Revue Roumaine de Chimie, **2006**, *51*(4), 317–322

Dedicated to Professor Dr. ALEXANDRU T. BALABAN, member of the Roumanian Academy on the occasion of his 75th anniversary

CHARACTERIZATION OF MULTI-SIGNALS ANALYTICAL OUTCOME BY MEANS OF THE INFORMATION ENTROPY AND ENERGY

Victor DAVID[a] and Constantin MIHAILCIUC[b]

[a] University of Bucharest, Faculty of Chemistry, Department of Analytical Chemistry,
Sos. Panduri nr. 90, Sect. 5, Bucharest, Roumania (Vict_David@yahoo.com)
[b] University of Bucharest, Faculty of Chemistry, Department of Chemical Physics,
Bd. Regina Elisabeta nr. 4-12, Bucharest 3, Roumania

Received January 17, 2006

Application of information concepts (information entropy, energy and temperature) to multi-signal analytical outcome was discussed. Chromatographic data sets were taken as real examples in order to calculate the information descriptors by simple mathematical procedures. Two possibilities were studied: the calculation of these descriptors for individual chromatographic peaks and finally by summing them, or the calculation information entropy and energy for the entire chromatogram and comparing to the first possibility. The entire mathematical procedure and data analysis was focused on chromatograms obtained for a mixture of 14 aromatic hydrocarbons by liquid chromatography with ultraviolet absorption detection.

INTRODUCTION

The information theory[1,2] evolved at the same time with the development of communication technology and cybernetics. The application of this theory in analytical chemistry started later as a result of the similarity between the communication and analytical systems. Major contributions to this topic have been brought by several authors soon after 1970.[3-8] Finally, one simple remark could be added to this outstanding theory: a very important role within the framework of information theory as Shannon had advanced it in 1948 is played, unexpectedly, by number 2. This is the base of logarithm that has been used for expressing the measure of information entropy. The concept of duality together with the meaning, origins, and uses of number two were thoroughly discussed by Pogliani, Klein and Balaban[9] in an attempt at showing how everyday concepts have much broader implications, and can be the source of unlimited curiosity.

The outcome of many analytical processes applied to multicomponent samples is usually multi-signal. For instance, the most known example is the chromatographic process, which is a powerful analytical technique providing both qualitative and quantitative information about multi-component samples. In this case, the evaluation of analytical data by means of information theory is now facilitated by the recent development in the field of data acquisition and data processing. The aim of this paper is to apply some information descriptors to the chromatographic outcome and to discuss their magnitude according to the specific computation procedures applied to the chromatographic data set.

THEORETICAL BACKGROUND

A set of given events, $\{X_i\}$, i=1,n, can be characterized from the information point of view by two major descriptors: information entropy introduced by Shannon (denoted by H),[1] and information energy introduced by Onicescu (denoted by E).[10,11] For this purpose a probability must be assigned to each event, denoted by $p(X_i)$. The information descriptors have been proposed as the following well-known relationships:

$$H = -\sum_{i=1}^{n} p(X_i) \cdot \log p(X_i) \tag{1}$$

$$E = \sum_{i=1}^{n} p(X_i)^2 \tag{2}$$

(The log-base is 2). Their dependences on $p(X_i)$ are opposite; H reaches a maximum value (log n) for equal probability for each events, $p(X_i) = 1/n$, which means a maximum disorder, while E attains a minimum value (1/n) for the same situation and a maximum value ($E_{max} = 1$) for a minimum disorder, *i.e.* all $p(X_i) = 0$, i = 1, n-1, excepting $p(X_n) = 1$. The connection between these two descriptors has been observed by Lepadatu,[12] and thus another information descriptor has been borrowed from thermodynamics and, consequently, called the information temperature (T_{inf}):

$$T_{\inf} = \frac{E}{H} = -\frac{\sum_{i=1}^{n} p(X_i)^2}{\sum_{i=1}^{n} p(X_i) \cdot \log p(X_i)} \tag{3}$$

The above remarks lead to the conclusion that the extreme value of T_{inf} is a minimum point, which is obtained for $p_i = 1/n$, for all events X_i. Consequently, the minimum value of T_{inf} becomes $\frac{1}{n \cdot \log n}$.

