

Christophe B. Y. Cordella · Julio S. L. T. Militão
Daniel Cabrol-Bass

A simple method for automated pretreatment of usable chromatographic profiles in pattern-recognition procedures: application to HPAEC–PAD chromatograms of honeys

Received: 27 January 2003 / Revised: 19 May 2003 / Accepted: 28 May 2003 / Published online: 19 July 2003
© Springer-Verlag 2003

Abstract In this article a simple method for automated pretreatment of chromatograms is presented. The resulting data matrix can be used as input for multivariate statistical analysis. Application of this method to high-performance anion-exchange chromatography–pulsed amperometric detection (HPAEC–PAD) chromatograms of honeys before canonical discriminant analysis results in very good performance in the data reduction with regard to the preservation of the original information content of the data. This pretreatment of the chromatogram allows for the use of all the peaks corresponding to the sugars present in the sample. This results in a high-quality discrimination between honeys of various types. A versatile program has been developed to apply this method. This serves as a starting point for software suited for food characterization and adulteration detection by semi-automatic pattern recognition applied to chromatographic analysis.

Keywords Chromatographic profiles · Chromatogram pretreatment · Pattern-recognition procedures · Multivariate statistical analysis · Discriminant canonical analysis · High-performance anion-exchange chromatography–pulsed amperometric detection (HPAEC–PAD) · Honey

Introduction and objectives

For over twenty years, the Bee Unit of the French Food Safety Agency laboratory at Sophia-Antipolis has been the laboratory used as the point of reference (OIE and FAO)¹ for the analysis of hive products. These include honey, royal jelly, bees' wax, and pollen. The quality of honeys is assessed through saccharide determinations (mono-, di-, tri-, and oligosaccharides) by high-performance anion-exchange chromatography with pulsed amperometric detection (HPAEC–PAD). This technique enables specific determination of all osidic compounds present in the samples. To date, the routine determination of 13 saccharides is carried out simultaneously following a COFRAC² method. However, this analysis is both tedious and time-consuming, requiring about 45 min for each run, and does not allow for the automatic treatment of the data using the standard software (Peaknet, Dionex) provided with the apparatus. Peaknet is a data-acquisition program which provides numerous tools for classical chromatogram treatment, consisting mainly of integrating each peak in the analytical profile in order to obtain the amount of each chemical substance detected during the analysis after standard calibration. This method is always applicable and does not permit disturbances to occur during the analysis due to the aging of any part of the system. This is because each series of analyses is preceded by a standard calibration. The data obtained by this preliminary treatment can be used in statistical procedures in order to find correlations between measured factors, or more generally to find the relationship(s) that explain(s) the relative behavior of the variables in a data set [1, 2]. This approach is now commonly adopted in pattern recognition/classification methods applied to characterization of food samples [3, 4, 5, 6, 7, 8, 9, 10, 11]. Another possible approach (not supported by the Peaknet Software) is the use of the entire chro-

C. B. Y. Cordella (✉)
Agence Française de Sécurité Sanitaire des Aliments
(AFSSA Sophia-Antipolis),
105 route des Chappes, 06902 Sophia-Antipolis Cedex, France
e-mail: c.cordella@afssa.fr

J. S. L. T. Militão
CNPq and Universidade Federal de Rondonia (UNIR),
Km8 Br364, 78.900 Porto Velho, Rondonia, Brazil

C. B. Y. Cordella · D. Cabrol-Bass
Laboratoire Arômes Synthèses Interactions,
Université de Nice Sophia-Antipolis,
28 avenue Valrose, 06108 Nice Cedex 2, France

¹OIE: Office International des Epizooties; FAO: Food and Agriculture Organization.

