Talanta 83 (2010) 225–232

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Application of multivariate image analysis in QSPR study of ^{13}C chemical shifts of naphthalene derivatives: A comparative study

Zahra Garkani-Nejad*, Marziyeh Poshteh-Shirani

Chemistry Department, Faculty of Science, Vali-e-Asr University, Rafsanjan, Iran

ARTICLE INFO

Article history:
Received 29 June 2010
Received in revised form 6 September 2010
Accepted 8 September 2010
Available online 16 September 2010

Keywords:
Multivariate image analysis (MIA)
Quantitative structure–property relationship (QSPR)
Principal component-artificial neural network (PC-ANN)
^{13}C chemical shift
Naphthalene derivatives

ABSTRACT

A new implemented QSPR method, whose descriptors achieved from bidimensional images, was applied for predicting ^{13}C NMR chemical shifts of 25 mono substituted naphthalenes. The resulted descriptors were subjected to principal component analysis (PCA) and the most significant principal components (PCs) were extracted. MIA-QSPR (multivariate image analysis applied to quantitative structure–property relationship) modeling was done by means of principal component regression (PCR) and principal component-artificial neural network (PC-ANN) methods. Eigen value ranking (EV) and correlation ranking (CR) were used here to select the most relevant set of PCs as inputs for PCR and PC-ANN modeling methods. The results supported that the correlation ranking-principal component-artificial neural network (CR-PC-ANN) model could predict the ^{13}C NMR chemical shifts of all 10 carbon atoms in mono substituted naphthalenes with $R^2 \geq 0.922$ for training set, $R^2 \geq 0.963$ for validation set and $R^2 \geq 0.936$ for the test set. Comparison of the results with other existing factor selection method revealed that less accurate results were obtained by the eigen value ranking procedure.

1. Introduction

Nuclear magnetic resonance (NMR) is a physical phenomenon based upon the quantum mechanical magnetic properties of an atom nucleus. Magnetic nuclei, like ^{1}H and ^{13}C absorb radiofrequency energy when placed in a magnetic field of strength specific to the identity of the nuclei. When this absorption occurs, the nucleus is described as being in resonance. Different atoms within a molecule resonate at different frequencies at a given field strength. The observation of the resonance frequencies of a molecule allows a user to discover structural information about the molecule. This phenomenon is known as the chemical shift and is the most important characteristic of a nucleus in terms of NMR. The shift of an individual atom depends on its atomic properties, such as the type of nucleus, its hybridization state and the overall electronic environment surrounding the nucleus [1].

In various fields of chemistry such as the investigation of natural products or the design of new compounds, scientists often need to either determine the structure of an unknown or new compound or to verify a hypothetical chemical structure. This process, known as structure elucidation, is based on the analysis of available spectral data. Nuclear magnetic resonance (NMR) spectroscopy is certainly one of the main analysis methods applied to these challenges and is a powerful technique for acquiring highly informative spectra associated with a structure.

Quantitative structure–activity/property relationship (QSAR–QSPR) studies, as one of the most important areas in chemometrics, give information that is useful for molecular design and medicinal chemistry [2–5]. QSAR/QSPR models are mathematical equations relating chemical structure to a wide variety of physical, chemical, biological and technological properties.

Jensen et al. [6] have used 33 polycyclic aromatic compounds with 24 different aromatic ring backbones, to generate linear regression models for ^{13}C chemical shift calculation. The chemical environment of each carbon atom was described by 2–11 parameters, obtained from the structural information (steric and electronic) calculated by Huckel method. Kvasnicka et al. [7] have published one of the first applications of using an artificial neural network for predicting and classifying ^{13}C chemical shifts based on functional group descriptors. The mathematical basis for ^{13}C NMR chemical shifts prediction using increments was discussed by Chen and Robien [8]. The incremental model was also used by Thomas et al. [9] for the prediction and assignment of the ^{13}C NMR spectra of substituted benzenes, naphthalenes and biphenyl compounds. Svozil et al. [10] have used artificial neural networks to predict ^{13}C NMR chemical shifts of alkanes. The topological description of each carbon atom was encoded using 13 descriptors that correspond to embedding frequencies of rooted sub-trees. Jurs and coworker [11–14] have published a series of papers comparing the results obtained by multiple linear regression analysis

* Corresponding author. Tel.: +98 391 3202416.
E-mail addresses: garakani@vru.ac.ir, z_garakani@yahoo.com (Z. Garkani-Nejad).

0039-9140/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2010.09.012

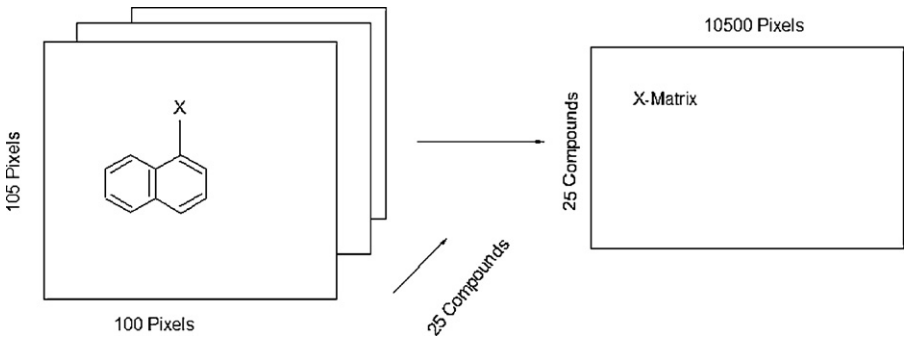


Fig. 1. 2D images and unfolding step of the 25 chemical structures to give the X-matrix. The arrow in structure indicates the coordinate of a pixel in common among the whole series of compounds, used in the 2D alignment step.

and feed-forward neural networks to simulate ^{13}C NMR spectra of keto-steroids [11], dibenzofurans [12], ribonucleosides [13], and monosaccharides [14].

A new strategy for prediction of the ^{13}C chemical shift is construction of the QSPR models using multivariate image analysis descriptors. Goodarzi et al. have reported a quantitative structure–property relationship study on the ^{13}C chemical shifts of methoxyflavonol derivatives using MIA-QSPR method [15]. They revealed that the predictive ability of MIA descriptors is comparable or even superior to the Gauge Included Atomic Orbital (GIAO) procedure for ^{13}C chemical shifts calculations.