A multi-signal outcome can be sampled into a number of values that can be used to estimate the probability $p(X_i)$. If the value of the signal in a particular point is denoted by A_i, then the event X_i becomes the value of A_i and the expression of $p(X_i)$ can be advanced as a normalized proportion of this point to the sum of all the signal points:

$$p(X_i) = \frac{A_i}{\sum_{i=1}^{n} A_i} \tag{4}$$

Recently, it has been shown that introducing the probability (4) in eq. (1) and applying Jensen's inequality, a minimum of the information entropy[13] (H_{min}) can be estimated more easily, according to the next relationship:

$$H_{\min} = \log \frac{(\sum_{i=1}^{n} A_i)^2}{\sum_{i=1}^{n} A_i^2} \tag{5}$$

On the other hand, by introducing $p(X_i)$ according to the eq. (4) into the formula of E (eq. 2), one obtains:

$$E = \frac{\sum_{i=1}^{n} A_i^2}{(\sum_{i=1}^{n} A_i)^2} \tag{6}$$

In this way, the dependence between H_{min} and E is given by the next simple relationship:

$$H_{\min} = -\log E \tag{7}$$

These descriptors can be computed for individual signals from a multi-signal chromatogram, knowing their amplitude at different time moments (t_i). In the case of spectrometric detection the signal amplitude is given by the absorbance at different time moments during the chromatographic run. The procedure of dividing a signal into several portions is known in signal theory as a sampling procedure, and in the case of modern chromatographic techniques this is achieved automatically by means of computer software for a chosen acquisition rate.[14]

APPLICATION

A common chromatogram obtained at the separation of 14 aromatic hydrocarbons by a reversed-phase high-performance liquid chromatography (HPLC) with diode array detection (DAD) is depicted in Fig. 1. Some experimental conditions are the following: isocratic elution with water/acetonitrile (40/60, v/v) on a C18 modified silica chromatographic column, at a temperature of 30°C, and detection at 254 nm. Data acquisition and primary processing were assisted by CHEMSTATION software used with the liquid chromatograph (Agilent Technologies). The entire chromatographic run lasted for approximately 30 min, with data acquisition every 0.007 min, resulting in about 4500 sampling points. The above information descriptors were computed or the chromatographic signal located within the interval [1.75; 30.0] min, owing to the fact that the dead time value for this chromatographic run was situated at about 1.75 min. Up to this point, no real signal could be assigned to analytes eluting from the column.

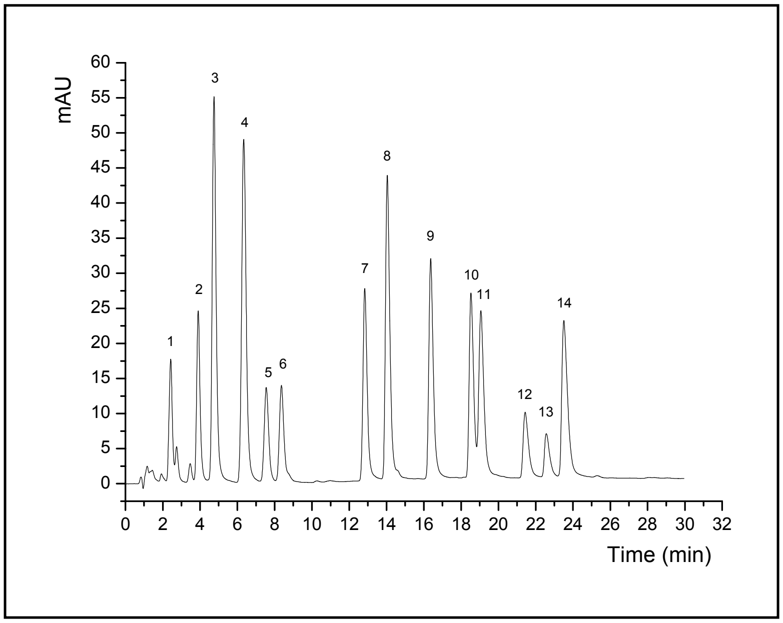

Fig. 1 – HPLC-DAD chromatogram of 14 polyaromatic hydrocarbons.