²COFRAC: COMité FRançais d'ACréditation, member of International Accreditation Forum (IAF: <http://www.iaf.nu>)

matogram profile. This provides a more effective data source than does classical peak integration. However, the use of entire chromatographic profiles is highly subject to retention time shifts. Therefore it is essential that a procedure for retention time correction is developed. For this purpose a recent innovation in gas chromatographic techniques, called retention time locking (RTL) has been developed by Agilent Technologies. This tool is based on hardware calibrations of the chromatographic system that permits the modification of hardware parameters (flow rate, column temperature, and/or other parameters) in order to maintain stable retention times. The main application of the RTL technique is associated with identification of chromatographic peaks. After having consulted several suppliers of analytical apparatus it was found that no commercial package allows for the automatic pretreatment of standard chromatograms and contains the building of a data matrix suitable for chemometric analysis as its final goal. Nevertheless, the use of liquid chromatography is increasing, particularly in the area of food analysis [11]. This is because this type of chromatography does not require a complex sample preparation and in most cases allows for analysis of the native food sample (solid or liquid).

Algorithms for chromatographic alignment and peak matching using time-warping have been described and applied to gas chromatography/ infrared/mass spectrometry [12] single and multiple wavelength HPLC chromatograms [13] and liquid chromatography/mass spectrometry [14]. The two algorithms most suitable for profile alignment, namely dynamic time warping and correlation optimized warping, were compared to chromatographic and spectroscopic profiles [15]. These methods are very effective for complete profile alignment of equal length prior to data modeling, a condition that is required by most multivariate chemometric methods. However, in our case, we used a much simpler approach for extraction of a fixed vector length corresponding to the 13 saccharides found in honey samples in a complete liquid chromatographic profile. This simple method for building a data matrix from the chromatographic profiles intended to serve as input for multivariate chemometric analyses (such as principal component analysis, PCA, or linear discriminant analysis, LDA) is presented. The discrimination of honeys of various origins serves as an illustration of its applications. Its user-friendly execution is also described.

Method

Pretreatment of the chromatograms

Each chromatogram file is arranged in two columns: retention time (Rt) and intensity (Int). These values are used to construct the raw chromatogram vector: $C = \{[Rt_1, Int_1], [Rt_2, Int_2], [Rt_i, Int_i], [Rt_N, Int_N]\}$. With this notation, the vector element $C(i)$ consists of a doublet where $[Rt_i, Int_i]$ i is the index of the i th value and N is the total number of data values. The value of N may differ for each chromatogram.

This raw chromatogram vector C was subjected to the following transformations:

1. Determination of the maximum value:
 - The maximum intensity value, $MaxInt$, was determined
2. Standardization of intensity values:
 - All the intensities, Int_i , were divided by $MaxInt$ and multiplied by 1000
3. Determination of threshold:
 - In order to eliminate the effect of noise on the detection of the peak, a threshold was subtracted from all the standardized intensity values. The optimal threshold value was found by systematic trials, varying the value between 0 and 20 on a set of standard chromatograms. A threshold value of 8 was found to be suitable for 0.8% of the most intense peaks.
 - Standardized threshold chromatographic vectors C are obtained after applying steps 1–3.
4. Determination of peak:
 - The intensity Int is related to the retention time Rt as a function called F . i.e. $Int=F(Rt)$. The local slope of this function may be found by using a classic numerical method. This indicates whether the chromatographic signal is increasing or decreasing. The peaks are detected by a change in the sign of the local slope. The method includes a heuristic rule, which distinguishes between separate peaks of the chromatogram. The retention times corresponding to the apex of each detected peak are kept in an array of real values of general element $Rtp(j)$, where j is the index of the detected peak; $j=1, \dots, Np$ and Np is the number of detected peaks.
5. Retention time correction:
 - The values of $Rtp(j)$ were adjusted using the retention times of fructose $Rt_{fructose}$ and sucrose $Rt_{sucrose}$. These were chosen to be used as reference peaks because these two sugars are always detected in all honey samples. Linear interpolations were performed separately on the three retention time ranges, i.e. $Rtp(j), <Rt_{fructose}; Rt_{fructose} < Rtp(j) < Rt_{sucrose}$ and $Rt_{sucrose} < Rtp(j)$. The values of retention times for fructose and sucrose along with those of other major sugars found in honeys are provided in Table 1. These values were obtained by using synthetic standard solutions containing all the sugars that were prepared using the COFRAC procedure. The chromatograms were obtained by following standard techniques. The preparation of the standards was repeated 15 times in order to estimate the standard deviation SD of the retention times for all trials.