Geladi and Esbensen [16] have demonstrated that image analysis may provide useful information in chemistry, though the descriptors do not have a direct physicochemical meaning, since they are binaries. In QSPR, images (2D chemical structures) have shown to contain chemical information [17,18], allowing the correlation between chemical structures and properties.

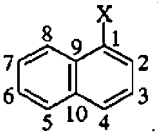
The present paper is focused on the application of 2D images, which are the proper structures of the compounds that can be drawn with aid of any appropriate program, as descriptors in QSPR. Then, multivariate image analysis-quantitative structure property relationship study (MIA-QSPR) is proposed to model and predict the ^{13}C chemical shifts of a series of naphthalene derivatives [19] using principal component regression (PCR) and principal component-artificial neural network (PC-ANN) modeling methods. Eigen value ranking (EV) and correlation ranking (CR) were used to select the most relevant set of PCs as inputs for PCR and PC-ANN modeling methods. Finally, obtained results using different methods are compared.

2. Experimental

2.1. Dataset

The ^{13}C NMR chemical shifts of 25 mono substituted naphthalenes (in ppm relative to TMS) were obtained from the literature [19]. The chemical structures of these compounds and their ^{13}C chemical shifts have been listed in Table 1.

Table 1
The structure of naphthalene derivatives and experimental ^{13}C chemical shifts for 10 positions.



Substituent *X*	C-1	C-2	C-3	C-4	C-5	C-6	C-7	C-8	C-9	C-10
–H	128	125.9	125.9	128	128	125.9	125.9	128	133.6	133.6
$-CH_3$	134	126.4	126.5	126.2	128.3	125.2	125.4	123.9	132.5	133.4
$-C(CH_3)_3$	145.9	123.1	125	127.4	129.6	123.6	123.6	126.4	132	135.8
$-CH_2Br$	132	127	125	129.3	128.5	125.8	126.2	125.2	130.8	133.7
$-CH_2OH$	136.2	125	125.3	128.1	128.5	125.6	126	125.4	131	133.6
$-CF_3$	128	124.6	124.1	133	129	126.7	127.9	129	134.6	129.7
–F	159.5	109.8	126	124.2	128.1	127.3	126.6	118.7	124.3	135.7
–Cl	131.9	126.1	125.7	127.1	128.2	129	126.7	125.2	130.8	134.6
–Br	122.6	129.5	125.7	127.5	127.9	126.3	126.9	126	131.6	134.2
–I	99.6	138.2	127.6	129.7	129.4	127.5	128.5	129.3	134.9	134.9
–OH	151.5	108.7	125.8	120.7	127.6	126.4	126.2	118.7	124.3	133.6
$-OCH_3$	155.3	103.6	125.7	120.1	127.3	126.2	125	119.9	125.5	134.4
$-OCOCH_3$	146.6	118	125.3	125.9	128	126.3	126.3	121.1	126.7	134.5
$-NH_2$	142	109.4	126.2	118.7	128.3	125.6	124.6	117.8	123.4	134.2
$-N(CH_3)_2$	151.7	114.7	126.5	123.4	129	126.3	125.6	124.1	129.7	135.7
$-NH_3^+$	124.2	121.3	125	131.4	129.4	128	128.7	120.6	126.2	134.8
$-NO_2$	146.5	123.8	123.9	134.5	128.5	127.2	129.3	119.3	124.9	134.2
–CN	108.8	131	123.5	131.8	127.3	126.1	127.1	125.2	130.8	131.4
–CHO	130.9	136.7	124.5	134.7	128.2	126.5	128.6	124.4	130	133.3
$-COCH_3$	134.9	128.8	124.2	132.9	128.3	126.3	127.9	124.5	130.1	133.8
–COOH	126.5	129.5	123.5	132.3	127.4	125	126.5	124.8	130.4	132.8
$-COOCH_3$	127.1	130.4	124.7	133.4	128.7	126.4	127.8	126.1	131.7	134.1
$-CON(CH_3)_2$	134.8	123.8	125.1	128.9	128.4	126.3	126.9	123.9	129.5	133.4
–COCl	129.2	136.5	125.4	137.3	129.9	128	130.4	125.9	131.5	134.6
$-Si(CH_3)_3$	137.8	131	125.5	129.7	129.2	125.1	125.2	131.8	137.4	133.8

Table 2
CR-PCR models for C1–C10 positions of naphthalene derivatives.