Data content of a chromatogram can be easily imported into an Origin program and processed in accordance with the above relationships. Each chromatographic peak was isolated to the baseline, such that to be separated to the previous or to the next peak, excepting the peaks #1, 10 and 11. In Fig. 2 the four sampled peaks from the above chromatogram are shown. The fit with a Gauss shape was more than acceptable for all 14 chromatographic peaks (correlation coefficient higher than 0.9). Before data processing the entire chromatogram was translated by a small absorbance in order to obtain only positive values of the absorbance in accordance with the conditions imposed to event probabilities (positive, and situated within the interval [0; 1]). The major parameters describing the chromatographic peaks and the computed information descriptors (entropy, energy and temperature) are given in Table 1. As can be seen, the value of the three information descriptors kept a constant value for all 14 chromatographic peaks. For comparison, in Table 2 the same descriptors were computed for the outcome that does not contain observable chromatographic peaks, unless a special attention is paid to the possible peaks distinguishable to the background noise by increasing the sensitivity of the data reporting. In the case of the chromatographic baseline the signal magnitude tends to be equal for all points belonging to it, and thus, their probability tends to be equal. Therefore, the information entropy reaches its maximum, while the information energy attains minimum values. This fact can be observed from the two tables. However, the baseline between 15.095 and 16.082 min is characterized by a lower value for H and a higher value for E than the other two retention time domains. In this case it is possible that some minor chromatographic peaks may be distinguishable from the baseline that contributes to this possibility. Indeed, as can be seen from Fig. 3, two minor peaks are slightly rising up from the background noise, and bringing their contribution to the information entropy and energy. Besides that, the number of points in this isolated signal is smaller than the other two baseline domains isolated from the chromatogram.

Table 1

Information parameters for the individual peaks given in Fig. 1

Peak #	Retention time t_R (min)	Standard deviation (σ)	A_{max} (mAU)	N_i	H_i (bits)	E_i	T_i
1	2.422	5.908	18.45	71	5.791	0.0207	$3.57 \cdot 10^{-3}$
2	3.902	8.172	25.326	112	6.109	0.0183	$2.99 \cdot 10^{-3}$
3	4.755	17.721	54.959	153	6.273	0.0168	$2.68 \cdot 10^{-3}$
4	6.349	16.494	49.035	148	6.387	0.0149	$2.33 \cdot 10^{-3}$
5	7.549	4.648	14.413	125	6.471	0.0137	$2.12 \cdot 10^{-3}$
6	8.362	4.642	14.032	100	6.298	0.0147	$2.33 \cdot 10^{-3}$
7	12.835	9.422	27.833	103	6.188	0.0163	$2.63 \cdot 10^{-3}$
8	14.042	14.877	43.968	99	6.163	0.0163	$2.64 \cdot 10^{-3}$
9	16.369	10.751	32.107	112	6.316	0.0149	$2.36 \cdot 10^{-3}$
10	18.529	8.862	27.201	88	6.143	0.0159	$2.59 \cdot 10^{-3}$
11	19.062	7.825	24.668	107	6.437	0.0132	$2.05 \cdot 10^{-3}$
12	21.435	3.162	10.221	119	6.653	0.0112	$1.68 \cdot 10^{-3}$
13	22.569	2.098	7.147	121	6.738	0.0104	$1.54 \cdot 10^{-3}$
14	23.509	7.584	23.272	139	6.728	0.0111	$1.65 \cdot 10^{-3}$