Peak attribution, i.e. sugar determination

The mean retention time of each sugar $Rts(k)$ from the above table is compared to the nearest retention time value of detected peaks $Rtp(j)$ as follows:

Table 1 Mean parameters used for recognition of peaks

k	Rts (min)	SD (min)	CV	Sugar
1	4.235	1.133	0.267	Trehalose
2	6.166	1.501	0.243	Glucose
3	6.831	1.675	0.245	Fructose
4	7.970	1.969	0.247	Melibiose
5	9.613	2.348	0.244	Isomaltose
6	11.027	2.762	0.250	Sucrose
7	13.832	3.356	0.243	Turanose
8	14.567	3.546	0.243	Palatinose
9	16.025	3.904	0.244	Melezitose
10	17.696	4.329	0.245	Raffinose
11	19.970	4.851	0.243	Nigerose
12	21.837	5.703	0.261	Maltose
13	29.024	9.502	0.327	Erllose

k , sugar index; Rts , mean retention time of sugars; SD , standard deviation; CV , coefficient of variance; number of trials=15

- If $\text{abs}(Rtp(j) - Rts(k)) \leq SD(k)$, then the peak j in the chromatogram is attributed to sugar k , and the corresponding intensity or peak area (see below) is associated to sugar k .
- If a sugar has not been attributed to a peak in the chromatogram, then it is given a value of 0.

A new vector S of a fixed size of 13 equal to the number of sugars results when applied to the chromatographic vector. Depending on the operator's choice, the values of $S(k)$ correspond to the standardized measure of threshold intensity or to the area of the peak which is attributed to the sugar k . Each area is calculated by summing the surface of elementary rectangles below the chromatographic peak.

Data matrix construction

The S vectors, corresponding to each sample, are arranged in rows in order to build the data matrix M . This is then used as input for multivariate statistical analysis.

Practical implementation

The application is written in Borland Delphi 5.0 object-oriented language and takes full advantage of the Microsoft Windows operating system.

Main interface and working processes

The program is built around a main interface that provides different sources of information. The interface was customized for the present study. Thus, the term "honey" appears in several places in the user interface. Figure 1

shows a copy of the user interface screen. The two main graphs display the raw chromatogram (top) and the transformed peaks diagram (bottom) respectively.

The program sequentially transforms all the sample chromatograms listed in the "Honey's files" list box at the top right-hand corner of the window. Before processing, the sample of honey is defined by a selection in the "Honey class" list box. Then the sample chromatogram is transformed as described above (Pretreatment of the chromatograms), i.e. after standardization and determination of its threshold. The retention times are adjusted using the retention times of fructose and sucrose. At this stage the program proceeds to the detection of 13 sugar peaks listed in the "Standard sugar retention times" list box at the bottom right-hand corner of the interface window and compares their retention times to those of the detected peaks in the sample chromatograms. This process produces a new line in the data matrix for each raw chromatogram vector as shown in Table 2.

Table 3 shows part of the resulting data matrix $M(13 \times N_s)$ generated by the program after processing all N_s raw chromatograms. This matrix can be used as a starting point for any chemometric treatment requiring input of equal size.