Position	Model
1	$134.620(\pm1.409)+0.478(\pm.165)PC_4-0.508(\pm0.217)PC_7+1.067(\pm0.247)PC_{10}+0.778(\pm0.371)PC_{17}-1.364(\pm0.483)PC_{21}+2.886(\pm0.543)PC_{22}$
2	$124.112(\pm0.930)+0.335(\pm0.100)PC_3-0.495(\pm0.109)PC_4+0.505(\pm0.143)PC_7-0.264(\pm0.155)PC_9-0.376(\pm0.171)PC_{11}-0.420(\pm0.245)PC_{17}-1.555(\pm0.358)PC_{22}$
3	$125.264(\pm0.107)-0.016(\pm0.009)PC_2+0.034(\pm0.013)PC_4+0.055(\pm0.014)PC_5-0.043(\pm0.016)PC_7-0.059(\pm0.020)PC_{11}-0.051(\pm0.021)PC_{12}+0.073(\pm0.027)PC_{16}+0.056(\pm0.028)PC_{17}+0.075(\pm0.031)PC_{19}$
4	$128.648(\pm0.499)+0.195(\pm0.054)PC_3-0.254(\pm0.059)PC_4-0.125(\pm0.064)PC_5-0.147(\pm0.072)PC_6+0.343(\pm0.077)PC_7+0.178(\pm0.099)PC_{12}-0.294(\pm0.143)PC_{19}-0.362(\pm0.192)PC_{22}$
5	$128.440(\pm0.074)+0.020(\pm0.006)PC_1+0.035(\pm0.008)PC_3-0.022(\pm0.012)PC_8+0.042(\pm0.013)PC_{10}-0.030(\pm0.014)PC_{11}+0.042(\pm0.015)PC_{13}+0.040(\pm0.018)PC_{15}+0.045(\pm0.025)PC_{21}-0.062(\pm0.028)PC_{22}$
6	$126.344(\pm0.133)-0.020(\pm0.010)PC_1-0.025(\pm0.012)PC_2+0.039(\pm0.022)PC_8+0.056(\pm0.022)PC_9+0.073(\pm0.031)PC_{14}+0.071(\pm0.035)PC_{17}+0.089(\pm0.036)PC_{18}-0.062(\pm0.038)PC_{19}+0.080(\pm0.041)PC_{20}-0.118PC_{21}-0.152(\pm0.058)PC_{23}$
7	$126.792(\pm0.175)-0.024(\pm0.013)PC_1+0.061(\pm0.19)PC_3-0.040(\pm0.021)PC_4-0.044(\pm0.027)PC_5+0.089(\pm0.027)PC_7+0.057(\pm0.029)PC_8+0.094(\pm0.041)PC_{14}+0.116(\pm0.047)PC_{18}-0.115(\pm0.050)PC_{19}+0.130(\pm0.053)PC_{20}-0.182(\pm0.077)PC_{23}$
8	$124.208(\pm0.358)+0.072(\pm0.027)PC_1+0.087(\pm0.032)PC_2+0.113(\pm0.038)PC_3-0.103(\pm0.042)PC_4+0.115(\pm0.046)PC_5+0.185(\pm0.055)PC_7-0.158(\pm0.063)PC_{10}-0.207(\pm0.084)PC_{14}-0.639(\pm0.138)PC_{22}$
9	$129.928(\pm0.367)+0.070(\pm0.028)PC_1+0.080(\pm0.032)PC_2+0.100(\pm0.039)PC_3-0.123(\pm0.043)PC_4+0.131(\pm0.047)PC_5+0.188(\pm0.057)PC_7-0.171(\pm0.064)PC_{10}-0.224(\pm0.086)PC_{14}-0.610(\pm0.142)PC_{22}$
10	$133.912(\pm0.128)-0.022(\pm0.010)PC_1+0.048(\pm0.019)PC_6-0.087(\pm0.020)PC_7+0.076(\pm0.022)PC_{10}-0.062(\pm0.024)PC_{11}+0.069(\pm0.026)PC_{13}+0.079(\pm0.034)PC_{18}-0.097(\pm0.044)PC_{21}+0.223(\pm0.056)PC_{23}$

Table 3
EV-PCR models for C1–C10 positions of naphthalene derivatives.

Position	Model
1	$134.620(\pm2.867)+0.144(\pm0.216)PC_1+0.099(\pm0.252)PC_2-0.299(\pm0.307)PC_3+0.478(\pm0.337)PC_4+0.119(\pm0.366)PC_5+0.235(\pm0.415)PC_6$
2	$124.112(\pm1.472)-0.031(\pm0.111)PC_1+0.042(\pm0.130)PC_2+0.335(\pm0.158)PC_3-0.495(\pm0.173)PC_4-0.114(\pm0.188)PC_5-0.213(\pm0.213)PC_6+0.505(\pm0.227)PC_7$
3	$125.264(\pm0.177)-0.005(\pm0.013)PC_1-0.016(\pm0.016)PC_2+0.004(\pm0.019)PC_3+0.034(\pm0.021)PC_4+0.055(\pm0.023)PC_5+0.024(\pm0.026)PC_6-0.043(\pm0.027)PC_7+0.018(\pm0.029)PC_8+0.026(\pm0.029)PC_9$
4	$128.648(\pm0.636)-0.002(\pm0.048)PC_1+0.042(\pm0.056)PC_2+0.195(\pm0.068)PC_3-0.254(\pm0.075)PC_4-0.125(\pm0.081)PC_5-0.147(\pm0.092)PC_6+0.343(\pm0.098)PC_7+0.020(\pm0.104)PC_8$
5	$128.440(\pm0.129)+0.020(\pm0.010)PC_1+0.000(\pm0.011)PC_2+0.035(\pm0.014)PC_3-0.012(\pm0.015)PC_4+0.015(\pm0.016)PC_5-0.007(\pm0.019)PC_6-0.004(\pm0.020)PC_7-0.022(\pm0.021)PC_8-0.007(\pm0.021)PC_9$
6	$126.344(\pm0.240)-0.020(\pm0.018)PC_1-0.025(\pm0.021)PC_2+0.020(\pm0.026)PC_3-0.005(\pm0.028)PC_4+0.008(\pm0.031)PC_5-0.018(\pm0.035)PC_6+0.029(\pm0.037)PC_7+0.039(\pm0.039)PC_8+0.056(\pm0.040)PC_9+0.019(\pm0.042)PC_{10}+0.024(\pm0.044)PC_{11}$
7	$126.792(\pm0.280)-0.024(\pm0.021)PC_1-0.011(\pm0.025)PC_2+0.061(\pm0.030)PC_3-0.040(\pm0.033)PC_4-0.044(\pm0.036)PC_5-0.046(\pm0.041)PC_6+0.089(\pm0.043)PC_7+0.057(\pm0.046)PC_8+0.041(\pm0.047)PC_9+0.041(\pm0.049)PC_{10}-0.016(\pm0.051)PC_{11}$
8	$124.208(\pm0.627)+0.072(\pm0.047)PC_1+0.087(\pm0.055)PC_2+0.113(\pm0.067)PC_3-0.103(\pm0.074)PC_4+0.115(\pm0.080)PC_5-0.066(\pm0.091)PC_6+0.185(\pm0.097)PC_7-0.021(\pm0.103)PC_8-0.064(\pm0.105)PC_9$
9	$129.928(\pm0.629)+0.070(0.047)PC_1+0.080(\pm0.055)PC_2+0.100(\pm0.067)PC_3-0.123(\pm0.074)PC_4+0.131(\pm0.080)PC_5-0.067(\pm0.091)PC_6+0.188(\pm0.097)PC_7-0.014(\pm0.103)PC_8-0.080(\pm0.105)PC_9$
10	$133.912(\pm0.249)-0.022(\pm0.019)PC_1-.018(\pm0.022)PC_2+0.003(\pm0.027)PC_3+0.010(\pm0.029)PC_4+0.028(\pm0.032)PC_5+0.048(\pm0.036)PC_6-0.087(\pm0.038)PC_7-0.042(\pm0.041)PC_8+0.001(\pm0.041)PC_9$

Table 4
Calculated values of ^{13}C chemical shifts using RC-PCR method for all carbon positions—training and validation sets.

