Babes-Bolyai University
Faculty of Chemistry and Chemical Engineering

Chemometric Methods Applied for the Processing of Analytical Signals

PhD thesis Abstract

PhD advisor
Costel SÂRBU

PhD student
Ioana-Anamaria TUHUŢIU (SIMA)

Cluj-Napoca, 2019

Contents

THEORETICAL ASPECTS

CHAPTER I – THEORETICAL CONSIDERATIONS

1.1 Chemometrics

1.2 Chemometric methods

1.2.1 Regression

1.2.2 Principal Component Analysis

1.2.3 Cluster analysis

1.2.4 Linear discriminant analysis

1.3. Applications of chemometric methods

1.3.1 Method validation

1.3.2 QSAR studies

1.3.3 Fingerprinting analysis

1.4 Data acquisition in HPTLC

1.5 Objectives of the thesis

ORIGINAL CONTRIBUTIONS

CHAPTER II – VALIDATION OF A QUANTITATIVE ANALYSIS METHOD FOR CATECHOLAMINES AND THEIR METABOLITES USING HIGH PERFORMANCE THIN-LAYER CHROMATOGRAPHY

2.1 Hypothesis and novelty of the study

2.2 Experimental setup

2.3 Results and discussion

CHAPTER III – QSAR STUDIES AND LIPOPHILICITY DETERMINATIONS

3.1 Hypothesis and novelty of the study

3.2 Prediction of catecolamines lipophilicity using the Cluj topological indices

3.2.1 Data set and correlating algorithm

3.2.2 Results and discussion

3.3 Assessment of lipophilicity indices of antioxidant compounds in RP-HPLC

3.3.1 Materials and methods

3.3.2 Results and discussion

CHAPTER IV – FINGERPRINTING AND AUTHENTICITY DETERMINATION OF WILD FRUITS AND DERIVED DIETARY SUPPLEMENTS

4.1 Hypothesis and novelty of the study

4.2 Fingerprinting of Romanian wild fruits

4.2.1 Materials and methods

4.2.2 Results and discussion

4.3 Authentication and fingerprinting of dietary supplements derived from berries

4.3.1 Materials and methods

4.3.2 Results and discussion

CHAPTER V – CONCLUDING REMARKS

References

List of original publications

Keywords:

Chemometrics

Method validation

QSAR/Lipophilicity

Fingerprinting/Authentication

Catecholamines

Antioxidants

Forest berries

Dietary supplements

Structure and objectives of the Phd thesis

The present Phd thesis is structured in four chapters. The first chapter presents theoretical elements regarding the chemometric and analytical methods used in the thesis. Chapters II–IV present the original contributions to the thesis, the results obtained during the experiments and their interpretation. The last chapter is devoted to the conclusions arising from the conducted experiments.

The thesis focuses on three main objectives, regarding the application of chemometrics in different fields of analytical chemistry, in order to improve the processing and interpretation of the instrumental signal.

The first objective was to demonstrate how chemometric methods represent important tools in the validation of newly developed methods. For this, a new method of analysis was proposed using digital thin-layer chromatography for the investigation of catecholamines' metabolites from biological samples.

Further, for the second objective the chemometric methods were used for modeling and predicting lipophilicity: on one hand for catecholamines and related compounds, using the algoritm proposed by the TopoCluj group, and on the other hand for antioxidant compounds with different structures, using various HPLC experimental conditions.

And finally, the third objective was to point out the necesity of chemometric tools for obtaining holistic and comprehensive fingerprints, characterisation and authentication of various samples. For this, the fingerprinting analysis was applied for wild berries and derived dietary supplements, using various analytical techiques assisted by different chemometric aproaches.

Chapter I – Theoretical considerations

General introduction - Chemometrics

Due to increasing analytical demands for higher specificity and sensitivity, a general trend in analytical chemistry is to produce more and more data per sample. This has been facilitated by developments in instrumentation and computer systems, making large amounts of data possible to produce and store with good economy.

In order to extract the relevant information from the acquired data and to perform good experiments there is a need for methods that can help the analytical chemists make efficient use of the sophisticated analysis systems. The solution is given by methods that are collected under the discipline of chemometrics.

The word “chemometrics” was invented by the Swedish organic chemist, Svante Wold in 1971, when he published a complex application of computational chemistry and thought that a new word will get his research founding. He also founded the International Chemometrics Society, together with an American chemist, Bruce Kowalsky.

Several definitions are given in order to explain the term “chemometrics”. Miller considers that a definition of chemometrics could be: "the application of multivariate, empirical modelling methods on chemical data" [1]. Another definition of chemometrics proposed by Massart et al is: "Chemometrics is a chemical discipline that uses mathematics, statistics and formal logic a) to design or select optimal experimental procedures; b) to provide maximum relevant chemical information by analyzing chemical data; and c) to obtain knowledge about chemical systems" [2]. And on the other hand, the International Chemometrics Society (ICS) defines chemometrics as: "the science of relating measurements made on a chemical system or process to the state of the system via application of mathematical or statistical methods" [3].

Since the 1980s, various research groups have gradually adopted the term "chemometrics" for the statistical methods of data interpretation. The applications of these methods are very comprehensive, covering various fields of chemistry, from analytical chemistry up to organic synthesis.

Chemometrics can now be described from several points of view. Many believe that this area includes applications of modern statistical methods: experimental design, calibration, pattern recognition and signal analysis tool for data in chemistry. Chemometrics can be generally described as an application of statistical methods in chemistry in order to improve the measurement process and to extract the most useful and complete information from raw data obtained from physico-chemical measurements (usually instrumental) [4].

Chemometrics represent an approach of chemical determinations that rely on the notion of indirect measurements. Thus, the measurements associated to the content of certain compounds of a set of samples are related to a property of the tested material, so that this property can be identified in an unknown sample by conducting measurements less laborious than those made initially on the set of well-known samples [5].

The value of chemometric techniques is based on the fact that the experimental data obtained nowadays in chemistry are essentially multivariate. For example, spectra or chromatograms contain hundreds or thousands of pairs of points, each wavelength or retention time representing an independent variable. The classic approach of chemical analysis is to treat data as univariate and interpret only one or two scales simultaneously. The access to modern computer controlled equipment has enabled chemists to achieve massive amounts of data (spectra, chromatograms, electropherograms, densitograms, etc.) in a relatively short time, and the interpretation of these data must be done by suitable statistical techniques.

The chemometric techniques have also found special applications in chemical experiment design. Thus, in case of screening tests, such techniques allow the elimination of factors (variables) which do not influence the investigated samples (or have an extremely low influence on them). The optimization of methods can be made faster through a multivariate analysis. These techniques also save time and money in the analysis of quantitative structure - property/activity relationships (QSAR/QSPR) for multicomponent systems. Last but not least, chemometric techniques allow quantitative modelling experiments in which many factors are involved. The chemometric methods have been already extensively applied in industries like petroleum, food and pharmaceutical, continuing to expand to new areas like authentication, classification and fingerprinting of different samples.

Chapter III – Experimental setup, results and discussions

1. Validation of a quantitative analysis method for catecholamines and their metabolites using high performance thin-layer chromatography

Results and discussion

Optimization of the TLC – image analysis procedure

The chromatographic investigation of the group of selected compounds raises several difficulties related to the separation resolution (when different reversed-phase layers were applied for their chromatography in preliminary experiments) and quantification after the derivatisation with DPPH$^{\bullet}$. To obtain a satisfactory resolution and symmetric spot shapes of the compounds, different chromatographic conditions such as various stationary phases (RP-18, RP-18W, CN and Diol), different compositions of the mobile phase (methanol : phosphate buffer (pH = 7.10) – 40:60 v/v and 20:80 v/v; acetonitrile : formic acid - 15:85 v/v; citrate buffer (pH = 3.00) : methanol – 96:4 v/v and citrate buffer (pH = 3.00) : methanol : formic acid - 96:4:10 v/v/v) and various DPPH$^{\bullet}$ spraying conditions, were investigated through several trials. As a result, a good chromatographic separation (retention factors: $R_{F(NMN)}$ = 0.63, $R_{F(MN)}$ = 0.46, $R_{F(3\text{-}MT)}$ = 0.38, $R_{F(DOMA)}$ = 0.84; $R_{F(E)}$ = 0.65; $R_{F(VMA)}$ = 0.59; $R_{F(DOPAC)}$ = 0.47; $R_{F(HVA)}$ = 0.22) and a sensitive detection was obtained using LiChrospher® RP-18 WF_{254S} HPTLC plates and a mobile phase consisted of citrate buffer (pH = 3.00) : methanol : formic acid (96:4:10 v/v/v).

The investigated compounds appeared as yellowish-white spots, produced by bleaching the purple color of the DPPH$^{\bullet}$ reagent. An increased concentration of the compounds resulted in more intensive area around the bleached spot on the TLC plate. While the separated spots on TLC plate were detected under visible light, the reaction with DPPH$^{\bullet}$ is very time-instable (the background starting to fade after 10 minutes from staining) and slit-scanning densitometry of such plates is almost unavailable because it brings unpredictable results. For this reason, the obtained chromatograms were scanned by a specialized flatbed scanner and also photographed with a digital camera.

Because the results of TLC - DPPH• method strongly depend on time that elapses between the background staining and data acquisition, in both cases the images were acquired every two minutes after spraying. Digital processing of captured images, optical density integrity, curves drawing, and peaks area calculation were made using the ImageDecipher-TLC image processing program in all cases. The best results were obtained using the images captured after four minutes. After this period a poor peak shape as well as decreased peak area and increased interference from the background was observed. In addition, the background strongly starting to fade after 10 minutes from staining.

Some problems related to the noise removal (present in both videoscans and photographs) and baseline drift (accentuated in case of photographed images due to the inhomogeneous illumination) corrections were encountered. In addition, in case of TLC combined with DPPH• application, image processing of the chromatograms denotes the investigating compounds as negative peak area (Figure 3).

While many complicated solutions have been reported in literature [6], the advantages of the ImageDecipher-TLC software solved the inconveniences related to the noise removal, baseline drift correction and negative peak area. This software allows some image processing operations as to invert images and also to select the colour (red, green, blue or grey) channel for analysis.

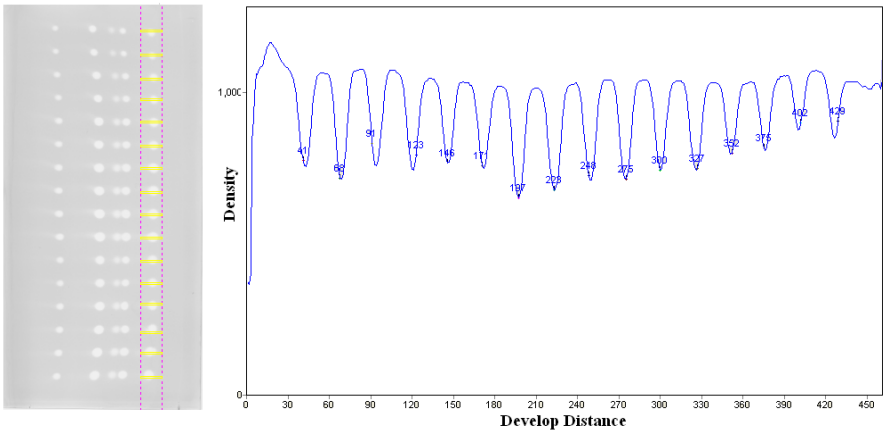
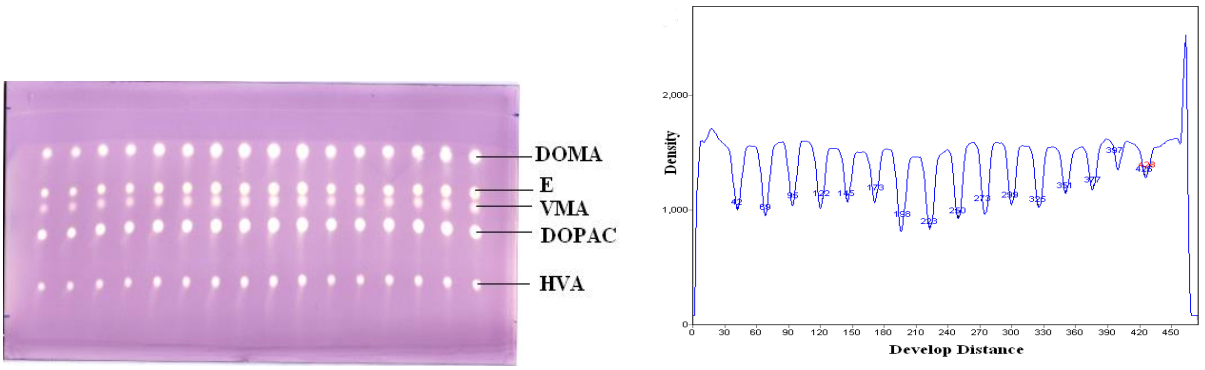
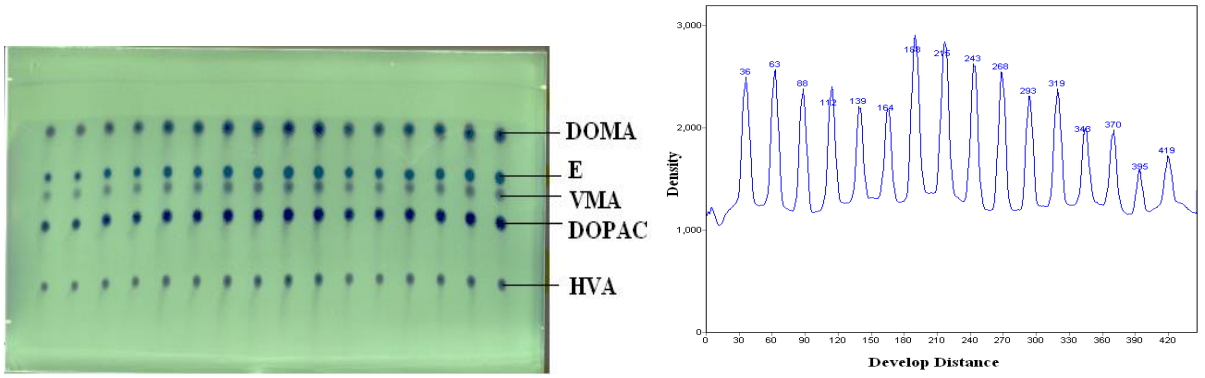


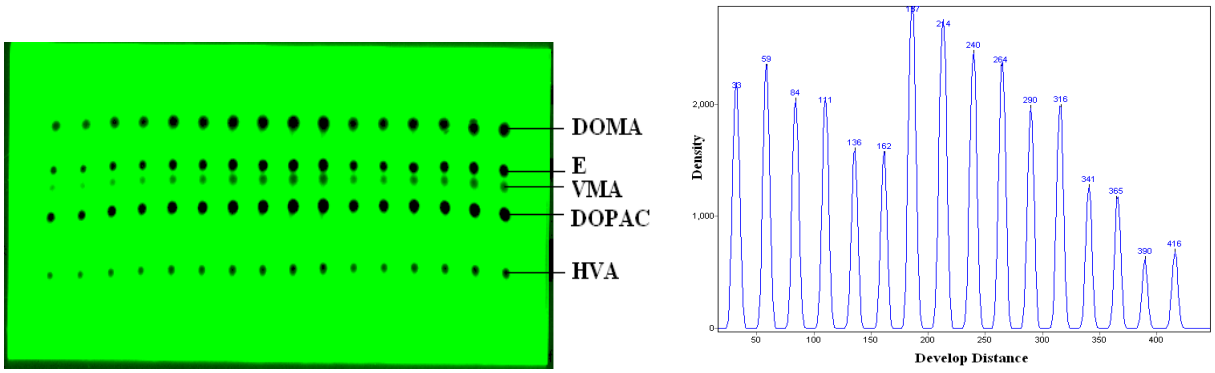
Figure 3 Example of quantitative integration (negative values) of selected spots area using ImageDecipher-TLC software in the case of unprocessed chromatogram.