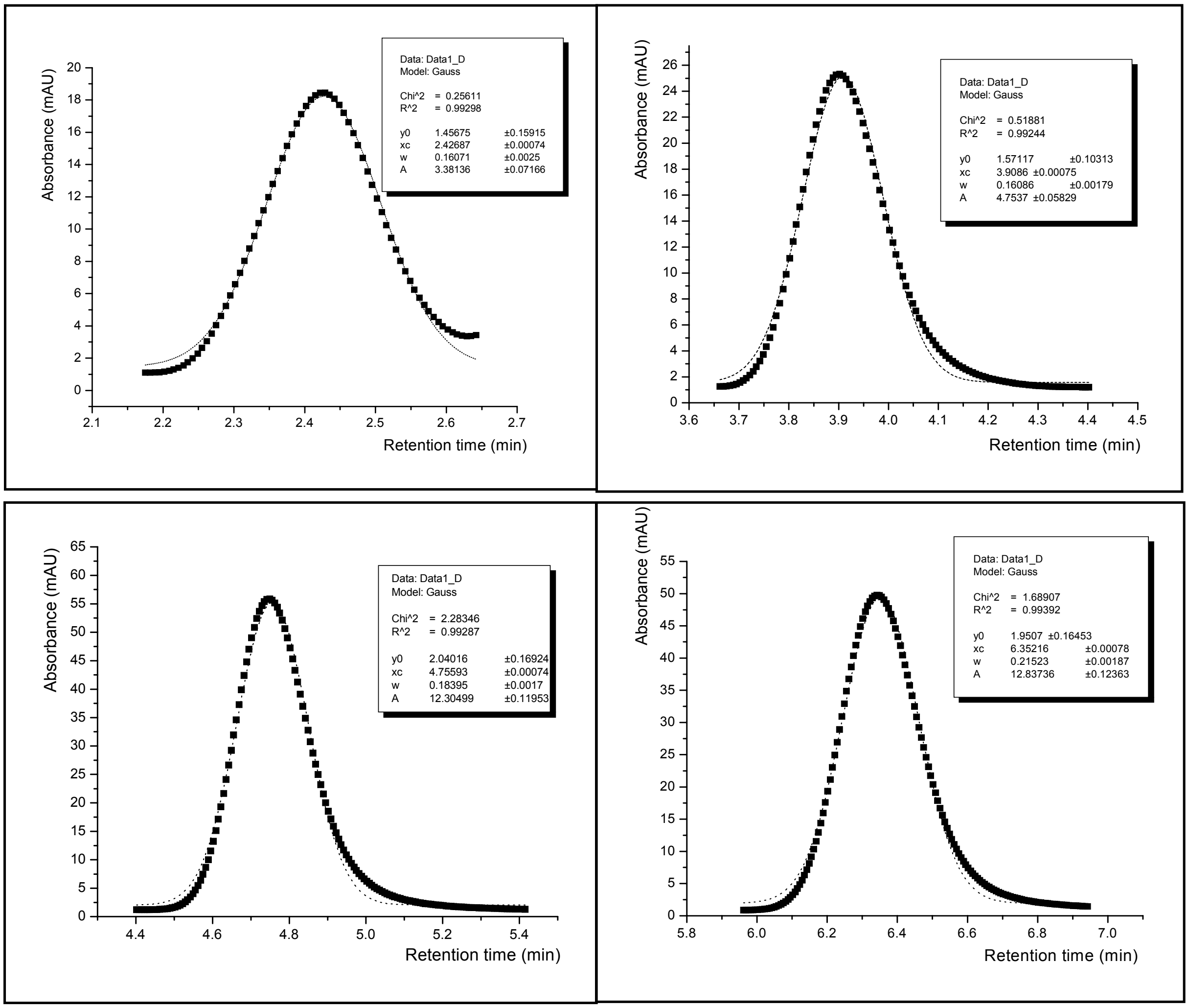

Fig. 2 – Details of the first four sampled chromatographic peaks, together with their Gaussian regression.

Table 2

Information parameters for the chromatographic baseline within three intervals given in Fig. 1

#	Retention time interval (min)	Standard deviation (σ)	A_{max} (mAU)	N_i	H_i (bits)	E_i	T_i
1	9.002 – 12.515	0.0778	0.440	528	9.040	$1.91 \cdot 10^{-3}$	$2.11 \cdot 10^{-4}$
2	15.095 – 16.082	0.0233	1.481	149	7.219	$6.71 \cdot 10^{-3}$	$9.29 \cdot 10^{-4}$
3	24.102 – 29.955	0.185	1.408	879	9.771	$1.15 \cdot 10^{-3}$	$1.18 \cdot 10^{-4}$

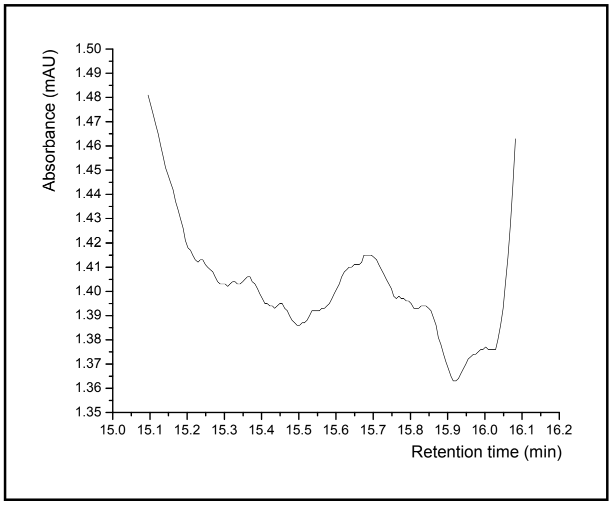

Fig. 3 – Background signal within 15.095–16.082 minutes illustrated at a small scale.

