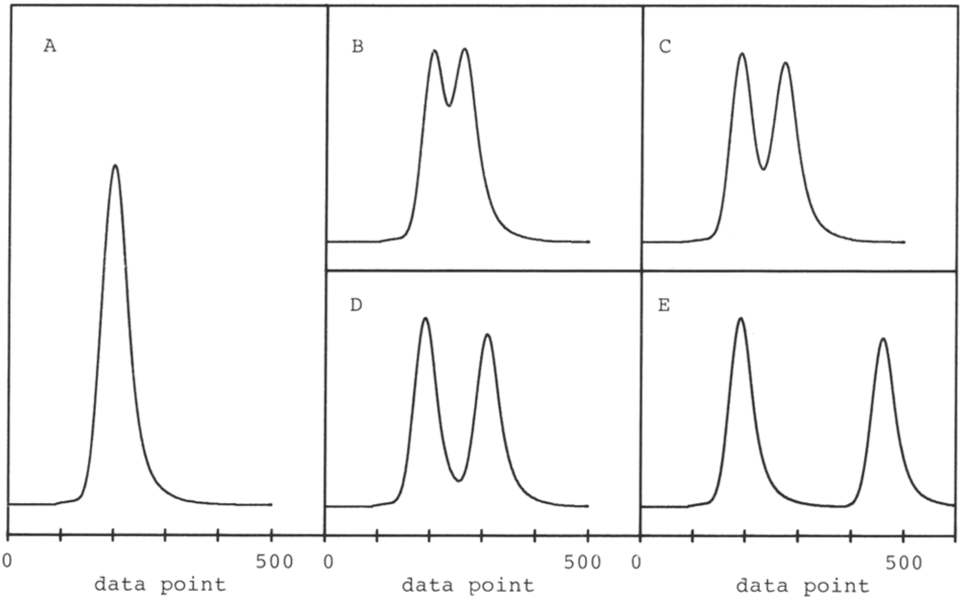

Fig. 4. Simulated chromatograms with various amounts of mutual information. (A) Peak interval $p = 19$, information loss $\delta I = 3.66$, R.S.D.$_{min}$ = 0.27%, period of the chromatogram $N = 436$; (B) $p = 61$, $\delta I = 0.60$, R.S.D.$_{min}$ = 0.013%, $N = 478$; (C) $p = 73$, $\delta I = 0.25$, R.S.D.$_{min}$ = 0.009%, $N = 490$; (D) $p = 119$, $\delta I = 0.006$, R.S.D.$_{min}$ = 0.007%, $N = 536$; (E) $p = 269$, $\delta I = 0.000$, R.S.D.$_{min}$ = 0.007, $N = 686$.

were reported to be *ca.* 4% and 11% (R.S.D.), respectively, and larger than those for the 100% materials shown in Fig. 2A and B[3]. This result indicates the necessity to use the suppression factor X_s in evaluating or designing versatile determination (see below). Some chromatograms with their own information are also shown in Fig. 4. For the more strongly overlapped peaks (A and B), the observation time is relatively short, but sufficient information cannot always be obtained. For the chromatograms with weakly overlapped peaks (D and E) the opposite situation applies.

FUMI and $I_E[1, k]$ have been derived on the assumption that the data processing involved in the calculation of the mutual information is the one-dimensional Kalman filter for peak resolution[5,6]. If another type of data handling with inferior peak-resolving power such as the commonly used perpendicular dropping is utilized, then the chromatograms with the weak peak overlap (D or E) would be the most efficient. The bias for the estimates provided by the common technique was shown to be not less than 10% for the overlapped peaks, whereas the Kalman filter gave a *ca.* 0.2% bias[3]. The superior data processing with the Kalman filter can provide a faster flow of mutual information than the commonly used technique.

The information loss δI is very important as it can serve to solve the problem of whether or not the rapidity of the chromatography shown in Fig. 1 can outweigh the incomplete peak separation in the elution process[3]. The loss δI in Fig. 1 is shown to be negligibly small in comparison with the total HPLC error in the following way. The information loss of the leading peak is 0.25 and the corresponding filtering error ⟨R.S.D.⟩ is 0.009% for a 100% concentration. These values for the trailing peak are the most favourable. On the other hand, the total error in the whole HPLC system covering the elution, detection and filtering was observed to be *ca.* 0.7% (R.S.D.) for the overlapped peaks; the error or reproducibility of the whole system, measured with a solution of diphenyl, was 0.24%[3]. The filtering error is far smaller than the observed total HPLC error; the information loss δI (= 3.538) corresponding to the HPLC error (= 0.24%) is far larger than δI for the filtering (= 0.25). We therefore conclude that the rapid analysis design comprising Kalman filtering of the overlapped peaks shown in Fig. 1 is excellent if the overall HPLC errors are acceptable.

When the suppressed signals F_i^* are input in *FUMI*, the values of the mutual information and the efficiency function decrease (see Figs. 2A′, 2B′ and 3B). The position of the efficiency maximum increases slightly with increasing X_s (see Fig. 3B) and the degrees of peak overlap of the best chromatograms are varied accordingly. The factor X_s seems to remain arbitrary in the function I_E, but is closely related to the linearity of the filter involved in *FUMI* and of the signals of HPLC systems[2]. X_s can be determined as characteristic of a particular HPLC system used. Incomplete linearity of HPLC signals has been observed in special cases and interfered with the successful application of the linear filter[5,6]. This is the case for *FUMI*. If the HPLC signals completely satisfied the linearity postulate, there would be no need for the above consideration and Fig. 4B (X_s = 1) would be given as the best in the ideal HPLC system. The factor X_s, therefore, was introduced to bridge the gap between theory and practice, and should be specified according to the HPLC linearity observance[2]. The necessity to use X_s in proposing the best chromatogram (Fig. 4C) can be interpreted by the same situation: the small peaks that would give poor precision should be appropriately separated in the chromatogram for a successful determination. The relationship between the suppression factor X_s and the peak-resolving powers of the one-dimensional Kalman filter for peak resolution was described previously[2].

The best chromatogram presented here holds true for the adopted experimental conditions involving linear Kalman filtering and an HPLC system with limited linearity. Its elution pattern may vary according to the conditions adopted. If the analysis is carried out or designed under more ideal situations, then more complicated chromatograms will be provided as the best (see above). The experimental design should be performed so that the mutual information may be maximal or most efficiently transferred to real situations.

CONCLUSION

The overall analytical system shown in Fig. 1 has been logically evaluated with the aid of *FUMI* and concluded to provide the maximal information flow. Information theory and the Kalman filter underlie the procedure proposed here and no experience has been utilized. *FUMI* is a simple, quantitative description of the Shannon mutual information in chromatography and will be useful in various areas of analytical chemistry.

REFERENCES

1 S. Arimoto, *Information Theory*, Kyoritsu Shuppan, Tokyo, 1976.
2 Y. Hayashi, S. Yoshioka and Y. Takeda, submitted for publication.
3 Y. Hayashi, T. Shibazaki, R. Matsuda and M. Uchiyama, *J. Chromatogr.*, 407 (1987) 59.
4 R. Matsuda, Y. Hayashi, M. Ishibashi and Y. Takeda, *J. Chromatogr.*, 462 (1989) 23.
5 Y. Hayashi, S. Yoshioka and Y. Takeda, *Anal. Chim. Acta*, 212 (1988) 81.
6 Y. Hayashi, R. Matsuda, S. Yoshioka and Y. Takeda, *Anal. Chim. Acta*, 209 (1988) 45.