(a)



(b)



(c)

Figure 4 Image of the chromatographic plate after the spot detection with DPPH• radical and the corresponding chromatogram obtained for DOMA quantification (integrated peak area) with ImageDecipher-TLC software in (a) visible mode; (b) inverted mode; (c) green scale.

To get more accurate integrated peak area, the saved colourful RGB image of the plate (consisted of yellowish-white spots on a purple background) was inverted and the pure colour and grey channel results were investigated. Symmetric peak shapes and accurate quantification of the chromatographic spots (accurate peak area versus applied concentration) with a positive value for the peak area were obtained by inverting the RGB images and conversion into pure green colour scale (Figure 4).

Method validation

The HPTLC method was validated in terms of linearity, precision, accuracy, limits of detection and quantification, for the mixture of HVA, VMA, DOMA, DOPAC and E and NMN, MN and 3-MT, respectively.

For the calibration procedure increasing volumes of the stock solution were applied on the chromatographic plate and the calibration function was constructed for the investigated compounds, by plotting the measured peaks area versus applied amount of compound.

The good linearity in the corresponding concentration range was evaluated by the linear regression equations and the values of the coefficient of determination (R^2) presented in Table 1 and Table 2 The limit of detection (LOD) and the limit of quantification (LOQ) were calculated based on the confidence bands of the calibration function in all cases.

Table 1 Method evaluation parameters for HVA, VMA, DOMA, DOPAC and E

Compound	Linearity range (ng/spot)	Regression equation	R^2	LOD (ng/spot)	LOQ (ng/spot)
DOMA	30-150	y = 87.33x - 2598	0.9986	33	36
E	30-150	y = 57.19x - 1817	0.9966	36	42
VMA	90-450	y = 12.64x - 1133	0.9965	103	120
DOPAC	30-150	y = 91.49x - 875	0.9985	13	17
HVA	60-300	y = 23.92x - 893	0.9978	45	54

Table 2 Method evaluation parameters for NMN, MN and 3-MT

Compound	Linearity range (μg/spot)	Regression equation	R^2	LOD (μg /spot)	LOQ (μg /spot)
NMN	0.10-1.30	y = 11077x + 504.91	0.9973	0.048	0.093
MN	0.10-1.30	y = 9857.8x + 598.49	0.9975	0.046	0.090
3-MT	0.10-1.30	y = 12911x + 3467.6	0.9977	0.045	0.087

The precision of the method, characterized as intra-day and inter-day precision, was expressed as relative standard deviation (RSD %) and determined at three concentration levels. The intra-day precision was assessed by analyzing six replicate spots for each concentration and the inter-day precision was assessed by performing the analysis by the same analyst during a period of 5 days.

The accuracy of the method, expressed as recovery, was investigated for standard solutions by analysing 3 replicate spots for each of the compounds, at three concentration levels. In this case the results (Tables 3 and 4) were included in range of 94.68% and 105.70% for HVA, VMA, DOMA, DOPAC and E and between 99.13%–106.59%, for NMN, MN and 3-MT, respectively.

Table 3 Accuracy and precision of the method for HVA, VMA, DOMA, DOPAC and E

Compound	Added amount (ng/spot)	Intra-day precision RSD (%)	Inter-day precision RSD (%)	Mean recovery (%)
DOMA	50	1.75	1.53	101.24
	75	1.31	1.20	99.27
	100	1.27	1.18	94.68
E	50	2.35	2.16	101.85
	75	1.19	0.99	99.79
	100	1.72	1.67	97.46
VMA	150	2.95	2.75	105.18
	200	1.35	1.31	103.00
	300	1.51	1.41	100.04

DOPAC	50	1.06	1.08	105.70
	75	1.25	1.20	99.49
	100	1.01	0.91	97.80
HVA	100	2.04	2.15	105.53
	150	1.86	1.60	103.93
	200	1.12	1.03	97.32

Table 4 Accuracy and precision of the method for NMN, MN and 3-MT

Compound	Added amount (μg/spot)	Intra-day precision RSD (%)	Inter-day precision RSD (%)	Mean recovery (%)
NMN	0.30	4.67	4.61	106.59
	0.70	2.57	2.34	103.34
	1.10	1.98	1.81	100.87
MN	0.30	4.59	2.69	105.43
	0.70	3.72	3.57	103.02
	1.10	1.99	1.98	99.13
3-MT	0.30	2.83	1.94	106.54
	0.70	1.52	1.36	104.31
	1.10	1.13	0.93	100.91

Application of the method to human urine sample analysis

The applicability of the proposed TLC – image processing method to determine the level of acidic metabolites of catecholamine in human urine sample was also investigated. The chromatographic separation, $DPPH^{\bullet}$ derivatisation and image processing analysis for spiked urine samples were carried out under the same conditions described for the standard mixtures. The specificity of the method was determined in relation to the interferences from other compounds in the urine samples.

- ***Determination of HVA, VMA, DOMA, DOPAC in biological sample***

All the acidic metabolites (HVA, VMA, DOMA, DOPAC) were confirmed by a good separation resolution on the basis of their retention by comparison with the spots of the standards

samples. Some interfering compounds detected as unidentified spot in the urine chromatogram ($R_{F(\text{unidentified peak})}$ = 0.64) were observed (Figure 5).

Due to the retention value of the unidentified interfering compounds, we were unable to quantify epinephrine ($R_{F(E)}$ = 0.65) in spiked urine samples using the chromatographic system mentioned above.

A matrix effect study and recovery test of the developed method on real samples was also performed. Spiked urine samples at three different concentration levels (60ng/spot, 120ng/spot and 180 ng/spot for DOMA and DOPAC; 120 ng/spot, 240 ng/spot and 360 ng/spot for HVA; 180 ng/spot, 360 ng/spot and 540 ng/spot for VMA respectively) were prepared by adding appropriate standard mixtures to urine sample.

The matrix effect was investigated by comparing spot area of the fortified urine sample with the spot area of standard solution at the same concentration using above mentioned three concentration levels for six replicate spots in all cases.

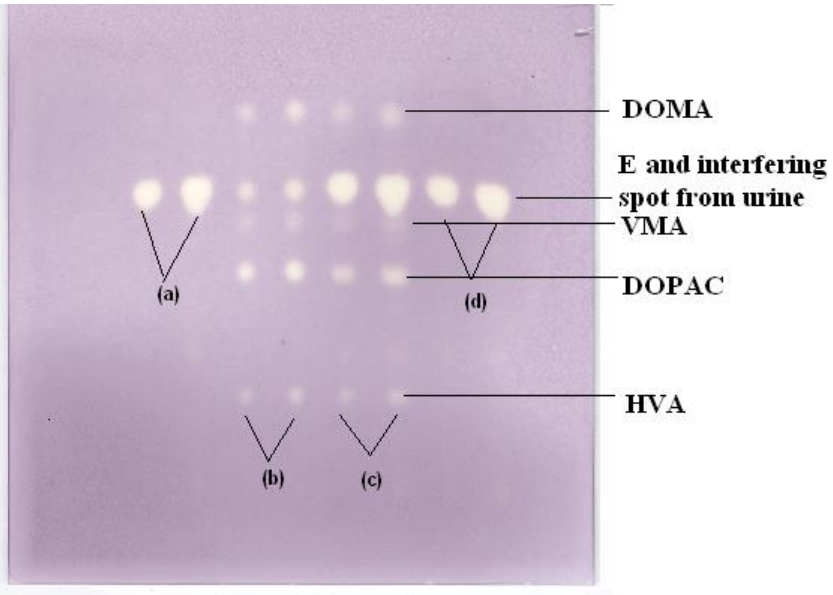


Figure 5 Image of the chromatographic plate presenting the analysis of urine and spiked urine samples: (a) female urine sample (3µL and 5µL); (b) standard mixture (3µL and 5µL); (c) spiked urine sample (3µL and 5µL); (d) male urine sample (3µL and 5µL).

The obtained results indicated the urine sample as a blank sample, the level of the endogenous metabolites in urine being under detection and quantification limit of the developed method. The results also reflect that matrix effects of urine are negligible after two times dilution.

Using the spiked urine samples, the amount of each compound calculated by the linear regression equation was compared to the fortified amount and the recovery rate of the method in urine was obtained. The results (Table 5) show recovery rates between 98% and 108% for all of the investigated metabolites.

Table 5 Results concerning the recovery of studied compounds from spiked urine samples

Compound	Fortified value (ng/spot)	Found amount (ng/spot)	Mean recovery (%)
DOMA	60	65	108
	120	122	102
	180	190	106
VMA	180	191	106
	360	354	98
	540	544	101
DOPAC	60	65	108
	120	122	102
	180	184	102
HVA	120	130	108
	240	242	101
	360	355	99

- ***Determination of NMN, MN and 3-MT in biological sample***

The chromatographic plates were spotted with standard mixture of the investigated compounds, in order to obtain a calibration curve, and also with urine and spiked urine. As it can be observed in Figure 6 MN and 3-MT were confirmed by a good separation resolution on the basis of their retention by comparison with the spots of the standards samples, but at the same R_F

as of NMN (R_F = 0.63) it can be observed that an interfering compound is present in the urine sample.

Based on the fact that our method cannot detect normetanephrine in urine of healthy subjects, another chromatographic system was employed to demonstrate this theory. Thus, standard solutions of metanephrine, normetanephrine, urine and spiked urine samples were spotted on TLC-silica gel 60 chromatographic plates and then they were developed using a mobile phase consisting of phosphate buffer : methanol 80:20 (v/v).

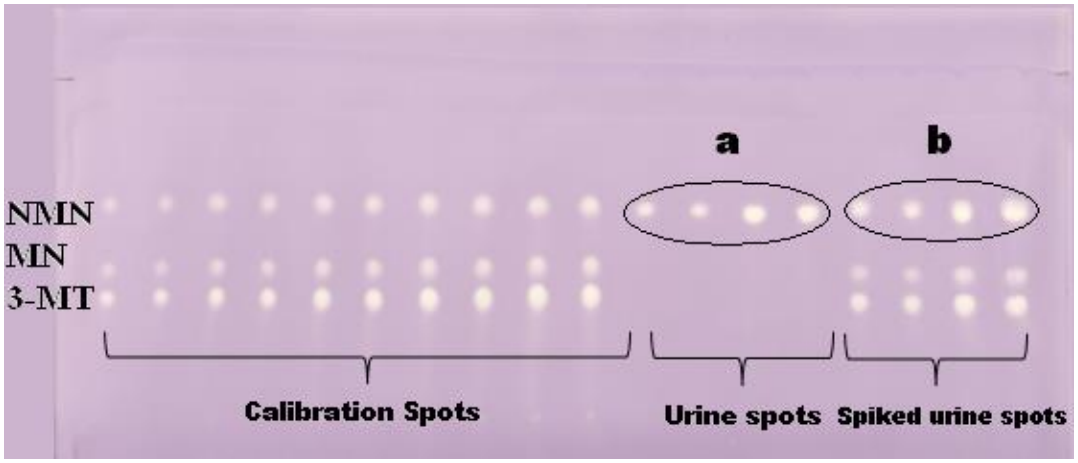


Figure 6 Image of the chromatographic plate presenting the analysis of urine and spiked urine samples of NMN, MN and 3-MT

In Figure 7 is shown the plate described before, after spraying it with 0.02% DPPH ethanolic solution, and it can be observed that the investigated compounds are no longer separated, but the unidentified compound is now well separated from the rest, thus demonstrating that it isn't normetanephrine.

Based on the linear regression equation obtained for all the investigated compounds, we were able to quantify the amount of metanephrine and 3-methoxytyramine directly using the area of the respective compound separate from the spiked urine sample and in case of

normetanephrine the quantification was made using the area obtained by the difference between the area of the spiked urine spot (Figure 6 – spots a) and the area of the unidentified compound (Figure 6 – spots b). The results (Table 6) show recovery rates between 97.96% and 103.48% for all of the investigated compounds.

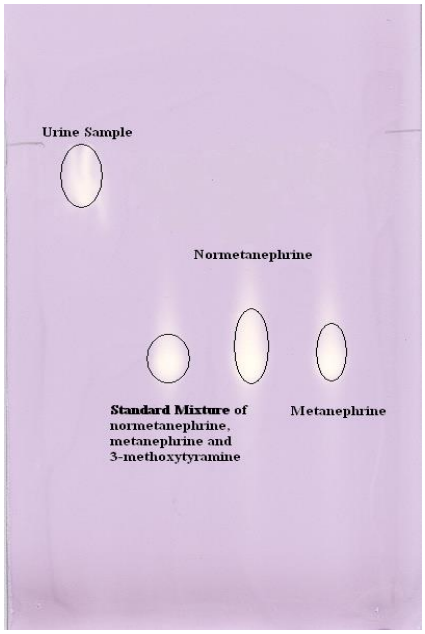


Figure 7 Image of the plate developed in modified chromatographic conditions

Table 6 Recovery results of spiked urine sample

Compound	Fortified value (μg/spot)	Found value (μg/spot)	Mean recovery (%)
NMN	0.400	0.394	98.49
	0.800	0.784	97.96
MN	0.400	0.405	101.24
	0.800	0.797	99.59
3-MT	0.400	0.419	104.86
	0.800	0.828	103.48

2. QSAR studies and lipophilicity determinations

2.1 Prediction of catecolamines lipophilicity using the Cluj topological indices

Data set and correlating algorithm

In this study a set of 38 catecholamines and related compounds was submitted to a novel QSAR approach based on weighting and alignment of the molecules over a hypermolecule, and prediction of lipophilicity using the Cluj topological indices, defined by Diudea in [7, 8].