Expectedly, these descriptors are not additive. That means that the information entropy and energy for the entire chromatogram can not be obtained by summing these descriptors for all individual chromatographic signals. If we consider the entire chromatogram to be a single signal and apply the above computation taking into consideration all 4494 sampled points we obtain values for the information entropy, energy and temperature given in Table 3.

Table 3

Information parameters for the entire chromatogram (0–30 minutes) sampled into 4494 points

σ	N_i	H (bits)	Maximum (H) (bits)	E	Minimum E	T	Minimum T
7.916	4494	11.011	12.134	$8.49 \cdot 10^{-4}$	$2.225 \cdot 10^{-4}$	$7.71 \cdot 10^{-5}$	$1.83 \cdot 10^{-5}$

As can be seen the values of the information entropy and energy approach their extreme values, which are obtained for a maximum disorder. Therefore, the entire chromatographic outcome corresponds to a data set that does not fit to a certain mathematical dependence. In this case, conditional probabilities can evaluate the dependence between peaks and events within the peaks, mainly when two detections are applied to the chromatographic separation. However, the question regarding the significance of these descriptors obtained for experimental data sets is still debatable, although the mathematical treatment is now well defined by the theory of probabilities.[15-17]

Data acquisition rate can influence the value of these descriptors. Generally, according to eqs. (1) and (2) it is expected that H may increase and E may decrease with the increase of the data acquisition rate. To prove this statement, the entire computation procedure was applied to the same chromatogram, but this time to a lower acquisition rate (0.015 min). For all 14 chromatographic peaks the information energy E doubled, while the information entropy decreased by 1 bit (*i.e.*, $\log_2 2$). Owing to the fact that the number of sampled points in each chromatographic peak was initially high (see Table 1), the new value of the acquisition rate decreased to the halved number of sampled points without affecting the major characteristics of the peak shape. However, in practice this experimental parameter is limited to a maximum value depending on the response speed of the chromatographic detector.

CONCLUSIONS

The multi-signal chromatographic outcome can be discussed by means of the information theory, using as information descriptors the entropy, energy, and, recently observed, the information temperature. The computation procedure requires the set of experimental values obtained as a digital outcome. It appears that the information entropy, as defined by Shannon's equation, or the information energy, as suggested by Onicescu, can be used to compare different portions of a multi-signal analytical outcome. These information descriptors can reveal the appearance of a signal which is distinguishable from the baseline noise. In this way, this theory may bring a significant contribution to the theory of detection, which is widely discussed in the literature of analytical chemistry.

REFERENCES

1. C.E.Shannon, *Bell Systm. Tech. J.*, **1948**, *27*, 379.
2. C.E.Shannon, *Bell Systm. Tech. J.*, **1948**, *27*, 623.
3. H.Kaiser, *Anal. Chem.*, **1970**, *42*, 24A.
4. K. Eckschalager and V.Stepanek, "Information Theory as Applied to Chemical Analysis", John Wiley & Sons, New York, 1979.
5. C.Liteanu and I.Rica, "Statistical Theory and Methodology of Trace Analysis", Ellis Horwood, Chichester, 1980, p. 79.
6. K.Danzer, M.Schubert and V.Liebich, *Fresenius' J. Anal. Chem.*, **1991**, *341*, 511.
7. D.L.Massart, *J. Chromatogr.*, **1973**, *79*, 157.
8. A.Eskes, F.Dupuis, A.Dijkstra, H. De Clercq and D.L.Massart, *Anal. Chem.*, **1975**, *47*, 2168.
9. L.Pogliani, D.J.Klein and A.T.Balaban, *MATCH-Commun. Math. Comp. Chem.*, **2004**, *51*, 213.
10. O.Onicescu, *Comp. Rend. Acad. Sci. Paris*, Seria A 26, **1966**, *263*, 841.
11. O.Onicescu and V.Stefanescu, "Elemente de Statistică Informațională cu Aplicații", Ed. Tehnică, București, 1979.
12. C.Lepadatu and E.Nitulescu, *Acta Chim. Slov.*, **2003**, *50*, 539.
13. V.David and A.Medvedovici, *J. Chemometrics*, **2005**, *19*, 16.
14. M.Otto, "Chemometrics. Statistics and Computer Application in Analytical Chemistry", Wiley-VCH, Weinheim, 1999, p. 51.
15. J.G.Roederer, *Entropy*, **2003**, *5*, 3.
16. M.Burgin, *Entropy*, **2003**, *5*, 146.
17. C.Menant, *Entropy*, **2003**, *5*, 193.