Journal of Chromatography A, 1216 (2009) 7055–7062

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Quantitative structure-retention relationship studies for taxanes including epimers and isomeric metabolites in ultra fast liquid chromatography

Pei-Pei Dong[a,b], Guang-Bo Ge[a,b], Yan-Yan Zhang[a,b], Chun-Zhi Ai[a,b], Guo-Hui Li[c,**], Liang-Liang Zhu[a,b], Hong-Wei Luan[a], Xing-Bao Liu[a], Ling Yang[a,*]

[a] Laboratory of Pharmaceutical Resource Discovery, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian 116023, China
[b] Graduate University of Chinese Academy of Sciences, Beijing 100049, China
[c] Laboratory of Molecular Modeling and Design, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian 116023, China

ARTICLE INFO

Article history:
Received 19 June 2009
Received in revised form 24 August 2009
Accepted 28 August 2009
Available online 1 September 2009

Keywords:
QSRR
Taxanes
Isomers identification
Ultra fast liquid chromatography
Monte Carlo variable selection
Artificial neural network

ABSTRACT

Seven pairs of epimers and one pair of isomeric metabolites of taxanes, each pair of which have similar structures but different retention behaviors, together with additional 13 taxanes with different substitutions were chosen to investigate the quantitative structure-retention relationship (QSRR) of taxanes in ultra fast liquid chromatography (UFLC). Monte Carlo variable selection (MCVS) method was adopted to choose descriptors. The selected four descriptors were used to build QSRR model with multi-linear regression (MLR) and artificial neural network (ANN) modeling techniques. Both linear and nonlinear models show good predictive ability, of which ANN model was better with the determination coefficient R^2 for training, validation and test set being 0.9892, 0.9747 and 0.9840, respectively. The results of 100 times' leave-12-out cross validation showed the robustness of this model. All the isomers can be correctly differentiated by this model. According to the selected descriptors, the three dimensional structural information was critical for recognition of epimers. Hydrophobic interaction was the uppermost factor for retention in UFLC. Molecules' polarizability and polarity properties were also closely correlated with retention behaviors. This QSRR model will be useful for separation and identification of taxanes including epimers and metabolites from botanical or biological samples.

1. Introduction

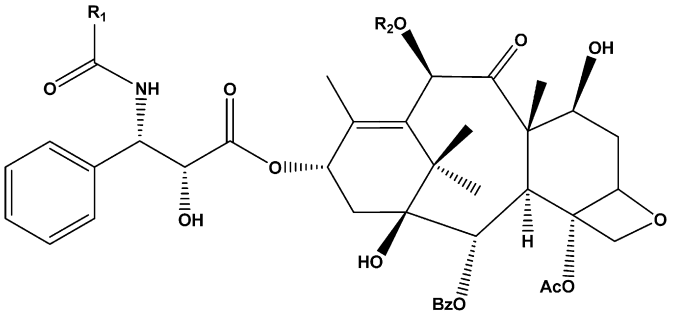
Taxanes, a class of biological compounds derived from yew trees, have been the target of intense investigations during the past 50 years [1–3]. It is well known that paclitaxel (Taxol™, Bristol–Myers Squibb Co.; Fig. 1) and docetaxel (Taxotere™, Sanofi–Aventis Co.; Fig. 1) are widely used for the treatment of many types of cancers [1]. Until now natural taxanes and their synthetic analogues are still the largest reservoir for compounds with better anti-cancer or anti-multidrug resistance activities [4,5]. Some natural taxanes such as 10-deacetylbaccatin III, 10-deacetyl-7-xylosylpaclitaxel and cephalomannine are valuable starting materials for production of paclitaxel and docetaxel to solve the supply crisis [6–9]. Thus, the isolation and identification of taxanes are quite important for comprehensive utilization of yew resources.

Up to now, liquid chromatography and its hyphened techniques were widely used for analysis of taxanes from botanical or biological samples [10,11], but some practical difficulties existed due to the complexity of practical samples. As for the isolation, the conventional HPLC usually suffered from long analysis time, low or medium resolution, and vast solvent consumption, as a result of the distinctions of polarity between taxanes deriving from structure diversity [12,13]. Concerning the identification, information from mass profiles may be insufficient owing to the existence of epimers and isomeric metabolites which have the same or similar mass spectra [4,14,15]. To some extent, the differences in retention times become an important factor for assignment of target taxanes in practice. Thus, obtaining retention times of target compounds by using analysis method with higher speed and better resolution was needed for analysis of taxanes. Recently, ultra performance liquid chromatography (UPLC) and ultra fast liquid chromatography (UFLC) have gained popularity due to their short analytical time and enhanced separation performance [16,17]. The UFLC system uses columns packed with fine particles (2.2 μm), and is operated at normal backpressure [18]. High separation speed and superior resolution make it suitable for the isolation and identification of taxanes from complex samples.