Substituent *X*	C-1	C-2	C-3	C-4	C-5	C-6	C-7	C-8	C-9	C-10
Training										
$-CH_2Br$	133.49	122.81	124.97	129.69	128.85	125.74	125.94	124.42	130.28	133.63
$-CH_2OH$	139.36	123.34	125.61	128.49	128.87	125.46	126.24	124.58	130.16	133.67
$-CF_3$	129.73	125.04	125.15	130.58	128.99	126.22	127.72	129.02	134.69	129.70
–Cl	134.26	128.82	125.66	130.32	127.92	128.32	126.68	126.21	132.30	134.07
–Br	123.03	129.68	125.84	127.21	127.73	127.17	126.72	123.67	129.86	134.14
–I	107.66	134.16	126.63	126.96	129.47	126.86	127.86	127.83	133.07	134.57
–OH	150.50	109.83	126.04	122.40	127.60	127.48	126.69	120.94	126.32	133.49
$-OCH_3$	145.34	111.57	125.31	122.83	127.82	126.3	125.52	121.62	127.23	134.36
$-OCOCH_3$	145.34	111.57	125.31	122.83	127.82	126.34	125.52	121.62	127.23	134.36
$-NH_3^+$	128.19	121.23	124.94	131.26	128.64	127.52	128.25	122.89	128.91	133.64
–CN	115.28	122.89	123.81	133.44	127.43	126.27	127.80	125.93	131.66	132.05
–CHO	138.95	130.57	124.86	131.19	128.21	126.50	128.42	123.57	129.12	133.87
$-COCH_3$	130.36	126.68	123.73	131.36	128.28	126.18	128.03	124.62	130.18	134.88
$-COOCH_3$	126.91	134.52	124.77	135.88	128.98	126.24	126.94	127.04	132.49	133.10
–COCl	143.31	137.51	125.07	137.60	129.65	127.00	129.55	125.40	131.01	134.89
Valid										
–H	128.81	130.16	125.80	128.95	127.87	125.60	126.52	126.55	132.35	134.26
$-CH_3$	127.27	126.08	126.71	127.39	127.86	125.72	125.89	124.56	130.99	133.41
$-C(CH_3)_3$	140.28	126.87	125.70	124.48	129.487	123.99	124.37	124.46	130.04	135.31
–F	163.82	110.86	126.14	123.99	127.95	127.60	127.66	118.17	124.08	135.12
$-NH_2$	148.75	110.02	126.40	121.39	128.54	125.65	124.11	115.80	121.26	134.37
$-N(CH_3)_2$	144.98	115.81	125.92	123.785	129.02	126.79	124.95	125.53	131.38	135.90
$-NO_2$	142.32	125.27	123.93	132.66	128.74	127.40	128.68	119.64	125.01	134.43
–COOH	118.74	128.75	123.51	130.01	127.85	125.21	126.24	124.06	130.35	133.10
$-CON(CH_3)_2$	129.22	130.54	124.51	129.68	128.46	126.52	128.50	126.57	132.14	133.75
$-Si(CH_3)_3$	129.61	128.23	125.27	131.81	128.94	124.48	124.99	130.47	136.06	133.72

Table 5
Calculated values of ^{13}C chemical shifts using EV-PCR method for all carbon positions—training and validation sets.

Substituent *X*	C-1	C-2	C-3	C-4	C-5	C-6	C-7	C-8	C-9	C-10
Training										
$-CH_2Br$	131.97	126.99	128.12	131.46	128.43	125.82	126.93	125.20	130.80	133.70
$-CH_2OH$	136.09	124.96	127.88	127.67	128.82	125.76	125.7	125.38	130.98	133.60
$-CF_3$	133.21	126.50	127.608	131.53	129.14	125.62	126.59	129.71	135.43	131.42
–Cl	133.34	126.63	122.13	128.37	128.32	128.41	126.09	125.40	131.03	135.00
–Br	122.86	129.59	130.51	126.57	128.11	126.22	127.10	126.03	131.64	134.28
–I	99.707	138.24	124.98	126.66	128.83	127.02	127.28	129.31	134.92	134.81
–OH	151.68	108.77	109.39	121.26	127.60	126.40	125.89	118.72	124.33	133.67
$-OCH_3$	150.88	110.77	109.88	123.28	127.59	126.25	125.4	120.49	126.09	134.42
$-OCOCH_3$	150.88	110.77	109.88	123.28	127.59	126.25	125.47	120.49	126.09	134.42
$-NH_3^+$	123.75	121.14	120.87	126.96	128.68	128.07	128.35	120.54	126.13	134.65
–CN	109.03	131.08	131.60	131.83	127.15	126.10	126.91	125.23	130.84	131.49
–CHO	130.80	136.66	135.80	134.79	128.38	126.51	128.28	124.39	129.98	133.26
$-COCH_3$	134.81	128.77	129.46	132.59	128.39	126.30	128.29	124.49	130.09	133.77
$-COOCH_3$	127.22	130.45	127.77	134.13	128.71	126.17	128.48	126.12	131.72	134.09
–COCl	129.27	136.53	136.57	137.62	129.73	128.04	130.02	125.91	131.51	134.64
Valid										
–H	128.32	126.02	125.83	126.08	127.82	125.72	126.53	128.04	133.65	133.68
$-CH_3$	133.50	126.22	125.96	128.14	128.02	125.42	125.10	123.83	132.42	133.27
$-C(CH_3)_3$	146.06	123.16	122.22	127.56	129.64	123.53	123.59	126.42	132.03	135.84
–F	159.28	109.72	124.58	128.64	128.75	127.95	127.81	118.67	124.26	135.79
$-NH_2$	141.44	109.19	111.05	121.23	128.48	125.81	125.46	117.72	123.31	134.04
$-N(CH_3)_2$	151.83	114.75	115.52	122.89	129.04	126.29	125.57	124.12	129.72	135.75
$-NO_2$	146.87	123.94	124.33	131.94	128.42	127.12	128.76	119.35	124.96	134.32
–COOH	125.75	129.23	129.08	131.50	127.88	125.23	126.54	124.70	130.28	132.57
$-CON(CH_3)_2$	134.60	123.73	123.91	128.43	128.34	126.34	126.91	123.87	129.47	133.33
$-Si(CH_3)_3$	132.33	129.001	127.87	131.74	129.12	126.25	126.64	131.06	136.52	132.00

2.2. Multivariate image analysis descriptors

In the MIA-QSPR method, the descriptors are the pixels of images that can be two or three dimensional. These pixels are correlated with dependent variables for making QSPR models. The 2D structures of each compound of Table 1 were systematically drawn in the Chem Sketch program [20], and then, converted to bitmaps in 100×105 pixels workspace, with resolution of 81×81 points in.$^{-1}$. All the drawn molecular structures were systematically fixed in a given coordinate. In our dataset, the pixel located at the 53×48 coordinate (on the carbon number 9), was used as reference in the alignment step, as illustrated in Fig. 1. Each 2D image was read and converted into binaries (double array in Matlab [21]). Each image of dimension 100×105 pixels was unfolded to a $1 \times 10{,}500$ row and then the 25 images were grouped to form a $25 \times 10{,}500$ matrix. Columns with zero variance were removed to minimize memory, reducing the size of matrix to 25×962.