In order to develop the QSAR model and to test its applicability, the set catecholamines and related compounds was divided randomly in two groups, 28 molecules for the training set (molecules no.: 1, 2, 4-6, 8-11, 13, 14, 16-18, 20, 22-25, 27, 28, 30-32) and 10 molecules for the test set (molecules no.: 3, 7, 12, 15, 19, 21, 26, 29, 33).

First the molecules were drawn in HyperChem and each of them was optimized at molecular mechanics (MM+) level of theory. The correlating algorithm followed a few steps: (1) generate the hypermolecule; (2) calculate the molecular descriptors; (3) find the best regression equations by correlating the topological indices with the chosen property (logP) and (4) test the predictive capability of the model.

In order to achieve the model, the structure is encoded in a numerical form. The arrangement of substituent groups, on the catecholamine derivatives, can be accounted for by the hypermolecule concept viewed as the union of the molecules forming the correlating space. A binary vector was assigned to each molecule by aligning them over the hypermolecule (Figure 8): 1-for a common feature in a given position of the hypermolecule and 0- for an empty position. Next, the binary vector was weighted by the mass of “hydride” fragments composing each molecule and the weighted vector was used in the data-reduction step and correlation weighting procedure, which are described in detail in recent papers from literature [9-12].

Then the topological indices were calculated using TOPOCLUJ software and the following descriptors are procured into consideration for developing the model: sumative

descriptor (SD), adjacency, connectivity, detour, distance, DS, IE max, IE min, IP max, IPmin and randic.

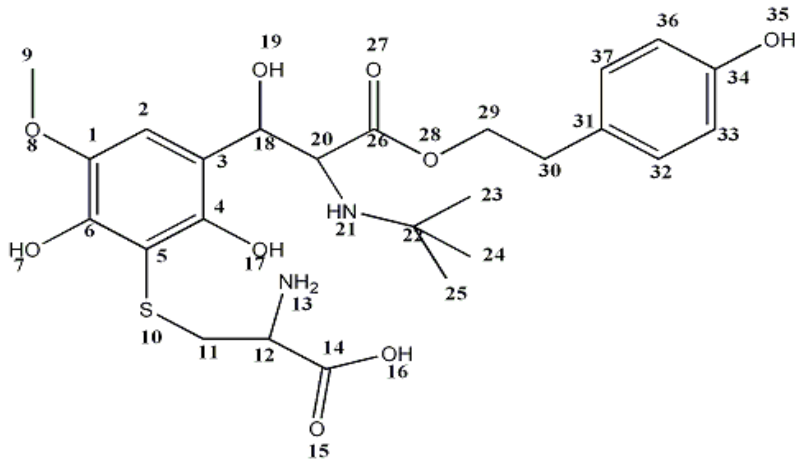


Figure 8 The hypermolecule

Results and discussion

QSAR models – by multivariate regression

The models were developed using the molecules from the training set and the best results are listed below (Table 11). At first, only one-dimensional models were selected. Then the most accurate model was picked up. After that, all two-dimensional models produced by adding a new attribute to the first model were experimented. Again the best model was chosen and supplemented by a new attribute. All the considered regression models must pass a test based on the value of Fisher statistics of all their regression coefficients (R^2).

If a term has the value of R^2 for the regression equation less than a specified threshold, this term is removed from the model. Thus, the process of adding new terms stops either when all attributes are included in the model or when no new term can be added without violating this criterion. The best equation is produced by the system based on square of correlation coefficient (R^2) and the best fit regression is that in which the values for R^2 are closer to 1.

Table 11 Best models in describing log P for the training set of molecules

No.	Descriptors	R^2	Adjust. R^2	St. Error	F
1	***SD***	***0.9118***	***0.9084***	***0.2477***	***268.74***
2	IP min	0.3611	0.3365	0.6666	14.696
3	IP max	0.3145	0.2881	0.6905	11.926
4	IE min	0.2696	0.2415	0.7128	9.5972
5	***SD, Adjacency***	***0.9121***	***0.9050***	***0.2522***	***129.64***
6	SD, Randic	0.9119	0.9048	0.2525	129.36
7	SD, DS	0.9119	0.9048	0.2525	129.35
8	SD, IE max	0.9119	0.9048	0.2525	129.35
9	SD, Distance	0.9119	0.9048	0.2525	129.33
10	SD, Connectivity	0.9118	0.9048	0.2525	129.30
11	SD, Detour	0.9118	0.9048	0.2525	129.29
12	***SD, Adjacency, Randic***	***0.9198***	***0.9098***	***0.2458***	***91.799***
13	SD, Adjacency, Distance	0.9188	0.9087	0.2473	90.544
14	SD, IE max, IP max	0.9178	0.9076	0.2488	89.356
15	SD, Distance, IP max	0.9176	0.9073	0.2492	89.059
16	SD, Detour, IP max	0.9163	0.9059	0.2511	87.595
17	SD, IE min, IP max	0.9148	0.9041	0.2533	85.916
18	SD, IP max, Randic	0.9129	0.9020	0.2562	83.872
19	SD, IP min, Randic	0.9129	0.9020	0.2562	83.826
20	SD, DS, IP max	0.9129	0.9020	0.2562	83.823
21	SD, Connectivity	0.9125	0.9016	0.2568	83.421
22	***SD, Adjacency, Randic, DS***	***0.9250***	***0.9119***	***0.2429***	***70.894***
23	SD, Adjacency, Randic, Connectivity	0.9242	0.9110	0.2441	70.128
24	SD, IP max, Distance, DS	0.9228	0.9094	0.2463	68.776
25	SD, IP max, Distance, Randic	0.9205	0.9066	0.2501	66.545

Model validation

For the applicability of the model, a step of external validation was performed on the best model obtained. Thus, the test molecules were submitted to the model in order to calculate log P using equation no. 22 from Table 11. The results are presented in Table 12 and Figure 9 represents the correlation log P-pred vs. log P-exp.

Table 12 Predicted log P for the test set of molecules using external validation

Compound	LogP- Pred	LogP-Exp
3	1.159	0.973
7	0.774	0.916
12	0.711	0.688
15	-0.598	-0.411
19	0.892	1.049
21	0.193	0.268
26	0.185	0.625
29	-0.094	-0.110
33	0.643	0.590
38	0.826	0.640

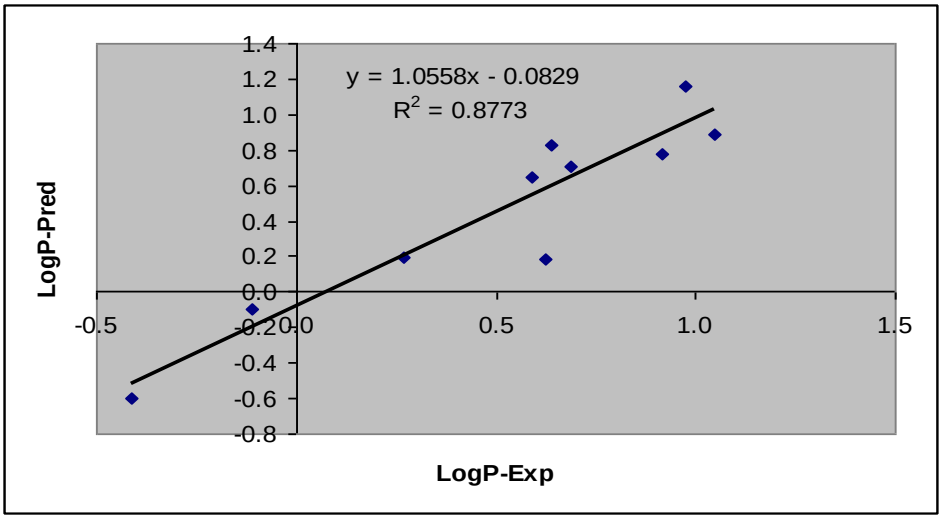


Figure 9 Plot of the regression log P-pred vs. log P-exp

The equations thus obtained were used to predict the values of log P for each molecule from the test set and data are listed in Table 13. It can be observed in Figure 10, that the correlation coefficient of log P-pred vs. log P-exp is far better in this case than in the external validation.

Table 13 Predicted log P for the test set using similarity cluster validation

Compound	LogP- Pred	LogP-Exp
3	1.006	0.973
7	0.744	0.916
12	0.569	0.688
15	-0.659	-0.411
19	0.971	1.049
21	0.267	0.268
26	0.359	0.625
29	-0.068	-0.110
33	0.599	0.590
38	0.820	0.640

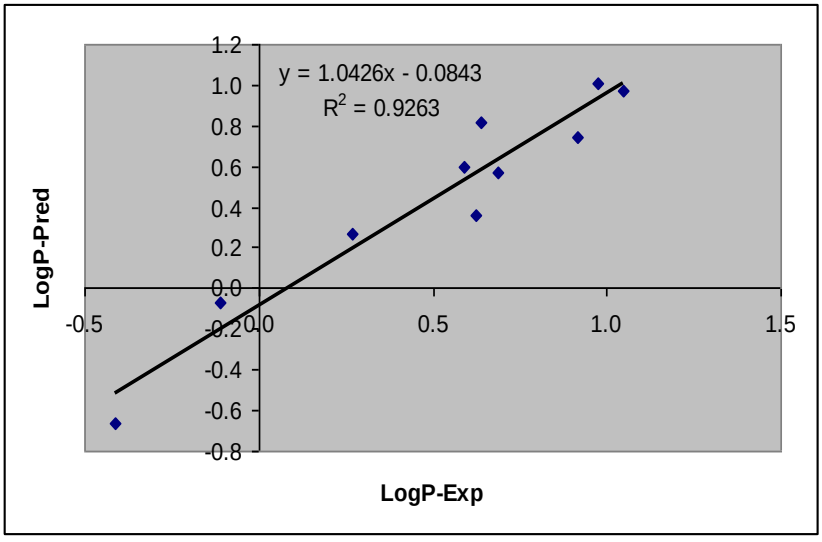


Figure 10 Plot of the regression logP-pred vs. logP-exp

QSAR models – by genetic algorithms

Genetic algorithms (GA) are an evolutionary method widely used for complex optimisation problems in several fields such as robotics, chemistry and QSAR [13-14]. Using the same topological indices (Table 10) QSAR models were generated for describing log P for the training set of molecules, using MobiDigs software, which allows searching for regression models by developing optimal model populations using genetic algorithms. The best obtained models are listed in Table 14.

Table 14 Best models in describing log P for the training set of molecules

No.	Descriptors	R^2	Adjust. R^2	St. Error	F
1	SD, Adjacency, Randic	0.9198	0.9098	0.246	91.77
2	***SD, Adjacency, DS, Randic***	***0.9250***	***0.9119***	***0.243***	***70.88***
3	SD, Adjacency, IE max, Randic	0.9198	0.9059	0.251	65.97
4	SD, Adjacency, Distance, Randic	0.9199	0.9059	0.251	66.00
5	SD, Adjacency, IE min, Randic	0.9200	0.9061	0.251	66.11
6	SD, Adjacency, IP max, Randic	0.9199	0.9060	0.251	66.08
7	SD, Adjacency, Detour, Randic	0.9198	0.9059	0.251	65.99
8	SD, Adjacency, IP min, Randic	0.9199	0.9060	0.251	66.05
9	SD	0.9118	0.9084	0.248	268.62
10	SD, Adjacency, Conectivity, Randic	0.9242	0.9110	0.244	70.11

Model validation

The best obtained model was submitted to external validation in order to test its applicability. Thus, for the test molecules log P was predicted using equation no. 2 from Table 14. The results are presented in Table 15 and Figure 11 represents the correlation log P-pred vs. log P-exp.

Table 15 Predicted log P for the test set of molecules using external validation

Compound	LogP- Pred	LogP-Exp
3	1.194	0.973
7	0.766	0.916
12	0.716	0.688
15	-0.491	-0.411
19	1.039	1.049
21	0.212	0.268
26	0.305	0.625
29	-0.100	-0.110
33	0.660	0.590
38	0.731	0.640

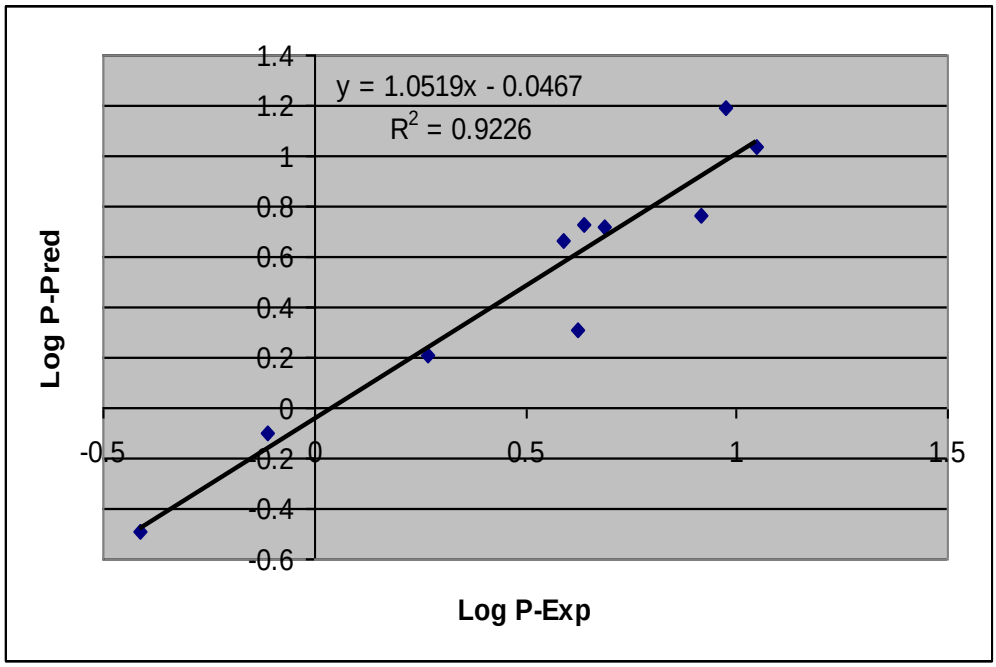


Figure 11 Plot of the regression log P-pred vs. log P-exp

2.2 Assessment of lipophilicity indices of antioxidant compounds in RP-HPLC

Results and discussion

The group of antioxidants investigated in this study includes compounds with very different structures, size and polarity, so it is expected that they have quite different chromatographic behavior. Therefore, the methanol fraction contained in the mobile phase was optimized so that all compounds have retention times between t_0 (dead time) and a maximum of 15 minutes; in order to shorten the analysis time and also to be able to compare the results for different temperatures (22 °C and 37 °C).