Example of application

We choose to present the results of a canonical discriminant analysis as an example of chemometric application using the resulting data matrix. The aim of this study was

Fig. 1 Main interface showing the raw chromatogram and peaks after transformation (standardization, threshold, and retention-time adjustment)

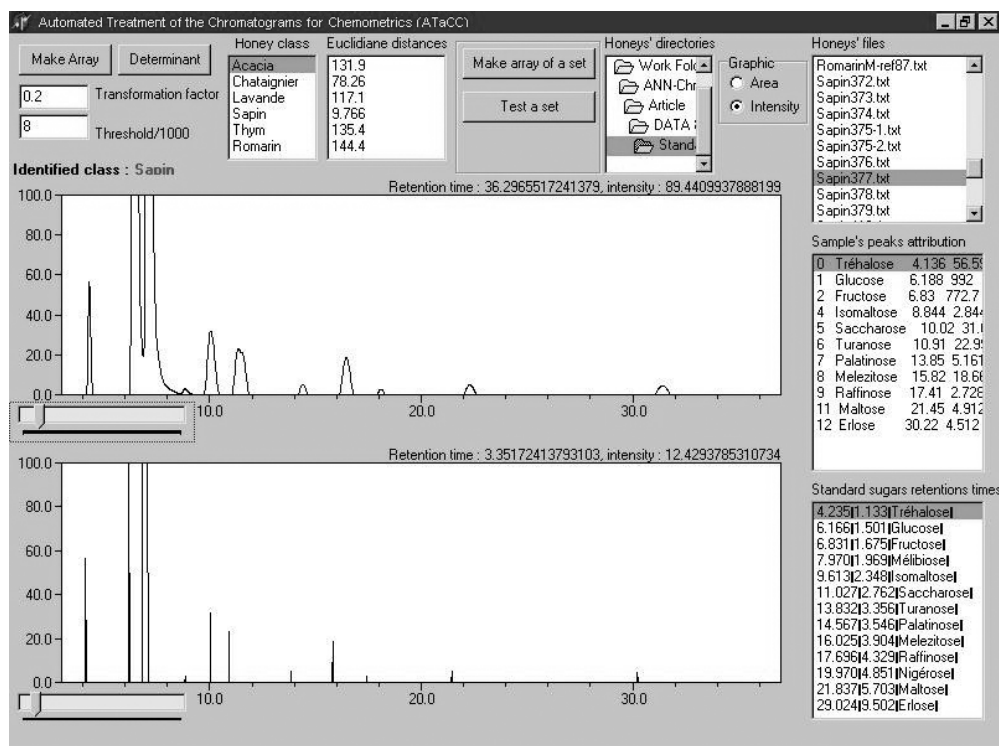


Table 2 Raw sample chromatogram vector C of general element $C(i)=[Rt_i, Int_i]$ and resulting vector S

Rt (s)	0.000	0.017	0.033	0.050	0.067	0.083	...	32.167	32.183	32.200	32.217		
Int (nC)	0	8	181	469	491	596	...	4019	2665	1405	192		
k	1	2	3	4	5	6	7	8	9	10	11	12	13
$S(k)$	61.8	992.0	783.9	0.0	2.5	31.1	27.3	4.5	15.5	3.5	0.0	4.9	11.2

Note that the maximum for $k=2$ (glucose) is equal to 992.0, i.e. 1000 as standardized maximum value minus 8 used as the threshold