2.3. Principal component regressions

In QSAR/QSPR studies, a regression model of the form $y = Xb + e$ may be used to describe a set of predictor variables (X) with a pre-

Table 6
The statistical parameters for RC-PCR, EV-PCR and RC-PC-ANN models.

Method	Set	C-1		C-2		C-3		C-4		C-5		C-6		C-7		C-8		C-9		C-10	
		R^2	SE	R^2	SE	R^2	SE	R^2	SE	R^2	SE	R^2	SE	R^2	SE	R^2	SE	R^2	SE	R^2	SE
RC-PCR	Training	0.856	5.690	0.804	4.475	0.799	0.467	0.830	0.329	0.830	0.329	0.667	0.568	0.875	0.505	0.836	1.300	0.833	1.309	0.827	0.603
	Valid	0.855	4.232	0.881	2.829	0.872	0.396	0.833	1.967	0.825	0.283	0.909	0.355	0.778	0.767	0.881	1.600	0.881	1.637	0.898	0.371
EV-PCR	Training	0.163	13.707	0.493	7.194	0.493	0.707	0.763	2.442	0.445	0.594	0.465	0.584	0.021	0.974	0.418	2.449	0.408	2.470	0.538	0.985
	Valid	0.128	10.375	0.587	5.281	0.471	0.804	0.554	3.216	0.622	0.416	0.630	0.411	0.375	0.930	0.679	2.624	0.699	2.601	0.315	0.961
RC-PC-ANN	Training	0.987	1.706	0.922	2.712	0.991	0.093	0.948	1.137	0.970	0.135	0.974	0.164	0.952	0.294	0.960	0.521	0.995	0.238	0.998	0.073
	Valid	0.980	2.132	0.997	0.527	0.999	0.016	0.998	0.129	0.999	0.001	0.985	0.012	0.963	0.149	0.999	0.192	0.997	0.287	0.999	0.033
	Test	0.954	2.115	0.999	0.063	0.999	0.038	0.999	0.041	0.999	0.025	0.936	0.329	0.982	0.304	0.994	0.360	0.998	0.197	0.998	0.054
ChemDraw	All	0.881	4.953	0.986	1.093	0.610	0.579	0.979	0.690	0.842	0.281	0.822	0.425	0.855	0.573	0.538	1.808	0.768	1.634	0.832	0.538

Table 7
Calculated values of ^{13}C chemical shifts using RC-PC-ANN method for all carbon positions—training, validation and test sets.

Substituent *X*	C-1	C-2	C-3	C-4	C-5	C-6	C-7	C-8	C-9	C-10
Training										
$-CH_2Br$	132.00	127.00	124.88	128.69	128.50	125.81	126.20	125.03	130.79	133.59
$-CH_2OH$	136.20	125.00	125.22	128.51	128.50	125.59	126.00	125.36	130.98	133.37
$-CF_3$	128.00	124.60	124.08	129.25	129.00	126.70	127.90	128.97	134.59	129.65
–Cl	131.90	126.10	125.68	128.36	128.20	128.97	126.70	125.29	130.76	134.56
–Br	122.60	129.50	125.69	128.42	127.90	126.33	126.90	126.00	131.58	134.16
–I	99.60	138.20	127.44	128.75	129.40	127.51	128.50	129.04	134.82	134.85
–OH	151.50	108.70	125.76	127.39	127.60	126.38	126.20	118.68	124.32	133.54
$-OCH_3$	150.95	110.80	125.44	127.74	127.65	126.36	125.65	120.61	126.12	134.42
$-OCOCH_3$	150.95	110.80	125.44	127.74	127.65	126.36	125.65	120.60	126.12	134.42
$-NH_3^+$	124.20	121.30	124.95	129.01	129.40	128.02	128.70	120.93	126.12	134.78
–CN	108.80	131.00	123.47	129.07	127.30	126.20	127.10	125.12	130.80	131.16
–CHO	130.90	136.70	124.46	129.51	128.20	126.55	128.60	124.48	129.98	133.26
$-COCH_3$	134.90	128.80	124.05	129.23	128.30	126.35	127.90	124.26	130.12	133.66
$-COOCH_3$	127.10	130.40	124.60	129.31	128.70	126.46	127.80	126.18	131.71	134.08
–COCl	129.20	136.50	125.39	129.90	129.90	128.00	130.40	125.71	131.48	134.59
Valid										
–H	126.06	125.99	125.89	128.87	128.00	125.88	125.88	128.00	132.97	133.63
–F	154.93	109.82	126.00	128.09	128.07	127.30	126.57	118.63	124.30	135.66
$-N(CH_3)_2$	145.78	114.70	126.47	128.00	128.98	126.30	125.51	123.75	129.65	135.67
$-CON(CH_3)_2$	133.75	123.83	125.11	129.04	128.39	126.30	126.89	123.81	129.23	133.37
$-Si(CH_3)_3$	137.81	129.74	125.53	129.19	129.17	125.10	125.19	131.39	137.33	133.84
Test										
$-CH_3$	115.14	126.32	126.47	128.44	128.30	125.19	125.28	123.89	132.26	133.38
$-C(CH_3)_3$	142.41	123.11	124.92	128.60	129.60	123.62	123.47	126.24	132.01	135.68
$-NH_2$	129.43	110.50	126.16	127.40	128.29	125.59	124.51	117.75	123.96	134.20
$-NO_2$	145.82	123.85	123.90	129.59	128.50	127.19	128.98	119.22	124.94	134.23
–COOH	107.03	129.13	123.52	129.29	127.40	125.00	126.35	124.80	130.45	132.76

dicted variable (y) by means of a regression vector (b). However, the colinearity, which often existed between independent variables, creates a severe problem in certain types of mathematical treatment such as matrix inversion [22]. A better predictive model can be obtained by ortogonalization of the variables by means of principal component analysis (PCA) and the consequent method is called principal component regression (PCR) [23–25].