Thus the fraction of methanol, for which a linear range was obtained for logk, ranged between 50 % - 60 % for the RP18 and CN columns, 60 % - 70 % for the C8 and C16-Amide columns, and 55 % - 65 % for the PFP column; and in all cases an increment of 2.5 % was used to obtain the specified 5 concentrations. The strong linear dependence of retention times on the methanol fraction was demonstrated by the values of coefficient of determination (R^2) higher than 0.99 in all cases.

Furthermore, by evaluating the profiles of k and logk values for all methanol fractions determined for both 22 °C and 37 °C, the regular changes in retention with increasing methanol ratios were observed in the case of C8, C16-Amide, PFP (except compound 22) and CN column, except RP18. In the case of the four columns, the mk and mlogk parameters were overlapping the intermediate (median) value corresponding to the middle concentration of methanol.

All the specific chromatographic lipophilicity parameters (arithmetic mean of k and logk- mk and mlogk, logkw, S, ϕ_0, scores corresponding to the first principal component obtained by applying PCA to the retention data - PC1/k and PC1/logk) were calculated and considered for all investigated columns at 22 °C and 37 °C. By a summary evaluation it can be observed that at 22 °C pterostilbene (19) has the highest lipophilicity index for the C8, C16-Amide and CN columns, pelargonidin (16) for the RP18 column and procyanidin C1 for the PFP column, while at 37 °C pterostilbene (19) has the highest lipophilicity index for the RP18, C8, and C16-Amide columns, pelargonidin (16) for the CN column and apigenin (16) for the PFP column. Also, the lowest lipophilicity index at 22 °C was found for epigallocatechin gallate (15)

on RP18 column, procyanidin C1 (22) on C8 column, protocatechuic acid (5) on C16-Amide and PFP columns, and chlorogenic acid (11) on CN column, while at 37 °C the lowest lipophilicity index was found for catechin (9) on RP18 and C16-Amide columns, and procyanidin C1 (22) on C8, CN and PFP columns.

In order to see how the temperature affects the lipophilicity we will refer only to the indices $logk_w$ and mlogk. First, matrices of correlation between the data obtained at 22 °C vs. 37 °C for all columns including also the computational lipophilicity values were calculated. Accordingly, it can be observed, considering firstly experimental $logk_w$ values for the two temperatures, the higher correlations were obtained for C16 (r = 0.969), C8 (r = 0.983) and CN (r = 0.828). A low correlation was obtained for RP18 (r = 0.463), and surprisingly a very low negative value resulted for PFP (r = -0.042). The statistical results concerning the computational lipophilicity descriptors indicate that at 22 °C the highest correlation were obtained on PFP (r = 0.918 with NCNHET, r = 0.873 with XLogP, and r = 0.855 with ALOGP2) and CN (r = 0.800 with CLogP and r = 0.620 with MLOGP). On the other hand, at 37 °C the best correlations were obtained on CN column (r = 0.533 with ALOGP98) and RP18 (r = 0.504 with CLogP).

A high correlation resulted also for RP18 column vs. Average value (r = 0.906) calculated for all experimental and computational data corresponding to each investigated compound; this value is used also in the Heberger algorithm [15-18] as it will be discussed below. In addition, the results illustrate a significant correlation between the results obtained on all columns (with some exceptions in the case of PFP and RP18) and the following computational descriptors: CLogP, MLOGP and Average.

The statistical evaluation of the correlation results considering the experimental data estimated as mlogk and also the computational indices showed that there is a high correlation between all experimental lipophilicity indices at the two temperatures, excepting the correlations between RP18 and CN (22 and 37 ^oC; r = 0.342 and r = 0.239), PFP at 37 ^oC (r = 0.358) and C16 at 37 ^oC (r = 0.384). A significant correlation has been observed between the mlogk values and CLogP ($0.525 < r < 0.723$), MLOGP ($0.423 < r < 0.679$).

A significant correlation can be observed (with some exceptions) in the case of Average, ALogP98 and XLogP2. In addition, the correlation between mlogk values at 22 ^oC and 37 ^oC for PFP becomes highly significant (r = 0.938). The large difference between the

correlation coefficients obtained for $logk_w$ and mlogk at the two temperatures in the case of PFP column can be clearly explained by the effect of extrapolation in the first case.

Moreover, the effect of temperature on the considered chemically bonded columns and the chromatographic behavior of the investigated compounds is clearly illustrated by box and whisker plot depicted in Figure 16. The larger difference is observed in both cases on the RP18 and PFP columns and the smaller effect on C16 and CN, two columns with higher polarity. Moreover considering the mlogk values a distinct difference is shown between the nonpolar C8 and C18 columns (positive effect) and the CN, C16-Amide and PFP (negative effect in order CN < C16-Amide < PFP).

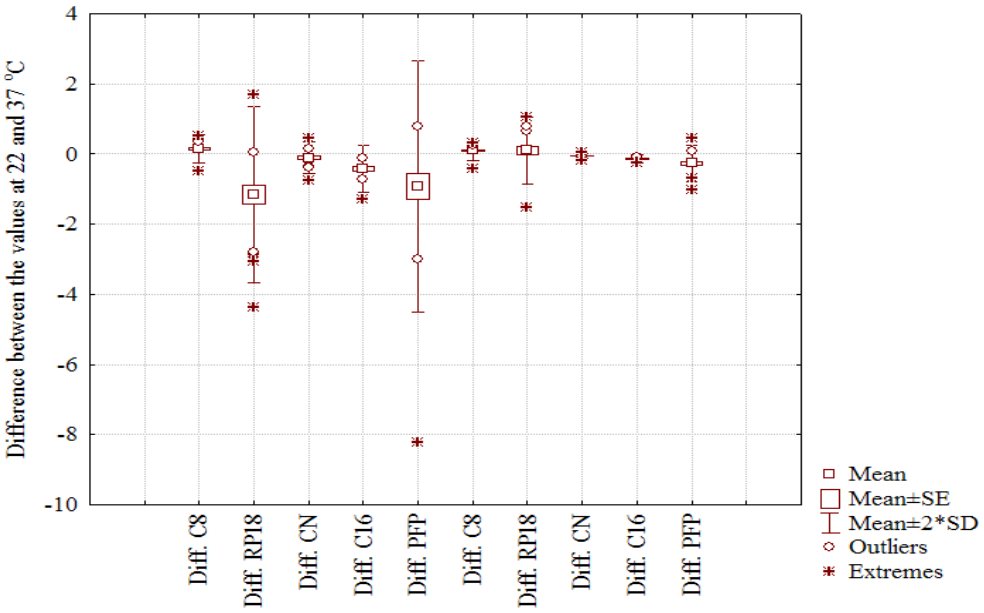
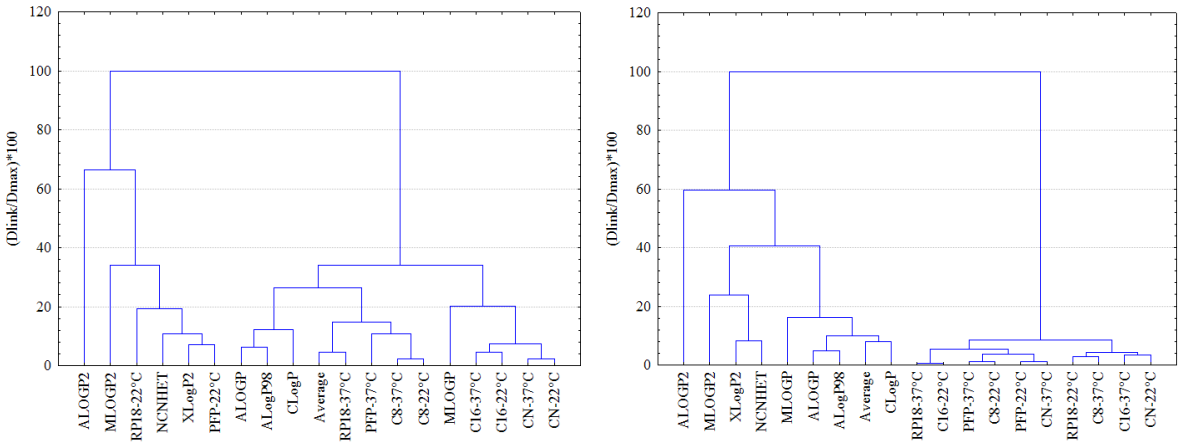


Figure 16 Box and wiskers corresponding to $logk_w$ values, (the first five boxes, from left to right) and mlogk values, respectively (the last five boxes)

The discrepancies observed in the case of $logk_w$ values can be explained once again by the effect of extrapolation and the different chromatographic behavior of some compounds (13, 16, 18, 19, 22). The statements above are well supported by the results obtained applying classical hierarchical cluster analysis (HCA) and PCA on the standardized datasets. The dendrogram obtained in the case of dataset including experimental $logk_w$ and computationally indices illustrates three well separated clusters (Figure 17a).



(a) (b)

Figure 17 Hierarchical cluster analysis dendrogram showing similarities among different chromatographic indices and computationally logP values: (a) $logk_w$, (b) mlogk

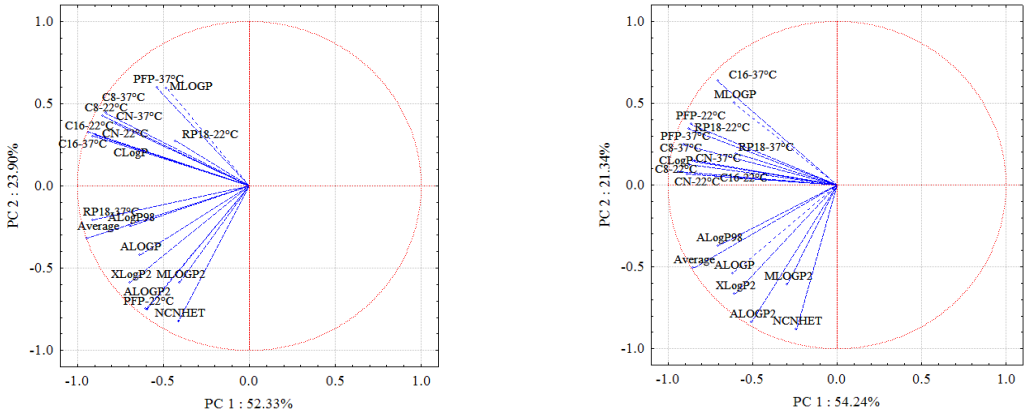
The $logk_w$ corresponding to CN and C16 columns at the two temperatures, including MLOGP, are in first group, the second combines the $logk_w$ obtained on C8 at the two temperatures, PFP and RP18 at 37 ^oC and some computational indices (ALOGP, ALogP98, ClogP and Average). The third cluster includes $logk_w$ corresponding to PFP and RP18 at 22 ^oC and XLogP2, NCNHET, MLOGP2 and ALOGP2. If the mlogk values are considered, a clear distinction between computationally estimated logPs and chromatographic indices is obtained. The high similarity of the mlogk is also clearly shown (Figure 17b).

Applying PCA on the $logk_w$ values, the first principal component explains 52.33 % of the total variance and the second component 23.90 %: a two component model thus accounts for 76.23 % of the total variance. The results from the PCA of mlogk values are slightly different.

The first two PCs account for 75.58 % of the total variance (PC1 54.24 % and PC2 21.34 %). The patterns obtained by two-dimensional representations of the loadings are more or less similar with the HCA-patterns discussed above. In the case of $logk_w$ (Figure 18a) two groups are clearly separated. The first includes the majority of the experimental $logk_w$ indices and two computational scales (ClogP and MLOGP), in the second group two $logk_w$ (RP18-37 ^oC and PFP-22 ^oC) appear in the vicinity of other computational scales. Two major groups are present also in the case of mlogk dataset. The first group includes all the mlogk indices and two

computational scales (CLogP and MLOGP) and in the second group we find only computational scales (Figure 18b).

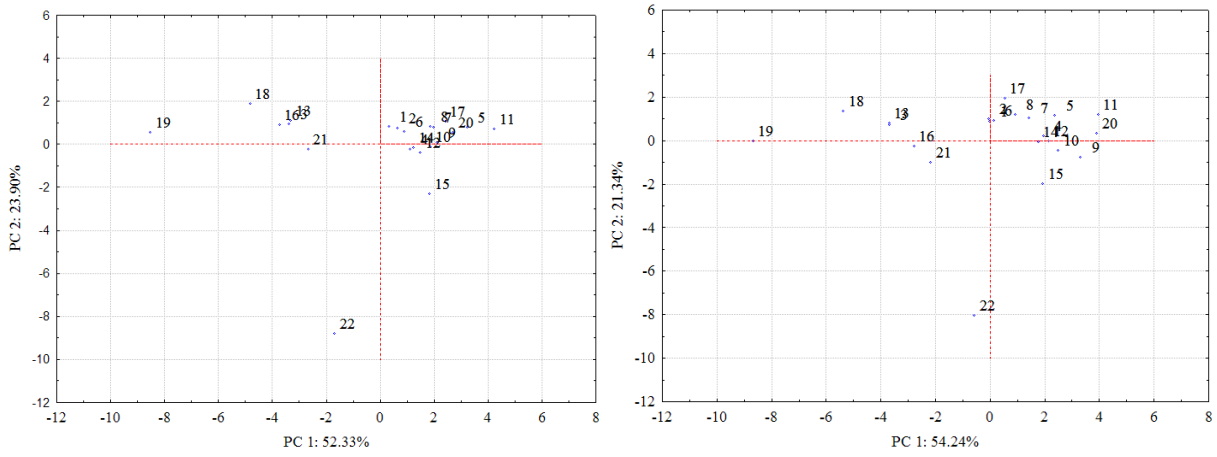
At the same time, the lipophilic character similarities existing between the investigated compounds may be illustrated by the lipophilicity charts ("holistic lipophilicity chart") obtained by 2-D scatterplots of the scores corresponding to the first two principal components. The score plots (Figure 19a-b) reveal two groups (more compacted in the case of $logk_w$) and identify two outliers: pterostilbene (19) and C1 type proanthocyanidin (22).