Table 3 Example of resulting matrix, M

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13
Lavender13.txt	0.0	992.0	719.2	0.0	7.2	81.8	0.0	0.0	0.0	0.0	0.0	7.6	0.5
Lavender15.txt	0.0	992.0	724.2	0.0	10.1	55.0	0.8	0.0	0.0	0.0	0.0	4.4	0.0
Lavender19.txt	0.0	992.0	734.0	0.0	9.1	91.3	0.0	0.0	0.0	0.0	0.0	8.3	0.0
Lavender63.txt	0.0	992.0	667.9	0.0	0.0	3.9	114.0	0.0	0.0	0.0	0.0	7.2	0.4
Lavender65.txt	0.0	992.0	592.3	0.0	0.0	4.4	35.4	0.0	0.0	0.0	0.0	3.3	0.0
Lavender89.txt	0.0	992.0	683.5	0.0	0.0	3.4	95.4	0.0	0.0	0.0	7.5	0.0	0.5
Fir372.txt	63.8	992.0	775.0	0.0	1.9	29.4	24.3	4.4	16.9	4.8	0.0	3.7	6.3
Fir373.txt	53.7	992.0	781.6	0.0	2.1	31.8	21.1	5.5	16.0	2.2	0.0	4.2	2.0
Fir374.txt	50.7	992.0	794.6	0.0	1.2	29.8	23.8	4.6	11.7	3.8	0.0	3.2	5.1
Fir378.txt	61.8	992.0	783.9	0.0	2.5	31.1	27.3	4.5	15.5	3.5	0.0	4.9	11.2
Fir379.txt	51.0	992.0	774.8	0.0	31.2	20.8	3.3	0.0	14.0	2.5	0.0	3.0	4.0
Fir413-1.txt	47.5	992.0	775.2	0.0	1.3	28.1	23.6	3.4	18.3	3.8	0.0	2.1	6.8

Columns=sugar index; Rows=sample chromatogram index
The row headers contain the chromatogram file name

Fig. 2 Score plots on the Root 2 vs Root 1 projection plane by canonical discriminant analysis applied to the 40x13 reduced data matrix

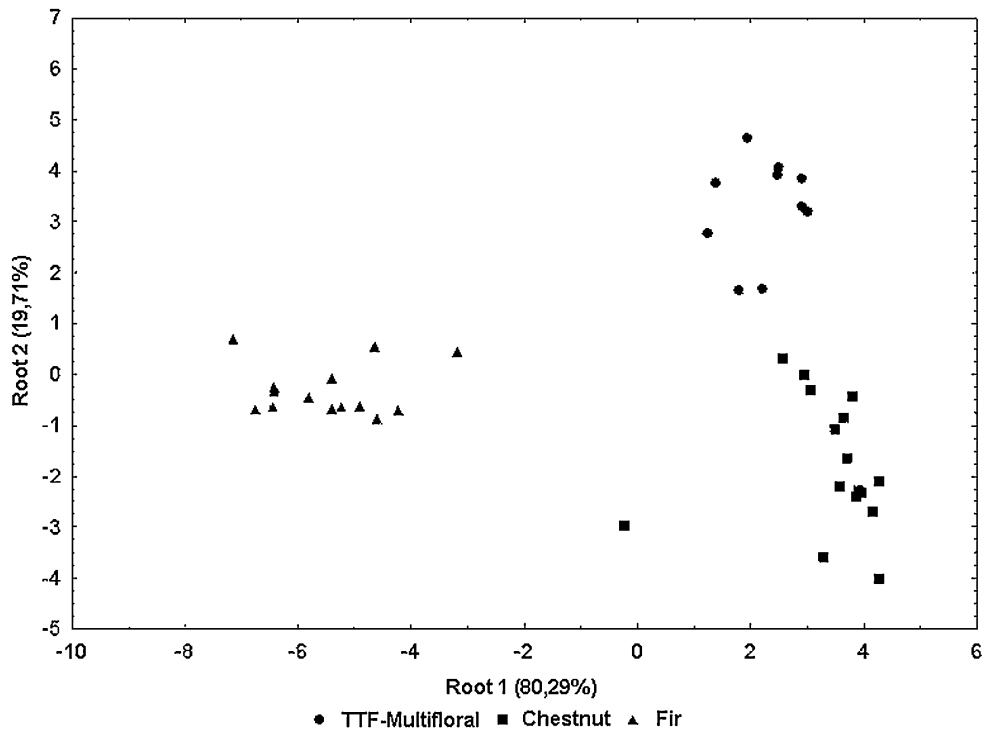


Table 4 Mahalanobis distances between the centroids of the groups

	Multifloral	Chestnut	Fir
Multifloral	0	5.407	8.835
Chestnut	5.407	0	9.341
Fir	8.835	9.341	0