To reduce the dimensionality of the independent variable space, a limited number of principal components (PCs) are used, and therefore a major question will arise after the PCA is, how many and which PCs constitute a good subset for predictive purposes? Hence, selecting the significant and informative PCs is the main problem in almost all of the PCA-based calibration methods [26–28].

Different methods have been addressed to select the significant PCs for calibration purposes. In the most common one which is called correlation ranking, the factors are ranked by their correlation coefficient with the property to be correlated (a dependent variable) [28]. The factor with highest correlation coefficient is considered as the most significant one and, subsequently, the factors are introduced into the calibration model until no further improvement of the calibration model is obtained. In the other method, which is called eigen value ranking, the factors are ranked in the order of decreasing eigen values and the factors with the highest eigen values are considered as the most significant factors.

In the present work, First PCA was carried out on data matrix using Minitab program [29]. After achieving PCs, two types of PCR analysis including correlation ranking based-PCR (CR-PCR) and eigen value ranking based PCR (EV-PCR) were employed. In the CR-PCR procedure, the scores of PCs were entered to the PCR model, consecutively, based on decreasing their correlation with the ^{13}C chemical shifts. A cut off value of $R^2 \geq 0.8$ was used to select the optimum number of PCs in the PCR models. The procedure for the EV-PCR method was similar to the CR-PCR method and the entrance of the PCs to the model was based on their decreasing eigen values.

For regression analysis, dataset was separated into two groups: training set including 15 compounds and validation set including 10 compounds. Training set was used for the construction of the PCR models and then the generated models were applied to the validation set. Obtained models were summarized in Tables 2 and 3 for CR-PCR and EV-PCR methods, respectively. In all 10 CR-PCR equations, the factor with highest correlation coefficient with the ^{13}C chemical shifts was considered as the most significant one and, subsequently, the factors were introduced into the cali-

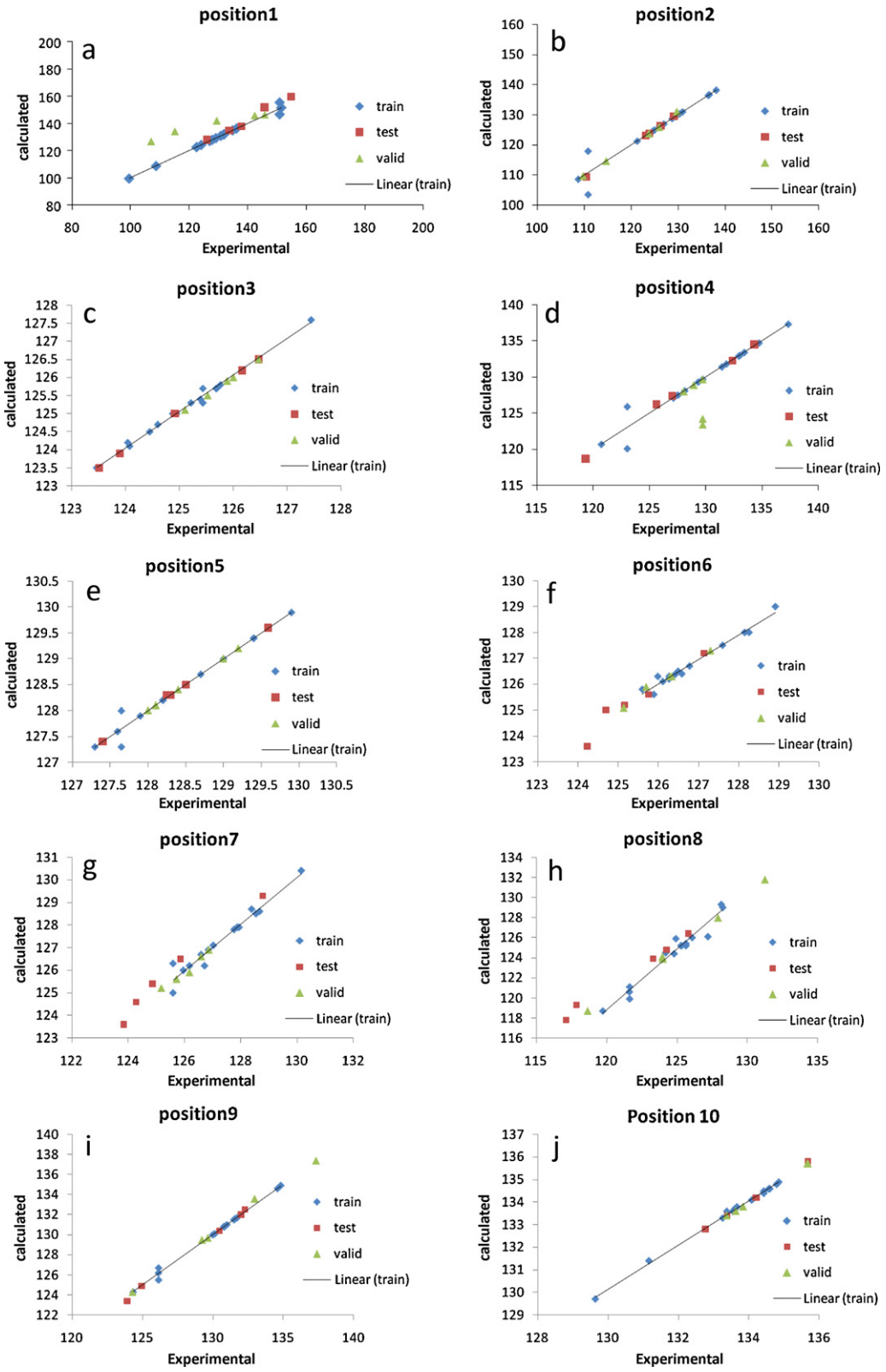


Fig. 2. (a–j) Plot of experimental ^{13}C chemical shifts of naphthalene derivatives against the calculated values using RC-PC-ANN model for C-1–C-10, respectively.

bration model until $R^2 \geq 0.8$ is achieved. PCs with higher correlation have greater information about the variation in the ^{13}C chemical shifts. In all 10 EV-PCR equations, the factors with highest eigen values were considered as the most significant factors and, subsequently, the factors were introduced into the calibration model until the number of factors in the EV-PCR models were identical to the number of factors in the CR-PCR models. Finally, obtained results using two CR-PCR and EV-PCR methods were compared. Calculated ^{13}C chemical shifts using CR-PCR and EV-PCR equations were shown in Tables 4 and 5, respectively.