(a) (b)

Figure 18 Scatterplot of loadings corresponding to the first two PCs (similar lipophilicity indices are positioned close to each other): (a) $logk_w$, (b) mlogk

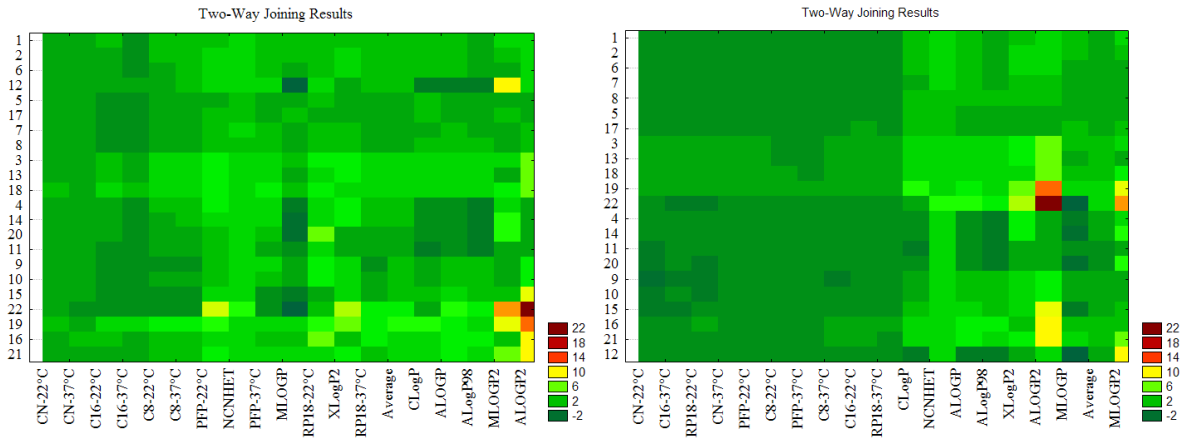
Two-way joining cluster analysis applied on a dataset formed by the $logk_w$ and mlogk values obtained for all compounds on all investigated columns at the two temperatures including also the computationally calculated indices provides similar conclusions regarding the effect of temperature and the chromatographic behavior of the compounds investigated (Figure 20a). The most similar results, considering $logk_w$ values and the computational scales, for example, are easily observed in the case of CN, C8 and also C16 at the two temperatures (green color), and it is also clearly pointed out the outlier position of the C1 type proanthocyanidin (22) (yellow color).



(a) (b)

Figure 19 Scatterplot of scores corresponding to the first two PCs (similar compounds are positioned close to each other in two distinct groups: (a) $logk_w$, (b) mlogk

The pattern in the case of mlogk values including also the computational scales illustrates a high similarity among all experimentally indices and CLogP, ALOGP and Average appear to be closer to them (Figure 20b).

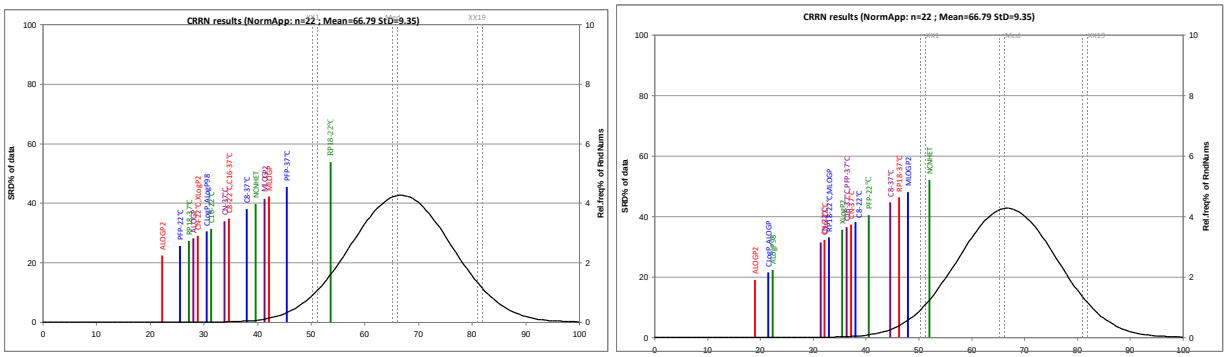


(b)

Figure 20 Two-way joining clustering of $logk_w$ (a) and mlogk (b) including computationally logP values for all investigated columns and both temperatures

In order to get more information and a better understanding of the experimental and computational estimation of lipophilicity we applied also a new non-parametric ranking method, sum of ranking differences-comparison of ranks by random numbers (SRD-CRRN) [15-18]. According to the SRD-CRRN, considering first the $logk_w$ values and computational scales, the best descriptors are obtained using PFP-22 ^oC, RP18-37 ^oC, CN-22 ^oC and C8–22 ^oC including ALOGP2 (the best), ALOGP and CLogP. Lower ranking values were obtained in the case of RP18–22 ^oC, PFP-22 ^oC, and MLOGP and MLOGP2 (Figure 21).

In the case of the dataset comprising mlogk values and calculated LogP values the results presented also in Figure 21 indicate ALOGP2, CLogP, ALOGP as the best computational scales followed by two groups of lipophilicity measures: (CN, C16 and RP18 at 22 ^oC and MLOGP) and (XLogP2, C16 and PFP at 37 ^oC, CN-37 ^oC and C8-22 ^oC). The farthest group includes C8 and RP18 at 37 ^oC, and MLOGP and NCNHET) and they be considered as the worst lipophilicity measures.



(a) (b)

Figure 21 SRD-CRRN Ranking of chromatographically estimated lipophilicity indices $logk_w$ (a) and mlogk (b), and computationally calculated logP values

3. Fingerprinting and authenticity determination of wild fruits and derived dietary supplements

3.1 Fingerprinting of Romanian wild fruits

Results and discussion

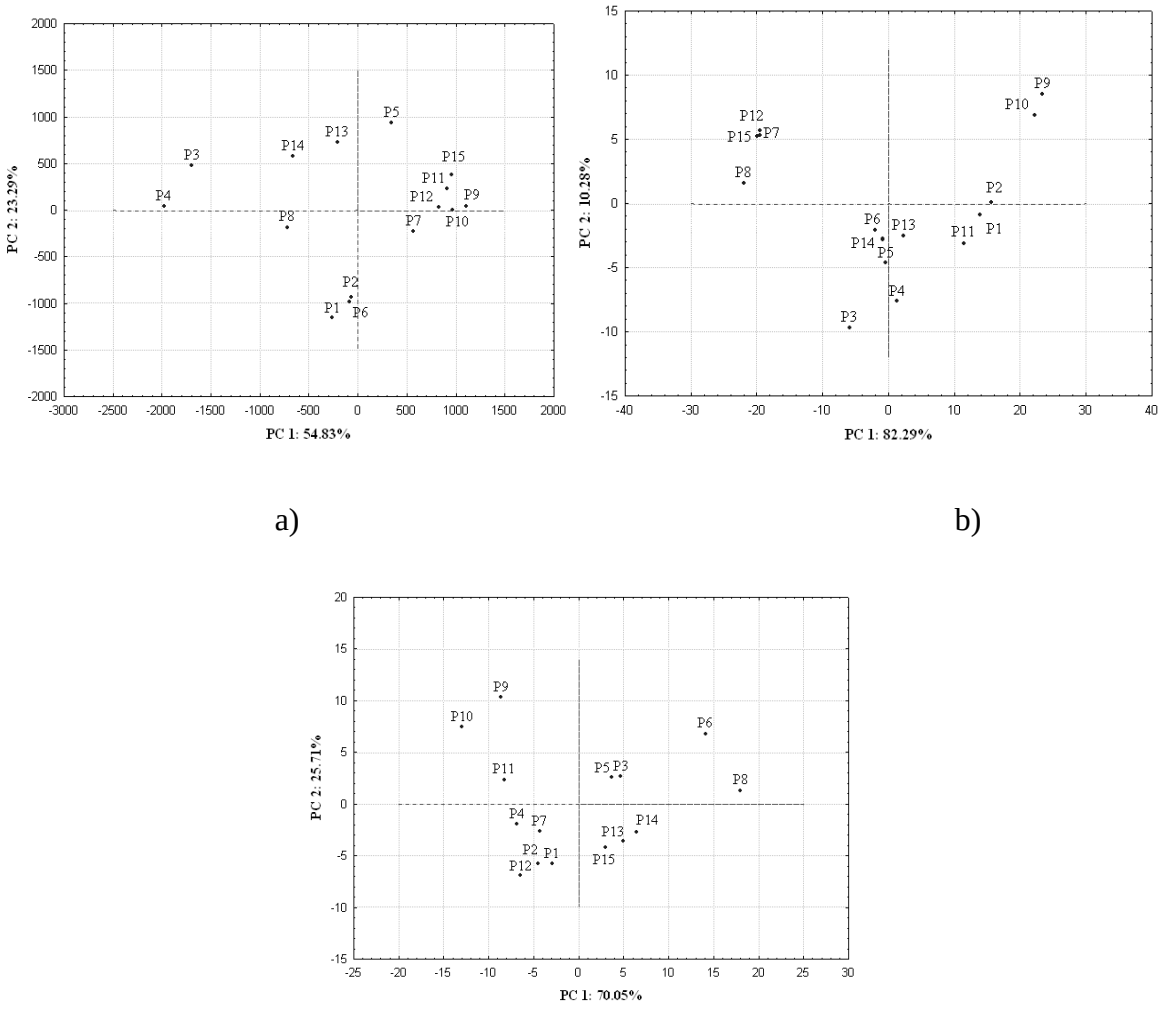
Classification of samples

TLC Analyzer 1.1 software was used to obtain the digital chromatograms from the images of the TLC plates. A plot of the brightness values versus scan distance (in pixels) is obtained and this is what the variables represent. In case of the UV and UV-Vis spectra the Spectra Manager software was used to digitize them and the variables represent the absorbance versus wave length.

The highly complex data obtained using HPTLC and UV-Vis spectrometry (digital chromatograms and digitized UV spectra, respectively) cannot be managed or handled by the simple visualization of the data matrix, therefore to be able to see trends and to compare samples, the chemometric analysis was required.

2D plots of PC1 vs. PC2 (Figure 23a-c) were obtained applying principal component analysis (PCA) to data matrices of digital TLC chromatograms (15 samples x 1000 variables), UV spectra profiles (15 samples x 400 variables) and UV-Vis spectra of DPPH• reduction profiles (15 sample x 1000 variables). Also, cluster analysis (CA) was applied to the same data matrices and dendograms were obtained using the tree single-linkage clustering (Figure 24a-c).

As it can be observed, the chemometric analysis, both principal component analysis and cluster analysis, managed to classify the samples in a comparable way regardless of the analytical technique used. The major clusters obtained using TLC were: *C1* – blackthorn (P3, P4); *C2* - sea-buckthorn (P1, P2); *C3* - cornelian cherry (P13, P14) and rose hip (P8); *C4* – bilberry (P5), rose hip (P6, P7), raspberry (P9, P10), cranberry (P11, P12) and blackberry (P15).



a) b)

c)

Figure 23 2D plots of PC1/PC2 obtained from data matrices of a) digital TLC chromatograms; b) UV spectra profiles and c) UV-Vis spectra of DPPH• reduction profiles

In case of UV spectrometry the clusters were: *C1* – rose hip (P7, P8), blackberry (P15), cranberry (P12); *C2* - raspberry (P9, P10); *C3* - blackthorn (P3, P4); *C4* – bilberry (P5), rose hip (P6) and cornelian cherry (P13, P14); *C5* - sea-buckthorn (P1, P2) and cranberry (P11).

And for the DPPH• reduction profiles the clusters were: *C1* – rose hip (P6, P8); *C2* - raspberry (P9, P10); *C3* – blackthorn (P3) and bilberry (P5); *C4* – cornelian cherry (P13, P14) and blackberry (P15); *C5* - sea-buckthorn (P1, P2), blackthorn (P4), rose hip (P7), and cranberry

(P12). The extracts were generally grouped depending on the fruit from which they were prepared, this suggesting that there is no significant difference in the content of bioactive compounds in fruits purchased from cultivators or natural sources.

Also, the chemometric methods were able to detect a small difference between the composition of sea-buckthorn (P1, P2), blackthorn (P3, P4), raspberry (P9, P10) and cornelian cherry (P13, P14), as they were grouped, in most cases, in small clusters of their own, apart from the larger group containing the remaining samples (which were classified more or less in the same group).

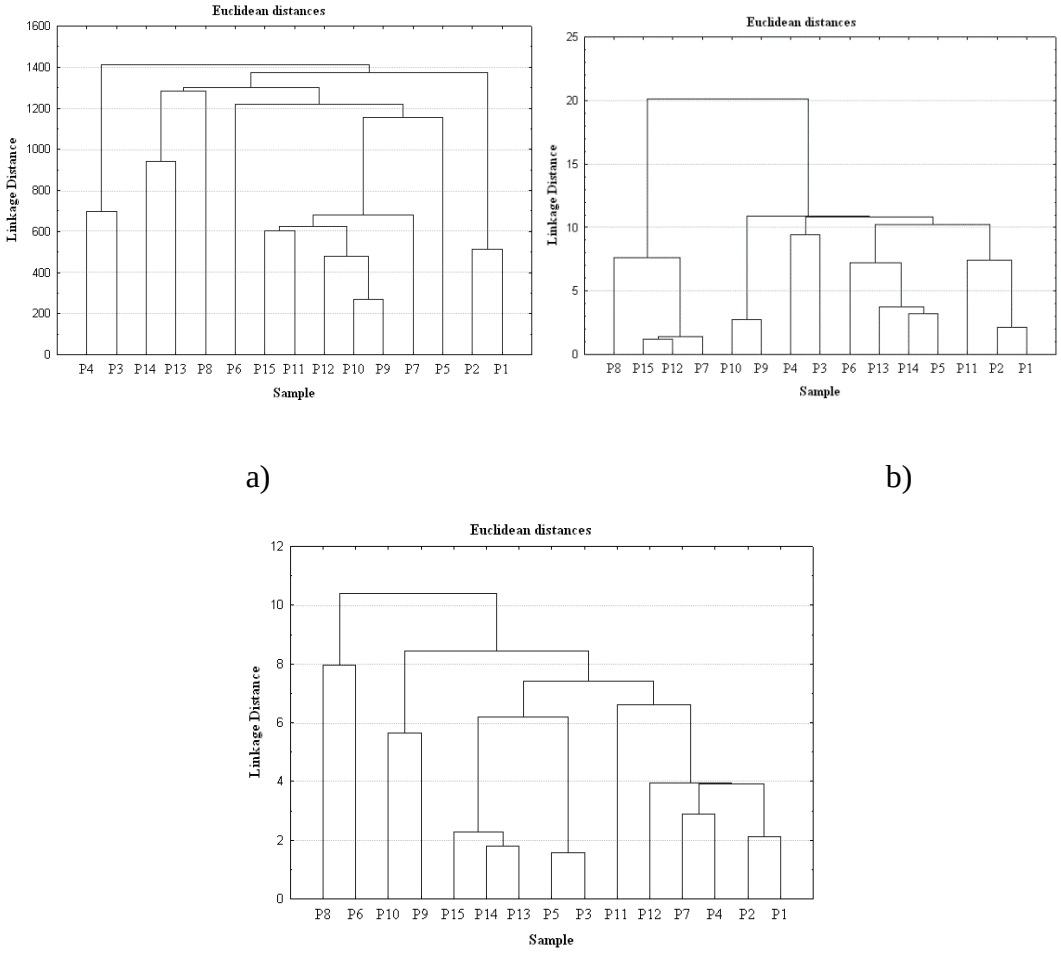


Figure 24 Cluster analysis dendograms obtained from data matrices of a) digital TLC chromatograms; b) UV spectra profiles and c) UV-Vis spectra of DPPH• reduction profiles;

Due to their numerous benefits for the human body less-known fruits like cornelian cherry and blackthorn should be introduced in a more balanced diet. That's why through this study we tried familiarizing these fruits among the population because they have the advantage of accessibility, being found in the wild flora of Romania.