According to general chromatographic theory, the differences in retention behaviors of analytes are due to their different interactions with chromatography system, including the interactions with solid phase, components of mobile phase and so on. Therefore,

* Corresponding authors. Tel.: +86 411 84379317; fax: +86 411 84676961.
** Corresponding authors. Tel.: +86 411 84379593; fax: +86 411 84675584.
E-mail addresses: ghli@dicp.ac.cn (G.-H. Li), yling@dicp.ac.cn (L. Yang).

0021-9673/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2009.08.079



1 Paclitaxel: R_1=Ph, R_2=Ac
2 Docetaxel: R_1=^{t}BuO, R_2=H

Fig. 1. Molecular structures of paclitaxel and docetaxel.

finding out the primary interactions involved in and quantifying the interacting mode are quite important to understand the retention mechanism and predict retention times for new compounds. Quantitative structure-retention relationship (QSRR) is a technique correlating the variations in chromatography retention behaviors to the variations of compound structures encoded in several descriptors [19,20]. Thus, QSRR model is not only able to predict the retention times of new compounds just from their structures; but also useful to elucidate the retention mechanisms according to the selected descriptors [21–23]. The combination of QSRR with UFLC will definitely improve the efficiency of isolation and identification of taxanes. With this investigation, we intended to build a QSRR model which could predict the retention times of taxanes under the given condition in UFLC.

In this article, first, we chose seven pairs of epimers and one pair of isomeric metabolites into the data set. The characteristic of each pair of isomers is that small structural difference corresponds to a large gap for the retention times. For example, despite of the only difference in hydroxyl orientation at C-7 site, the 7α-stereoisomers always have longer retention times than their corresponding 7β-taxanes [24]. The same situation exists for the taxanes metabolites hydroxylated at parent ring and side chain [25]. These isomers were included to help find out structural information closely related to retention behaviors in UFLC, and confirming the interactions pivotal for chromatography separation process. Second, upon the optimized geometry of selected compounds, a large number of descriptors were calculated to characterize the structure differences of taxanes. Then a new variable selection method named Monte Carlo variable selection (MCVS) was adopted to select suitable descriptors subsets [26]. With the selected descriptors, linear and nonlinear models were built to mimic the interacting mode of UFLC system. The chosen descriptors were analyzed to explain the differences in retention behaviors of isomers. The retention mechanism was also discussed according to the selected descriptors.

2. Experimental

2.1. Solvents and chemicals

Twenty nine taxanes were used in this research (structures shown in supporting information). Taxanes authentic standards (purity >98%), including 10-deacetylbaccatin III (10-DAB), baccatin III (B), docetaxel (D) and paclitaxel (P), 3′-para-hydroxyl paclitaxel (C3′-p-OH-P) were purchased from Sigma–Aldrich Co. 6-hydroxyl paclitaxel (C6-OH-P) was from BD Gentest Corp. Some other taxoid standards (purity >95%), including 7-*epi*-10-deacetylbaccatin III (*epi*-DAB), 9-dihydro-13-acetylbaccatin III (9-DHB), 10-deacetyl-7-xylosylpaclitaxel (10-DAXP), 10-deacetylcephalomannine (10-DAC), 10-deacetylpaclitaxel (10-DAP), cephalomannine (C), 7-*epi*-10-deacetylcephalomannine (*epi*-DAC), 7-*epi*-10-deacetylpaclitaxel (*epi*-DAP), taxol C (T,c), 7-*epi*-baccatin III (*epi*-B), 7-*epi*-paclitaxel (*epi*-P) were purchased from Shanghai Jinhe Bio-technology Co. Ltd. Taxanes including 7-*epi*-docetaxel (*epi*-D), 7-*epi*-cephalomannine (*epi*-C), 2′-acetyl-paclitaxel (2′-AC-P), 10-deacetyl-10-propionylpaclitaxel (10-DAPP) were synthesized as reported [27]. 10-deacetyl-7-xylosyltaxol C (10-DAXT,c) and 10-deacetyl-7-xylosylcephalomannine (10-DAXC) were purified and characterized in house from needle extract of *T. Mairei*. 10-deacetyl-7-xylosyl baccatin III (10-DAXB), 4,10-dideacetylbaccatin III (4,10-DDAB), 4,10-dideacetyl-7-xylosyl baccatin III (4,10- DDAXB), 13-Oxo-10-deacetylbaccatin III (13-Oxo-DAB), 13-Oxo-10-deacetyl-7-xylosylbaccatin III (13-Oxo-DAXB) were synthesized according to published schemes [28,29]. 4″-hydroxylcephalomannine (C4″-OH-C) was prepared as reported previously [30]. Millipore water (Millipore Corp., Bedford, MA, USA) and LC grade methanol and acetonitrile (Tedia Co. Inc., USA) were used throughout; other reagents were analytical grade. Stock solutions (1.0 mg ml^{-1}) were prepared by dissolving accurately weighed 1 mg of each taxane in 1 ml of HPLC-grade methanol and stored at 4 °C. These were used to prepare the standard working samples by appropriate dilution.