Table 8
Calculated values of ^{13}C chemical shifts using ChemDraw program.

Substituent *X*	C-1	C-2	C-3	C-4	C-5	C-6	C-7	C-8	C-9	C-10
–H	128	125.9	125.9	128	128	125.9	125.9	128	133.6	133.6
$-CH_3$	134.9	126.8	126.9	126.5	128.6	125.6	125.8	125.5	132.6	133
$-C(CH_3)_3$	145.9	120.9	125.4	127.7	129.9	124.9	124.9	128.4	132.1	135.9
$-CH_2Br$	132	127.4	125.4	129.6	128.8	126.2	126.6	125	130.9	133.8
$-CH_2OH$	136.2	125.1	125.7	128.4	128.8	126	126.4	125.1	131.1	133.7
$-CF_3$	126.1	124.6	124.5	133.3	129.3	127.1	128.3	126.2	131.8	129.8
–F	159.2	109.6	126.4	124.5	128.4	127.7	127	121.2	124.3	135.4
–Cl	132.1	127.2	126.1	127.4	128.5	129.4	127.1	124.7	130.6	134.7
–Br	122.6	130.4	126.1	127.8	128.2	126.7	127.3	127	131.7	134.3
–I	98.6	138.3	128	130	129.7	127.9	128.9	132.7	134.8	135
–OH	153.7	110.8	126.2	121	127.9	126.8	126.6	123	126.2	134.7
$-OCH_3$	156.9	104	126.1	120.4	127.6	126.6	125.4	123.5	125.5	134.5
$-OCOCH_3$	146.6	118.7	125.7	126.2	128.3	126.7	126.7	122.7	128.6	134.6
$-NH_2$	143.7	109.6	127.6	119	128.6	126	125	121	123.6	134.3
$-N(CH_3)_2$	151.3	117.6	127.9	123.7	129.3	126.7	126	125.1	131	135.8
$-NH_3^+$	148.4	122.7	127.4	128.3	128.3	126.3	126.3	128.3	133.9	133.7
$-NO_2$	146.5	123.8	125.3	134.8	128.8	127.6	129.7	123.2	125.6	134.3
–CN	110	132.5	124.9	132.1	127.6	126.5	127.5	123.8	131.7	131.5
–CHO	132.5	137.7	126	135	128.5	126.9	129	124.8	130.1	133.4
$-COCH_3$	134.8	128.8	125.7	133.2	128.6	126.7	128.3	126.3	130.2	133.9
–COOH	127	129	125	132.6	127.7	125.4	126.9	125.1	132	134.5
$-COOCH_3$	127.6	129.9	126.2	133.7	129	126.8	128.2	126.5	133.3	134.2
$-CON(CH_3)_2$	134.8	123.3	126.6	129.2	128.7	126.7	127.3	128.4	129.6	133.5
–COCl	128	136	126.9	137.6	130.2	128.4	130.8	126.2	131.6	134.7
$-Si(CH_3)_3$	138.7	133.5	125.9	130	129.5	125.5	125.6	128.4	137.1	135.2

2.4. Artificial neural network modeling

Because of the complexity of the relationships existed between the activity/property of the molecules and the structures, nonlinear modeling methods are often used to model the structure–activity/property relationships. Artificial neural networks (ANNs) as non-parametric non-linear modeling techniques have attracted increasingly interest in the recent years [30,3]. Multilayer feedforward neural networks (MLF-ANN) trained with back-propagation learning algorithm become increasingly popular techniques [3,30–32]. The flexibility of ANN for discovering a more complex relationship causes that this method find wide application in QSAR/QSPR studies, which reviewed by Duch et al. [33].

The principal component-artificial neural network (PC-ANN), which combines the PCA with ANN and models the non-linear relationships between the PCs and dependent variable, was proposed by Gemperline to improve the training speed and decrease the overall calibration error [34].

At the present work, comparison of the statistical parameters for two RC-PCR and EV-PCR methods showed the superiority of RC-PCR method over the EV-PCR method (Table 6). Therefore, we used the PCs which were selected by RC-PCR method as input variables of ANN.

An artificial neural network with back-propagation algorithm was constructed. Our network had an input layer, a hidden layer and an output layer. The input vectors were the set of PCs which were selected by correlation ranking procedure. The number of nodes in the input layer depended on the number of PCs in the PCR equations. The number of nodes in the hidden layer was optimized through learning procedure. The training, validation and test datasets including 15, 5, and 5 compounds, respectively, were used to optimize the network performance. Obtained results using RC-PC-ANN method were shown in Table 7. For comparison, R^2 and standard error (SE) of different models for training, validation and test sets were summarized in Table 6.

3. Results and discussion

Table 1 lists the names of the compounds used in this study and their corresponding experimental ^{13}C chemical shift values. In this list the experimental ^{13}C chemical shift values for 10 carbon positions have been accessed. In order to find a correlation between MIA descriptors and these spectroscopic data, after eliminating the descriptors with zero variance, 962 MIA descriptors were remained. Then, PCA was applied on the descriptors data matrix. Twenty-three PCs were generated which were considered as the input variables of PCR and PC-ANN models. For each carbon position, separate PCR models based on eigen value ranking and correlation ranking were obtained. Obtained models were shown in Tables 2 and 3. Calculated values of ^{13}C NMR chemical shifts using these RC-PCR and EV-PCR equations were indicated in Tables 4 and 5, for training and validation sets, respectively. The statistical parameters of these models were summarized in Table 6. It should be noted that the obtained results using CR-PCR procedure shows superior qualities than those obtained by EV-PCR models. CR-PCR models show good performances and could predict the ^{13}C chemical shifts of the related molecules with low error.

To increase the predictive ability of the obtained models, a non-linear modeling method was employed. Typically, superior models can be found using ANNs because they implement non-linear relationships and because they have more adjustable parameters than the linear models. Therefore, we suggested the use of ANN as the non-linear model. As previously mentioned, obtained results by the CR-PCR procedure were more accurate than the EV-PCR procedure. The order of PCs based on their decreasing correlation was shown in equations of Table 2. Thus, these subsets of PCs were used as input of ANN models. The calculated values of ^{13}C chemical shifts using ANN models were represented in Table 7 for training, validation and test sets, respectively.