Therefore, it can be observed from the chemometric analysis of the TLC digital chromatograms and UV spectra, that these fruits are clustered in well delimited groups with no apparent similarity to other fruits. But after the determination of the DPPH• scavenging activity profiles and chemometric analysis of these, we can notice that the profile of blackthorn is similar to that of rose hip, sea-buckthorn and cranberry, this suggesting that blackthorn contains similar quantities of anthocyanins and ascorbic acid as rose hip, sea-buckthorn and cranberry.

On the other hand the cornelian cherry seems to be similar to blackberry, according to the chemometric analysis, which points out that the composition of these two fruits are similar, possibly with similar effects on the human body.

RSA% - time profile method

To monitor the scavenging profile of the samples, the absorbance at 517 nm was recorded for all concentrations (C0, C1, C2 and C3) at different time intervals (T1–T7: 1, 5, 10, 15, 20, 25 and 30 minutes after the reaction was started) and the RSA% (% relative scavenging activity) was calculated for each obtained absorbance using the formula: RSA % = $[(A_{control} - A_{test})/A_{control}]\times 100$, where $A_{control}$ is the absorbance of the DPPH• solution without test sample and A_{test} is the absorbance of the DPPH• solution plus test sample.

Simply displaying the values of RSA% in a table isn't enough when the aim is to compare different plant extracts. Thus the representation of the data in a plot (Figure 25) is more suitable, managing to outline simultaneously the time-profiles for different concentrations of all samples.

Figure 25 shows that the measured radical scavenging activity and reaction kinetics of different extracts depend on the applied sample concentration. Thus the DPPH• is gradually

consumed for lower concentrations (C0, C1 and C2) with an ascending RSA%-time profile, but in most cases, for higher concentration (C3) of extract the reaction reaches an equilibrium after the first minute. Also the results indicate that the samples which have a steeper slope of the time-profile are samples with higher antioxidant activity, for example P3, P4, P6 and P7.

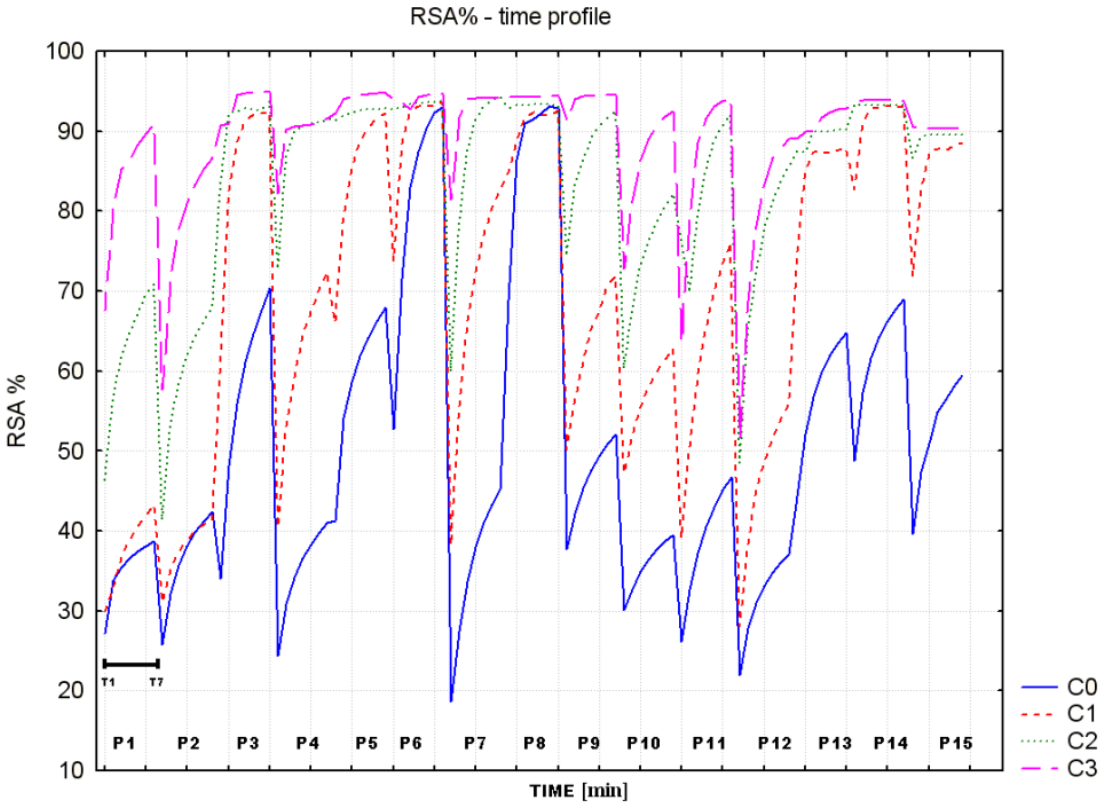


Figure 25 DPPH• scavenging activity-time profiles, expressed as RSA% at four different concentrations; C0 = 3.33% extract, C1 = 8.33% extract, C2 = 16.66% extract and C3 = 25% extract respectively.

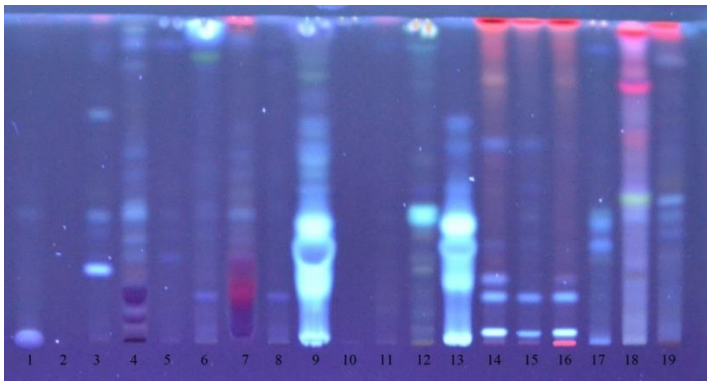
3.2 Authentication and fingerprinting of dietary supplements derived from berries

Results and discussion

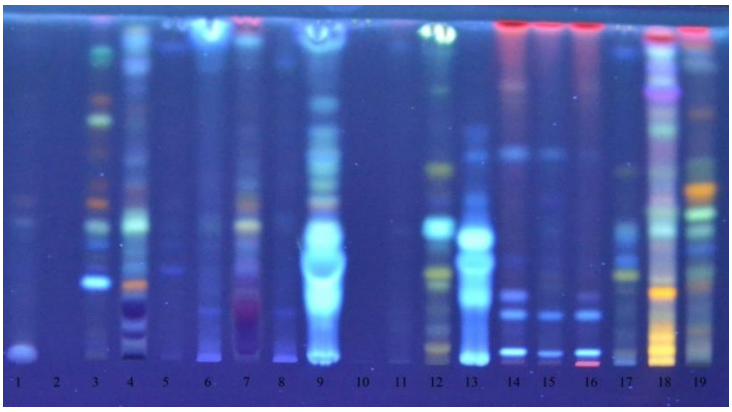
Analytical information acquirement

Considering the investigated samples, (herbal medicines based on cranberry, bilberry and sea-buckthorn extract/fruit) their therapeutic use is derived mainly from their high content of polyphenols. Consequently the analytical techniques used in this study were selected and optimized accordingly.

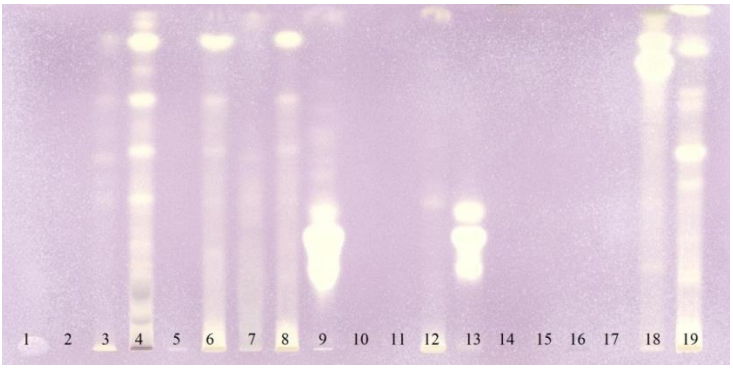
The images of the chromatographic plates are presented in Figure 26. There can be observed how the chromatographic image is changing while the plate is sprayed with NTS reagent for highlighting the polyphenols (image (a) vs. image (b)).



a)



b)



c)

Figure 26 The images of the TLC plates: a) UV observed compounds; b) polyphenols and c) antioxidants

There can be seen that the samples presents significant amount of polyphenols, but their positive reaction with DPPH was lower than expected. However, some of the samples are very poor in chemical composition, so their biological activity is expected to be very low. By carefully examination of the TLC chromatograms obtained with TLC Analyzer and also the UV–Vis spectra corresponding to the investigated berries there may be appreciated that they present some significant differences (Figure 27).

The cranberry chromatograms and spectra present the lowest level of specific characters. On the other side, the bilberry and sea-buckthorn present specific peaks in similar region. The UV-Vis spectra are more specific, since there may be observed the peaks associated to berries fruit (around 450 nm for sea-buckthorn and 550 nm for bilberry). The cranberry supposed to present a specific peak around 650 nm, corresponding to red region. However, it cannot be identified on the spectrum and this is associated to lower extraction efficiency in case of cranberry.

Anyway, the extraction conditions were very good for the rest of the investigated samples so they were maintained and used for dietary supplements samples.

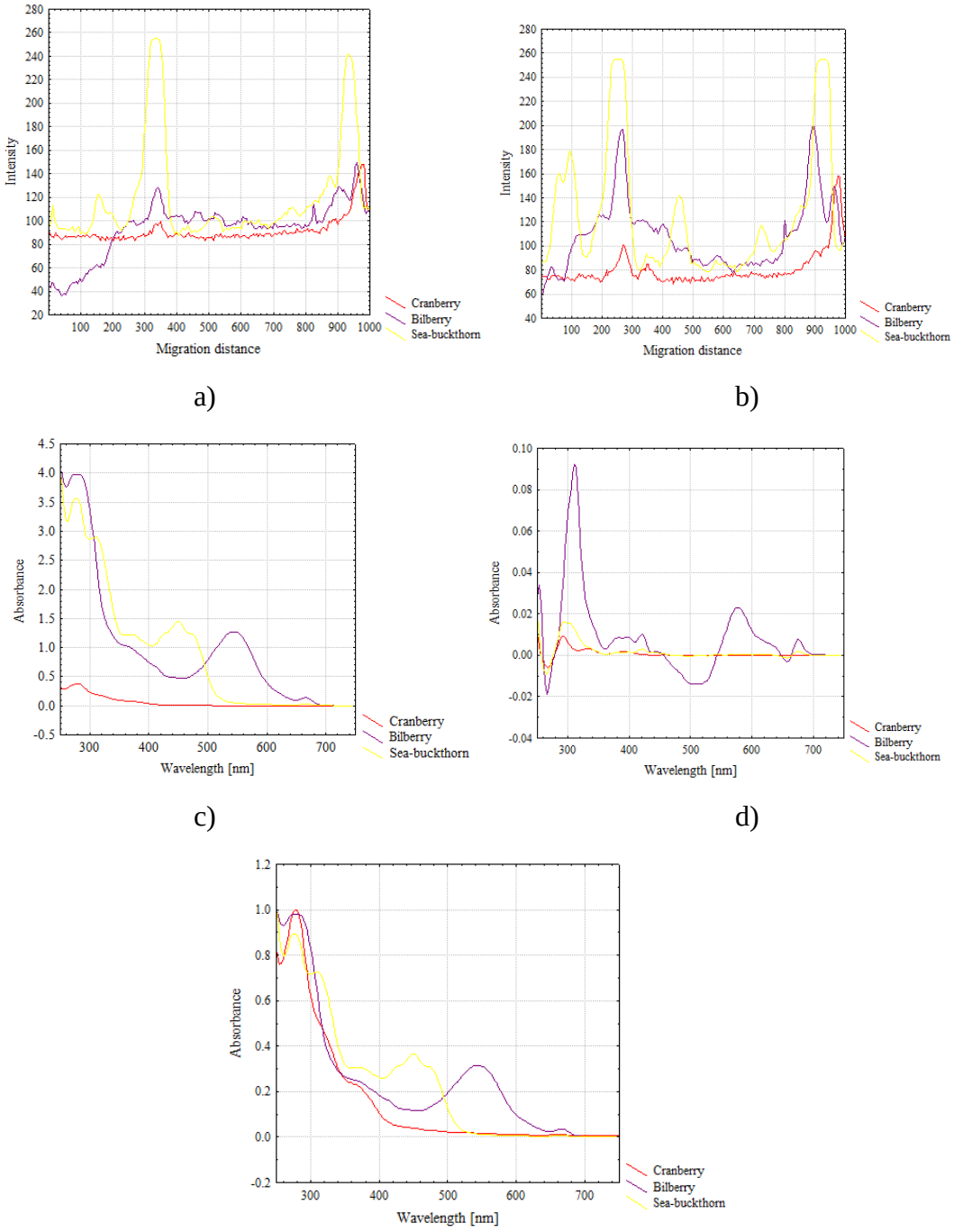


Figure 27 The TLC chromatograms (UV observed compounds (a); polyphenols (b)) and UV-Vis spectra (zero order derived spectra (c); first order derived spectra (d); normalized spectra (e)) obtained for cranberry (red), bilberry (purple) and sea-buckthorn (yellow)

In case of HPLC analysis, to achieve efficient resolution and as many signals as possible, mixtures of solvent A – water or acidified water (0.1% formic acid) and solvent B – methanol or methanol : acetonitrile (1:1), as mobile phase were tested. Finally, solvent A – acidified water (0.1% formic acid) and solvent B –methanol : acetonitrile (1:1), were selected as the best elution system, as the acidified water gave better peak shapes and the mixture of methanol and acetonitrile had better elution power.

Also, isocratic and several gradient conditions, selected from literature and adapted, were tested to optimize the HPLC separation. As the sample's components were barely separated under isocratic conditions, gradient elution was used instead and the best results were obtained using an adapted version of our previous work.