Table 5 Maximum, mean and minimum Mahalanobis distances from the samples of each group to the centroids

Samples		Centroids		
		Multifloral	Chestnut	Fir
Multifloral	Maximum	4.237	7.019	10.225
	Mean	2.586	5.803	8.885
	Minimum	1.486	4.146	7.693
Chestnut	Maximum	8.147	5.890	10.956
	Mean	5.927	2.931	9.480
	Minimum	3.744	1.677	6.959
Fir	Maximum	11.118	11.667	5.837
	Mean	8.973	9.429	2.711
	Minimum	6.974	7.702	1.142

to determine whether the HPAEC–PAD sugar profiles could be a relevant way to distinguish the honey category. In this paper we present only the results obtained for three typical honeys namely: multifloral honeys (TTF), chestnut and fir. For further information about the experimental conditions used to obtain the sample files and for the complete results see Ref. [16]. Statistical analysis was performed using the Statistica 6.0 [17] package.

It is to be noted that raw chromatogram files exported from the analytical instrument are made up of 2400 data points (time×intensity) and cannot be imported directly by Statistica which requires that the variables (factors) be arranged in columns and the observations (samples) in rows. Furthermore, the maximum number of variables that Statistica can manage is limited to 1000. Hence the necessity to develop a procedure to read, reduce, and merge such raw data files in order to produce a data matrix suitable for chemometric analysis.

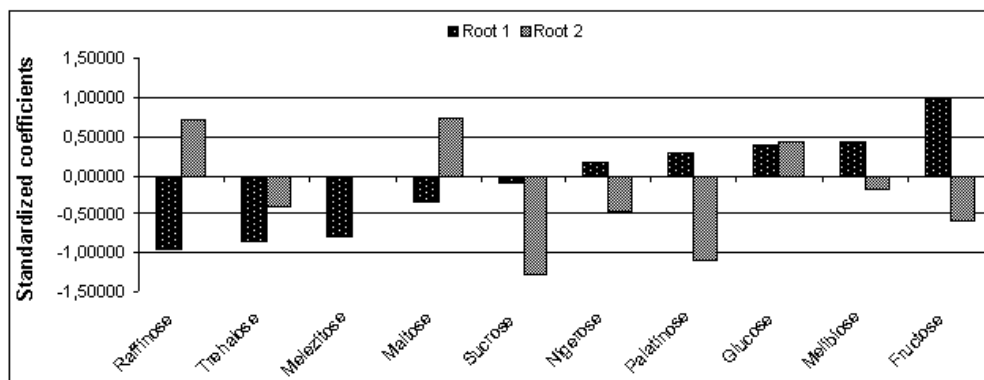
In the example presented 40 honey samples of the three above-mentioned categories (10 multifloral, 16 chestnut and 14 fir) were analyzed. Figure 2 shows the score plots on the Root 2 vs. Root 1 projection plane resulting from canonical discriminant analysis applied on the reduced data matrix (13×40).

It can be seen on the score plots that the projections of samples belonging to the same category fall into three well-separated groups. Moreover, the distances between the centroid of each group (Table 4) are large enough compared to the mean distances of each sample to the centroid of its group (italicized values in Table 5) to distinguish each group without ambiguity (for details see Ref. [12]).

From these results it is possible to classify new honey samples of unknown origin by assigning them to the nearest group. Validation by the leave-one-out cross validation method gave a 100% correct classification. Usually, these three categories of honeys are classified as “dark honeys” and are considered as having similar chemical composition, particularly in terms of their sugar profiles, which shows variations for only one or two sugars.