R^2 and SE values using three different methods (RC-PCR, EV-PCR and RC-PC-ANN) were summarized in Table 6. As can be seen from this table, RC-PC-ANN model shows more predictive ability than the PCR models. This indicates that there are nonlinear relationship between PCs and ^{13}C chemical shifts. Plots of experimental ^{13}C chemical shifts versus calculated values using RC-PC-ANN method for all 10 carbon positions are shown in Fig. 2(a–j), respectively. As it is observed, obtained models by the RC-PC-ANN method indicate high qualities. This means that there are non-linear relationships between the proposed MIA descriptors and the ^{13}C chemical shifts of the naphthalene derivatives.

Although the main aim of the present study was to investigate relationship between 2D images and ^{13}C chemical shifts, ^{13}C chemical shifts of the studied compounds were calculated using ChemDraw program [35]. Obtained values were shown in Table 8. For comparison, statistical parameters of these values were indicated in Table 6. As can be seen from this table, calculated ^{13}C chemical shifts using RC-PC-ANN models are more accurate than the calculated values by ChemDraw program.

Also, obtained results in this work indicated that though MIA descriptors do not have a direct physicochemical meaning, but may provide useful information and are capable to predict the ^{13}C chemical shifts of studied compounds.

4. Conclusion

As the conclusion, the proposed correlation ranking procedure for factor selection in PC-ANN algorithm produced perfect models for MIA-QSPR study of ^{13}C chemical shifts of naphthalene derivatives. In comparison with eigen value factor selection method, it was obtained that the EV-PCR method could not predict ^{13}C chemical shifts, accurately. It can be concluded that factor selection for ANN by the correlation ranking is more straightforward than the eigen value ranking. Also, obtained results indicated that MIA descriptors are capable to recognize the physicochemical information and may be useful to predict ^{13}C chemical shifts.

References

[1] K.A. Blinov, Y.D. Smurnyy, T.S. Churanova, M.E. Elyashberg, A.J. Williams, Chemometr. Intell. Lab. Syst. 97 (2009) 91.
[2] B. Hemmateenejad, K. Javadinia, M. Elyasi, Anal. Chim. Acta 592 (2007) 72.
[3] B. Hemmateenejad, M. Shamsipur, R. Miri, M. Elyasi, F. Foroghinia, H. Sharghi, Anal. Chim. Acta 610 (2008) 25.
[4] P. Gedeck, R.A. Lewis, Curr. Opin. Drug Des. 11 (2008) 569.
[5] M.P. Freitas, Curr. Comput. Aided Drug Des. 3 (2007) 235.
[6] K.L. Jensen, A.S. Barber, G.W. Small, Anal. Chem. 1991 (1991) 1081.
[7] V. Kvasnicka, S. Sklenak, J. Pospichal, J. Chem. Inf. Comput. Sci. 32 (1992) 742.
[8] L. Chen, W. Robien, Anal. Chem. 65 (1993) 2282.
[9] S. Thomas, D. Stro¨hl, E. Kleinpeter, J. Chem. Inf. Comput. Sci. 34 (1994) 725.
[10] D. Svozil, J. Pospichal, V. Kvasnicka, J. Chem. Inf. Comput. Sci. 35 (1995) 924.
[11] L.S. Anker, P.C. Jurs, Anal. Chem. 64 (1992) 1157.
[12] D.L. Clouser, P.C. Jurs, Anal. Chim. Acta 321 (1996) 127.
[13] D.L. Clouser, P.C. Jurs, J. Chem. Inf. Comput. Sci. 36 (1996) 168.
[14] B.E. Mitchell, P.C. Jurs, J. Chem. Inf. Comput. Sci. 36 (1996) 58.
[15] M. Goodarzi, M.P. Freitas, T.C. Ramalho, Spectrochim. Acta Part A 74 (2009) 563.
[16] P. Geladi, K. Esbensen, J. Chemometr. 3 (1989) 419.
[17] M.P. Freitas, Org. Biomol. Chem. 4 (2006) 1154.
[18] M.P. Freitas, S.D. Brown, J.A. Martins, J. Mol. Struct. 738 (2005) 149.
[19] E. Pretsch, P. Buhlmann, C. Affolter, Structure determination of organic compounds, Tables of spectral data, Ch. 4, Page 100.
[20] ACD/ChemSketch version 11.02, Advanced Chemistry Development, Inc., Toronto, Ont., Canada, 2008.
[21] MATLAB Version 7.1 Mathworks Inc., 2005 www.Mathworks.com.
[22] D.C. Montgomery, E.A. Peck, Introduction to Linear Regression Analysis, Wiley, New York, 1982.
[23] I.T. Jolliffe, Principal Component Analysis, Springer, New York, 1986.
[24] J.H. Kalivas, P.M. Lang, Mathematical Analysis of Spectral Orthogonality, Marcel Dekker, New York, 1994.
[25] G. Puchwein, Anal. Chem. 60 (1988) 569.
[26] Y.L. Xie, J.H. Kalivas, Anal. Chim. Acta 348 (1997) 19.
[27] J.M. Sutter, J.H. Kalivas, P.M. Lang, J. Chemometr. 6 (1992) 217.
[28] J. Sun, J. Chemometr. 9 (1995) 21.
[29] Minitab version 15.1.0.0, (2006) www.Minitab.com.
[30] H.F. Chen, Anal. Chim. Acta 609 (2008) 24.
[31] B. Hemmateenejad, M.A. Safarpour, F. Taghavi, J. Mol. Struct. (Theochem.) 635 (2003) 183.
[32] D.T. Manallack, B.G. Tehan, E. Gancia, B.D. Hudson, M.G. Ford, D.J. Livingstone, D.C. Whitley, W.R. Pitt, J. Chem. Inf. Comput. Sci. 43 (2003) 674.
[33] W. Duch, K. Swaminathan, J. Meller, Curr. Pharm. Des. 13 (2007) 1497.
[34] P.J. Gemperline, J.R. Long, V.G. Gregoriou, Anal. Chem. 63 (1991) 2313.
[35] ChemDraw program, www.cambridgesoft.com.