On the other hand, the MEKC method is generally applicable for the separation of neutral components and this technique also needed some optimisation. The bioactive compounds from the investigated samples have in their structure aromatic rings and differ in their pattern of hydroxylation, methylation and glycosylation. Accordingly they could be ionisable or neutral. The interaction between the polyphenols and the negatively charged micelles of the buffer depends on the charge value and the hydrophobicity of the compounds.

The MEKC has a resolving effect on the neutral polyphenols, while the charged ones have small interactions with the micelles. Also, as the polyphenols strongly interact with the micelles due to the hydrophobic properties therefore the resolution may be varied by modifying the micellar phase. The addition of organic solvents to the background electrolyte containing surfactant (SDS) is commonly used modifier in order to improve the selectivity [19].

At pH 9.3 in borate buffer polyphenols are negatively charged due to the dissociation of phenolic groups (pK≈9), thus they migrate according to their charge-to-size ratio. The 50 mM concentration of SDS resolves the neutral components. The borate has a complexation effect on the glycosides enhancing selectivity. The post-conditioning procedure was also optimized, based on our previous work [20-21], thus the capillary was flushed with 100 mM of SDS for 10 min, after daily use in order to remove adsorbed components from the capillary wall.

As for the detection of the bioactive compounds, DAD detection was considered to be a good choice, as the structure of these compounds allows them to have strong UV absorbance at different wavelengths. Therefore, in both MEKC and HPLC analysis, different UV wavelengths were tested: 200, 214, 250, 280, 365 nm and 360 and 280 respectively, and the best was chosen based on the highest density of detected peaks.

Thus, the best results were obtained using 200 nm for the MEKC analysis and 280 nm for the HPLC analysis, examples of electropherogram and chromatogram are presented in Figures 28 and 29.

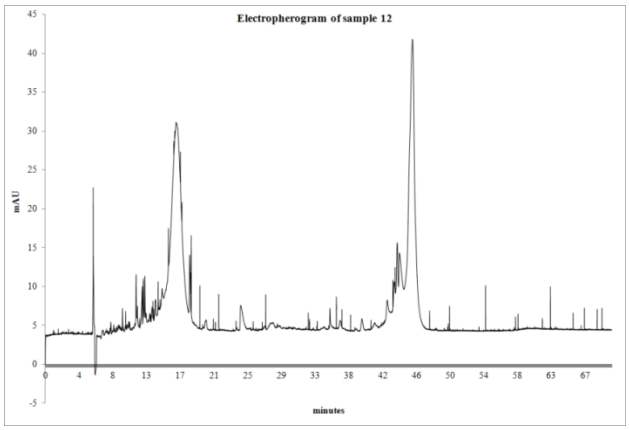


Figure 28 MEKC polyphenolic profile of sample 12, regisrered at 200 nm, using 50 mM disodium tetraborate and 50 mM sodium dodecyl sulfate (pH = 9.3) as BGE

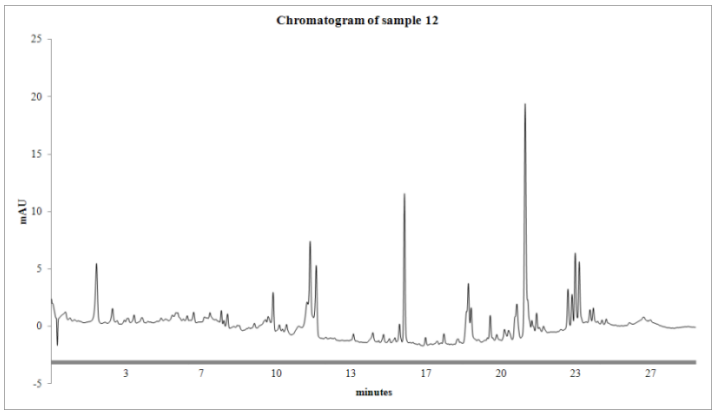


Figure 29 RP-HPLC polyphenolic profile of sample 12 regisrered at 280 nm, using a mixture of A (H_2O with 0.1% formic acid) and B (Metanol : Acetonitrile, 1:1 v/v)

Chemometric data analysis

The herbal medicines are processed products that retain and concentrate a part of the bioactive compounds of the raw material that they are made of. However, if trying to classify them according to this aspect, a visual differentiation among different samples could be done, but the process would be subjective and also small differences between related samples might be missed.

Thus, in order to obtain the authentication of herbal medicines, more advanced methods of discrimination are required, and several chemometric approaches were tested (Cluster Analysis (CA), Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA)), considering the sample's polyphenolic profiles (HPTLC or HPLC chromatograms, UV spectra and electropherograms, respectively) as analytical information.

The first multivariate exploratory technique used was CA. The dendrograms were obtained by applying CA on the digitized HPTLC chromatograms (19 samples × 1000 variables), HPLC chromatograms (19 samples × 900 variables), UV-Vis spectra (19 samples × 501 variables) and MEKC electropherograms (19 samples × 854 variables).

Ward's method of amalgamation for cluster building has been selected, because it uses an analysis of variance approach to evaluate the distance between clusters. Moreover, the distance between clusters was computed by squared Euclidean method, which is not affected by the addition of new objects to the analysis or by outliers. These selections should lead to the best classification offered by CA.

The dendrograms obtained by applying the CA on the digitized electropherograms, HPTLC and HPLC chromatograms and zero order UV-Vis spectra offer some information about the similarities/dissimilarities observed between the analyzed samples which are being mostly associated according to the nature of barriers used for production.

Generally good clustering was obtained for samples containing cranberry and sea-buckthorn, regardless of the analytical technique that was used for the separation.

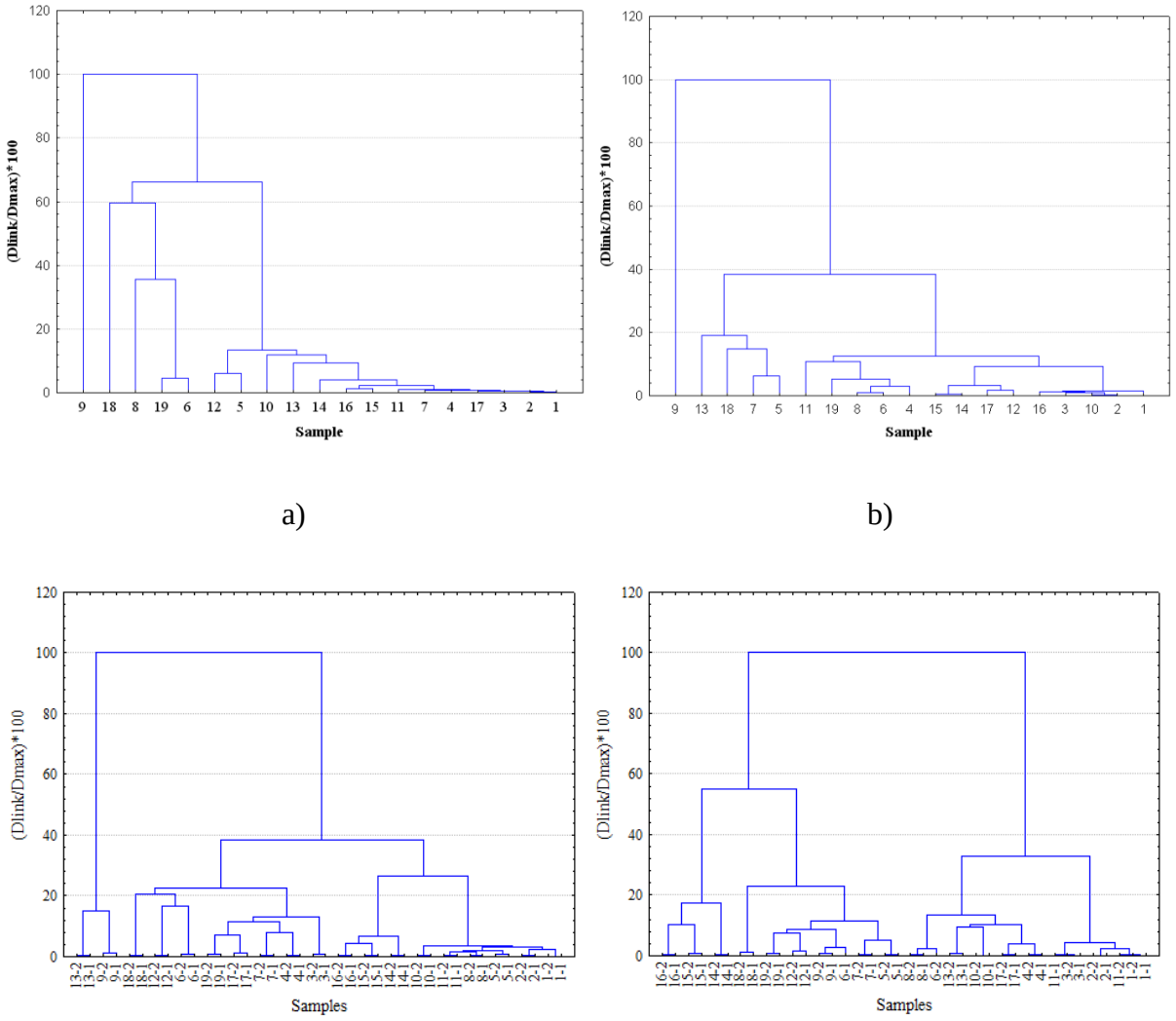


Figure 30 Dendograms obtained by applying Cluster Analysis on data matrices of digitized: a) electropherograms, b) HPLC chromatograms, c) HPTLC chromatograms and d) zero order UV-Vis spectra

But, as it can be observed in case of HPLC and MEKC analysis, sample 9 was the most differentiated from the others, forming a group of its own, and this can be attributed to the fact that this sample contains the least amount of extract/tablet, and these techniques are apparently more sensitive.

Furthermore, it is interesting to observe the classification of the herbal medicines obtained by combination of the selected berries. Sample 19 is a mixture of different species, amongst which may be mentioned the cranberry and bilberry. According to all classification, the sample was stronger associated to the cranberry based products.

Samples 17 and 18 contain both bilberry and sea-buckthorn, but their classification was quite different. For example, sample 18 was weekly associated to a group, while 17 was classified differently by HPTLC and HPLC chromatograms vs. spectra and electropherograms. The dendrograms obtained on the HPTLC and HPLC chromatograms was placing sample 17 more closely to the cranberry samples, while the dendrograms corresponding to the spectra and electropherograms was indicating that the sample is closer to sea-buckthorn.

In order to confirm the CA observation, and for a better visualization of samples distribution the PCA was applied on the covariance matrices of the digitized spectra and chromatograms. PCA was used to reduce the dimensionality of the original dataset by explaining the correlation among a large number of variables on the basis of a smaller number of principal components (PCs) without much loss of information. The projected dots of the electropherograms, HPTLC and HPLC chromatograms and spectra were localized in a confined cluster in the 2D-projection plot of PCA (Figure 31).

As it is presented, the first two PCs obtained from the HPLC chromatographic data account for more than 84% of the variance, more than 78% of the variance for the HPTLC chromatographic data, more than 89% of the variance for the UV-Vis spectrometric data, while the two PCs corresponding to electrophoretic data account for approximately 56% of the variance, respectively. The obtained PCA patterns, although more illustrative, are in good agreement with those obtained with CA. Also, is easy to observe that the samples are generally shuffled, this being a result of the large abundance of flavonoids in the herbal medicines, which is leading to a lower discrimination of the samples.

Finally, it is interesting to see that sample 18 has the same tendency of being separated from the rest of the samples regardless of the analytical technique used for separation, this confirming once more that it is highly different, and it cannot be associated to any of the raw materials used for its preparation.

Furthermore, the low discrimination obtained in all cases, can be attributed to the fact that most of the samples are mixtures of different ingredients (extracts/fruits) which contain large amounts of flavonoids and polyphenolic compounds and they cannot be discriminated by any of the multivariate classical methods (CA and PCA). Regarding PCA it is well documented that in many cases, more than two or three significant PCs are necessary to adequately characterize the data. In these cases there are more possible graphs and, as a direct consequence, the information retained in a larger number of PCs is dissipated.

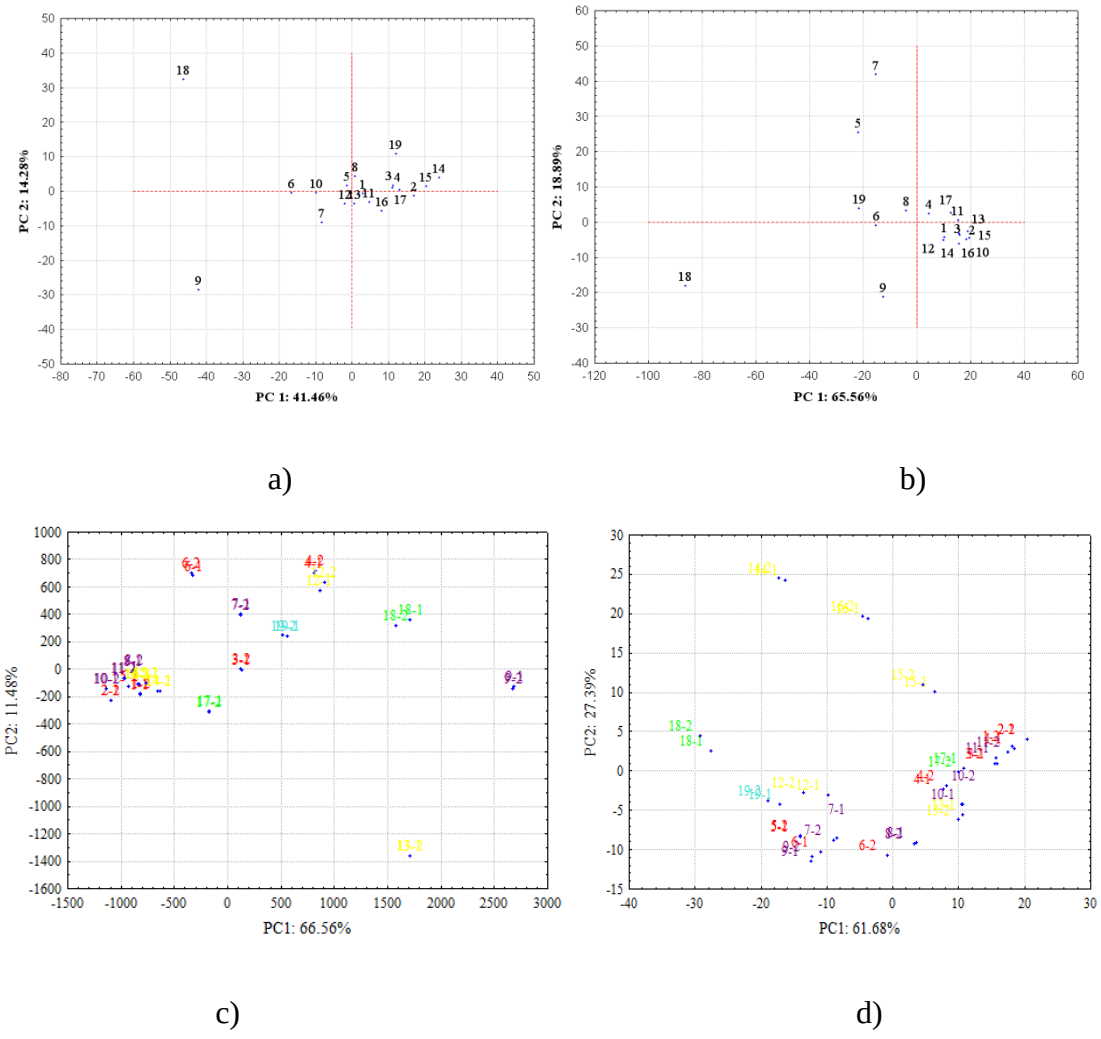


Figure 31 2D projections of PC1 vs. PC2 obtained by applying PCA on data matrices of digitized: a) electropherograms, b) HPLC chromatograms, c) HPTLC chromatograms and d) zero order UV-Vis spectra

However, this situation can be proficiently resolved by using a combination of PCA with LDA which could lead to a more efficient discrimination of the investigated samples, according to our previous work and other relevant applications [22, 23]. In this way, the variance covariance matrix of the new variables becomes a diagonal matrix, because the scores are orthogonal and the number of PCs is less than or equal to the number of samples.