Figure 3 shows the coefficients of the variables in the canonical discriminant roots. A high contribution of a variable does not denote a high amount of the corresponding sugar in the sample but indicates a high discriminating power i.e. a strong contribution to the differentiation of the groups of samples. Only 10 of the 13 sugars are retained to build the discriminant roots, (isomaltose, turanose and erlose having a too low capacity of discrimination).

The adopted approach is not limited to the particular list of sugar present in honeys, the “Standard sugar retention times” list found at the bottom right of the main window can be augmented with other sugars that may be subsequently identified, or with new compounds subject to the constraint that the mean retention time and corresponding standard deviation be previously determined. In this way, the attributed peak list will be enriched with new chemical information present in the original chromatograms and the number of columns of the reduced data matrix generated by the program will be equal to the number of compounds in the standard reference list.

Fig. 3 Contribution of variables to Root 1 and Root 2 of canonical discriminant analysis

Conclusions

This example shows that the proposed simple pretreatment of chromatograms achieves a considerable reduction of the original data file while maintaining the essential chemical information and allowing for the discrimination of samples between each honey category.

This program is well suited for the complete automatic pretreatment of gas or liquid chromatograms and the creation of the corresponding reduced data matrix. This data matrix could be used as input for multivariate statistical methods requiring equal length input vectors. The program produces a data matrix, which is expandable in terms of peaks that can be identified by the analyst. Furthermore, unknown peaks can also be included in the matrix building process in order to preserve the richness of the original data. This program is the first part of a software-suite designed for semi-automatic pattern recognition of food samples from chromatographic analyses. This software-suite is used for the characterization of honeys and for the detection of their adulteration by exogenous sugars from the HPAEC-PAD profiles.

A freeware version and a brief description of the software are available upon request to the authors.

Acknowledgments We thank the referees for useful suggestions, which helped us to improve the manuscript. Dr Tom Forrest, Dalhousie University, Halifax Canada and Marguerite McQuire are also gratefully acknowledged for their help with linguistic revision of the manuscript.

References

- Jambu M (1978) Classification automatique pour l'analyse des données. Dunod, Paris
- Jambu M (1991) Exploratory and multivariate data analysis. Academic Press, London
- Krauze A, Zaleswski RI (1991) *Z Lebensm Unters Forsch* 192: 19–23
- Peña-Creciente R, Herrero-Latorre C (1993) *J Agric Food Chem* 41:560–564
- López B, Latorre MJ, Fernández García MA, García S, Herrero C (1996) *Food Chem* 55:281–287
- García-López C, Grané-Teruel N, Berenguer-Navarro V, Efigenio García-García J, Martín-Carratala ML (1996) *J Agric Food Chem* 44:1751
- Moshonas MG, Shaw PE (1997) *J Agric Food Chem* 45:3968
- Martin-Carratala ML, Garcia-Lopez C, Berenguer-Navarro V, Grané-Teruel N (1998) *J Agric Food Chem* 46:963–967
- Mateo R, Bosch-Reig F (1998) *J Agric Food Chem* 46:393–400
- Siverstsen HK, Holen B, Nicolayen F, Risvik E (1999) *J Sci Food Agric* 79:107–115
- Cordella C, Moussa I, Martel A-C, Sbirrazzuoli N, Lizzani-Cuvelier L (2002) *J Agric Food Chem* 50:1751–1764
- Wang CP, Isenhour TL (1987) *Anal Chem* 59:649–654
- Nielsen N-PV, Carstensen JM, Smedsgaard J (1998) *J Chromatogr A* 805:17–35
- Bylund D, Danielson R, Malmquist G, Markides KE (2002) *J Chromatogr A* 961:237–244
- Pravdova V, Walczak B, Massart DL (2002) *Anal Chim Acta* 456:77–92
- Cordella C, Militão JSLT, Clément M-C, Cabrol-Bass D (2003) *J Agric Food Chem* 50:3234–3242
- StatSoft (2001) Statistica (data analysis software system) version 6. www.statsoft.com