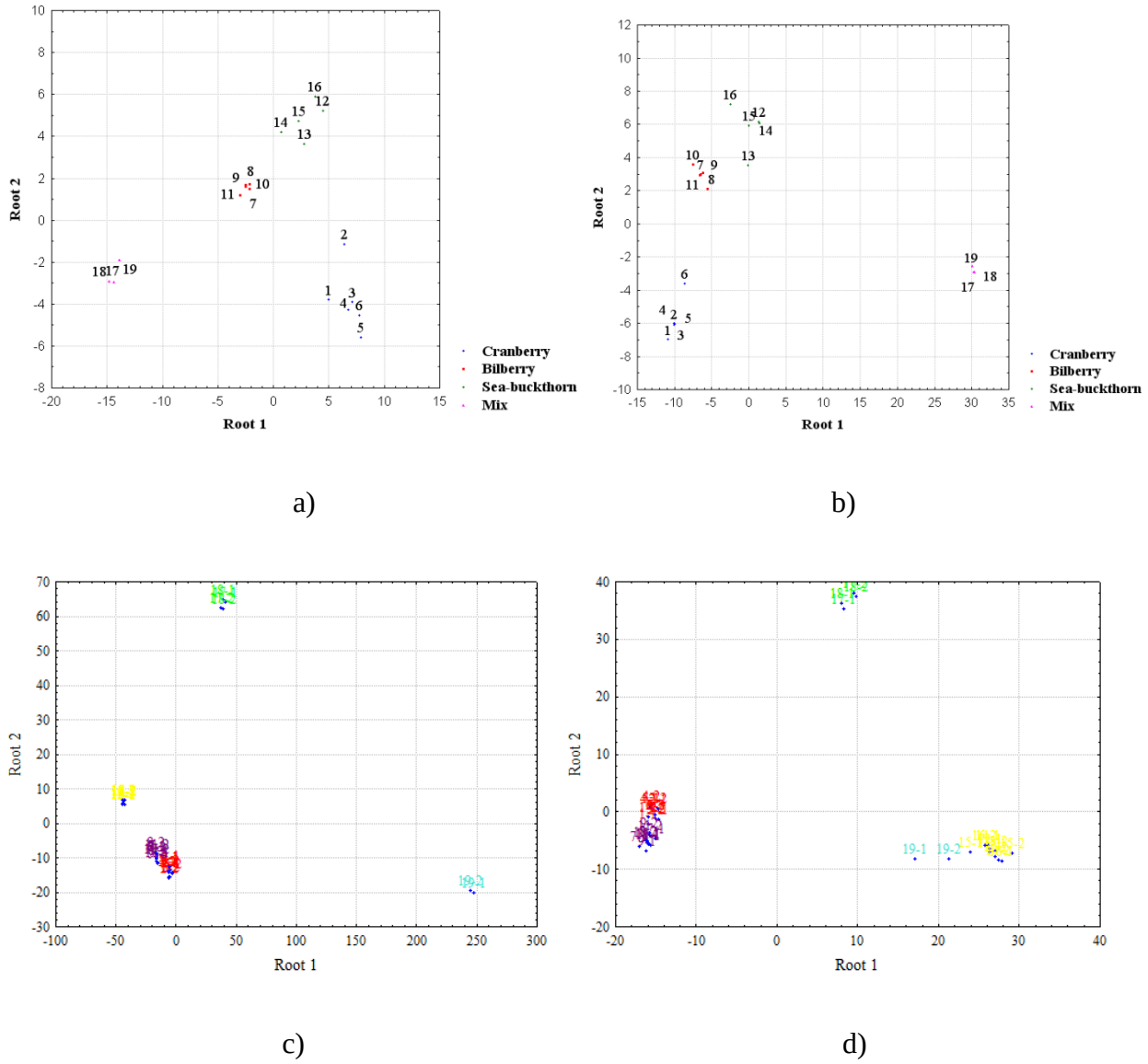


Figure 32 Plot of Root1 vs. Root2 scores obtained by applying PCA-LDA methodology on data matrices of digitized: a) electropherograms, b) HPLC chromatograms, c) HPTLC chromatograms and d) zero order UV-Vis spectra

LDA is a supervised classification technique based on linear discriminant functions which maximize between class variance and minimize within-class variance. The Euclidean distance was used in the LDA algorithms to classify unknown samples and the stepwise algorithm was used to extract the most important variables.

The results obtained by applying LDA to the scores corresponding to the first 15 principal components for the HPLC and MEKC analyses and the first 12 principal components for the HPTLC and UV-Vis spectrometry analyses, respectively, indicate a total separation of samples (100%) within four groups, in good agreement with the nature of the raw material used for their preparation and independently of the separation technique.

The Root1–Root2 score plots (Figure 32a-d) illustrate well differentiated groups of samples based on cranberries, bilberries, sea-buckthorn and mixtures, without any overlapping.

Although the concentration of fruit/extract was very different in each sample, the proposed combination of PCA-LDA was able to successfully classify the samples according to the nature of their raw material. Also, the results indicate that regardless of the separation technique, the classification of samples was made along the Root1 axis for the mixtures (samples 17, 18 and 19) and mostly along the Root2 axis for the other samples. Moreover, all the above presented analyses are indicating that the combination of PCA with LDA is leading to more powerful classification and discrimination of samples.

Chapter IV – Concluding remarks

The present PhD thesis focuses on three main objectives, regarding the application of chemometrics in different fields of analytical chemistry, in order to improve the processing and interpretation of the instrumental signal.

The first objective was to demonstrate how chemometric methods represent important tools in the validation of newly developed methods. For this, a new method of analysis was proposed using digital thin-layer chromatography for the investigation of catecholamines' metabolites from biological samples.

Therefore, a convenient, sensitive and rapid TLC - image processing method was developed and validated for quantitative evaluation of acidic catecholamine metabolites (HVA, VMA, DOMA, DOPAC) and also the metabolites that are associated with adrenal tumours when hyper-secreted in urine (NMN, MN and 3-MT). The advantages of the proposed method are the rapid measurement of metabolites using inexpensive equipment and, in addition, this method does not require laborious sample clean-up procedures or a complicated pre-derivatization step.

The high sensitivity with $DPPH^{\bullet}$ detection and the high quality of the validation parameters (accuracy and precision, LOD and LOQ) showed that this method should be useful for rapid preliminary biomedical investigations of the acidic metabolites of catecholamine in case of diseases associated with a hyper excretion of these compounds in urine. Furthermore, the simple sample preparation, inexpensive equipment and short analysis time are grounds that recommend this method for the rapid screening of pheochromacytomas and other adrenal tumours.

Further, for the second objective the chemometric methods were used for modeling and predicting lipophilicity: on one hand for catecholamines and related compounds, using the algoritm proposed by the TopoCluj group, and on the other hand for antioxidant compounds with different structures, using various HPLC experimental conditions.

In the first study of this chapter, a set of thirty eight catecholamines and related compounds was submitted to a novel QSAR method based on the alignment of all structures over a hypermolecule, thus obtaining a powerful topological descriptor, the summative descriptor (SD), for the prediction of lipophilicity (log P). The set of molecules was divided in two groups, the first group (training set) was used to develop the QSAR models by multivariate regression and also by genetic algorithms, and the second group (test set) was used to validate the obtained models.

The results indicate that the QSAR model obtained using multivariate regression has good predictive capacity in case of external validation but in case of validation by similarity clusters the results were significantly improved, from a coefficient of correlation of 0.8773 in the first case, to 0.9263 in the second case, respectively. Also the QSAR model obtained using genetic algorithms provided similar results, with a coefficient of correlation of 0.9226, thus supporting the idea that the new QSAR approach is of great use in predicting the lipophilicity of catecholamine related compounds.

The second study of this chapter consisted of investigations concerning the lipophilicity of a group of antioxidant compounds using reversed-phase high-performance liquid chromatography. Different mixtures of methanol-water as mobile phase and several stationary phases, such as RP18, C8, C16-Amide, CN and PFP were tested, and the results indicated pterostilbene as the most lipophilic compound.

Significant correlations were observed between different experimental indices of lipophilicity at the two temperatures and some computed log*P* scales (CLogP, MLOGP, ALogP98). The mlogk values were the most correlated with the computed indices. In addition, the results obtained in this study by applying multivariate exploratory techniques, like HCA, PCA or the two-way joining clustering and profile representation illustrated more or less the same (dis)similarities of the stationary phases and they were well supported by the ranking scales generated applying SRD-CRRN algorithm. Overall, the results (mainly mlogk indices) illustrate a similar and small effect of temperature on the chromatographic behavior of the investigated compounds in all cases. In consequence, we concluded that the mean (mlogk) is a better lipophilicity estimator, as it is not affected as much by experimental and model errors like in the

case of the extrapolation estimator ($logk_w$), conclusion which was also pointed out in the literature and well supported by these results.

And finally, the third objective was to point out the necesity of chemometric tools for obtaining holistic and comprehensive fingerprints and for characterisation and authentication of various samples. For this, the fingerprinting analysis was applied for wild berries and derived dietary supplements, using various analytical techiques assisted by different chemometric aproaches.

Fingerprinting of Romanian known and less-known berries was carried out based on thin-layer chromatography profiles, spectrometry using the UV spectra and DPPH• scavenging profiles. The chemometric analysis, which involved cluster analysis (CA) and principal component analysis (PCA), was successfully coupled with analytical techniques in order to classify the berry samples. Also, the time profiles of the antioxidant activity, expressed as RSA%, were determined for the first time for these types of samples, at four different concentrations. In addition, the less-known berries, cornelian cherry and blackthorn, were found to have similar antioxidant profiles to blackberry and rose hip, sea-buckthorn and cranberry, respectively.

Furthermore, it has been proved that the dietary supplements can be classified according to the raw material used for their production. The simple chemical methodologies may not offer information regarding the herbal medicines nature, but combined with adequate chemometric methodologies the samples may be discriminated and authenticated.

In addition, there may be concluded that CA and PCA may offer some preliminary results, but the combination of PCA with LDA leads to more powerful classification and discrimination of samples, according to their raw material composition. Also, because the results did not show significant differences by using different separation techniques, it is suited to use either of them for similar experiments, with respect to their advantages/disadvantages.

Moreover, the simple and efficient methodology developed in this chapter might be used for screening and authenticity control of different products (herbal medicines, drugs, food, etc.) and can be implemented in any quality control laboratory.

References

[1] Miller C, Chemometrics in Process Analytical Chemistry. In: Process Analytical Technology: Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical Industries, pp. 226–328. Ed. Bakeev KA, Blackwell Publishing Ltd, 2005.

[2] Massart D, Vandeginste B, Buydens L, De Jong S, Lewi P, Smeyers-Verbeke J: Handbook of chemometrics and qualimetrics: Part A. Elsevier, 1997

[3] Hibbert DB, Minkkinen P, Faber NM, Wise BM, Anal Chim Acta , 2009, 642, 3–5.

[4] Brereton RG, Chemometrics: applications of mathematics and statistics to laboratory systems – E. Horwood, New York, 1990.

[5] Workman J Jr, Chem. Intell. Lab. Syst., 2002, 60, 13-23.

[6] Kaczmarek K, Walczak B, de Jong S, Vandeginste BGM, ActaChromatogr., 2005, 15, 82–96.

[7] Diudea MV, MATCH Commun. Math. Comput. Chem., 1997, 35, 169-183.

[8] Diudea MV, J. Chem. Inf. Compu. Sci., 1997, 37, 300-305.

[9] Harsa AM, Studia UBB Chemia, 2014, 1, 111-123.

[10] Harsa TE, Studia UBB Chemia, 2014, 1, 99-110.

[11] Maties R, Szefler B, Ionut I, Tiperciuc B, Studia UBB Chemia, 2012, 4, 121-133.

[12] Harsa TE, Harsa AM, Szefler B, Cent. Eur. J. Chem., 2014, 12(3), 365-376.

[13] Goldberg DE, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, Massachusetts, 1989.

[14] Wehrens R, Buydens LMC, Trends Anal. Chem., 1998, 17(4), 193-203.

[15] Héberger K, Trends Anal. Chem., 2010, 29, 101–109.

[16] Andric F, Heberger K, J. Chromatogr. A, 2015, 1380, 130–138.

[17] Andric F, Heberger K, J. Pharm. Biomed. Anal., 2015, 115, 183–191.

[18] Andric F, Bajusz D, Racz A, Segan S, Heberger K, J. Pharm. Biomed. Anal., 2016, 127, 81-93.

[19] Tonin GF, Jager AV, Micke GA, Farah JPS, Tavares MFM, Electrophoresis, 2005, 26, 3387-3396.

[20] Andrási M, Gáspár A, Klekner Á, J. Chromatogr. B, 2007, 846, 355-358.

[21] Andrási M, Zékány L, Gáspár A, J. Anal. Chem., 2015, 70: 1360-1367.

[22] Jombart T, Devillard S, Balloux F, BMC Genetics, 2010, 11, 94-109.

[23] Brereton GR, Applied Chemometrics for Scientists. John Wiley & Sons Ltd, Chichester, 2007.