2.2. Instrument and analytical conditions

A Shimadzu Prominence UFLC^{TM} system, equipped with a CBM-20A communications bus module, an SIL-20ACHT autosampler, two LC-20AD pumps, a diode array detector (DAD), a DGU-20A3 vacuum degasser, and a CTO-20A column oven was used for analysis of taxanes. A Shim-pack XR-ODS (50.0 mm × 2.0 mm I.D., particle size: 2.2 μm, Shimadzu) analytical column was used. The taxanes stock solutions was diluted to 100 μg ml^{-1} and used for working samples. The mobile phase consisted of water (A) and CH_3CN (B), and the gradient elution was carried out with the following profile: 0–17 min, 25%–60% B; 17–20 min, 90% B; 20–25 min, balanced to 25% B. The flow rate was 0.2 ml min^{-1} and the column temperature was kept at 40 °C, the injection volume was 1 μl. The scan wavelength was set from 190 to 370 nm and the detection wavelength was set at 230 nm. Data were processed with LCsolution software (Shimadzu, Japan) and the retention times were collected for further analysis.

3. Computation method

3.1. Data set

The name and relative retention times (normalized) of taxanes were listed in Table 1 (Chromatography in supporting information). The whole data set was grouped into two subsets based on the *k*-means clustering method [31]. About 80% of the data set was classified into the training set to select descriptors subset and build models; the remaining 20% was used as prediction set in multilinear regression. This 20% data set was separated into validation and test set for artificial neural network modeling.

3.2. Molecular descriptors

The molecular descriptors were calculated based on the geometry optimized with Hyperchem 6.03 Professional software (Hypercube, Gainesville, Florida). The optimization procedures were as follows: the starting conformations were obtained from the similar geometry of the crystal structure of paclitaxel [32]. Then the semi-empirical PM3 optimization was done using the Polak-Ribie're conjugate gradient algorithm with an RMS gradient of 0.001 kcal mol^{-1} Å^{-1} as convergence criterion [33]. Based on this optimized conformation, 1666 different molecular descriptors

Table 1
Experimental and predicted retention times (normalized) of taxanes.

No.	Compounds	$_N$RT[a]	MLR[b]	ANN[c]
Training set[d]				
1	2′-AC-P	0.99905	0.9726	1.0294
2	*epi*-D	0.96702	0.886	0.8835
3	*epi*-C	0.96336	0.9039	0.9098
4	T, c	0.95725	0.9357	0.9227
5	10-DAPP	0.94395	0.9574	0.9547
6	P	0.85168	0.7744	0.8531
7	*epi*-DAP	0.84408	0.9103	0.8940
8	D	0.82155	0.8323	0.8440
9	*epi*-DAC	0.79875	0.8318	0.8227
10	10-DAXT, c	0.71129	0.695	0.6774
11	10-DAP	0.69657	0.7414	0.6878
12	C3′-p-OH-P	0.64106	0.6977	0.6627
13	10-DAXP	0.62587	0.6664	0.6648
14	10-DAXC	0.57932	0.6239	0.6227
15	*epi*-B	0.56636	0.4821	0.5466
16	C4″-OH-C	0.53678	0.6024	0.5602
17	9-DHB	0.48093	0.473	0.4490
18	B	0.36464	0.3795	0.3536
19	13-Oxo-DAB	0.2919	0.3165	0.2976
20	13-Oxo-DAXB	0.22914	0.1926	0.2280
21	10DAB	0.14581	0.1681	0.1424
22	4,10-DDAB	0.02273	0.015	0.0232
23	4,10-DDAXB	0	−0.0187	0.0022
Prediction set[e]				
24	*epi*-P[f]	1	0.8312	0.9497
25	C[g]	0.81022	0.7847	0.8100
26	C6-OH-P[g]	0.72676	0.795	0.7685
27	10-DAC[f]	0.64785	0.8079	0.7594
28	*epi*-DAB[g]	0.33342	0.2753	0.3393
29	10-DAXB[f]	0.11304	0.1167	0.1100

[a] Normalized retention times by $_N\mathrm{RT} = (\mathrm{RT} - \mathrm{RT}_{min})/(\mathrm{RT}_{max} - \mathrm{RT}_{min})$.
[b] $_N$RT predicted by MLR model.
[c] $_N$RT predicted by ANN model.
[d] Molecules "1–23"were used as "Training set" in both MLR and ANN models.
[e] Molecules "24–29" were used as "Prediction set" in MLR model.
[f] Molecules used as "validation set" in ANN model.
[g] Molecules used as "test set" in ANN model.

were generated from E-DRAGON web server [34–36]. Additional 16 molecular descriptors were calculated using Hyperchem software as shown in Table 2 (Values of these descriptors listed in supporting information).

3.3. Monte Carlo variable selection (MCVS)

MCVS is based on Monte Carlo cross validation which performs the leave-group (n_v)-out cross validation over a large number of times ($100 \leq N \leq 100000$) [26,37]. It is especially useful when the number of descriptors is much greater than that of compounds. The average root-mean-squared error of prediction (RMSEP) of N times cross validation is then used to evaluate the predictive ability of models. Two kinds of heuristic search algorithms (simulated annealing and genetic algorithm) can be applied to select suitable combination of descriptors subset. The probability of chosen subset to achieve lowest average RMSEP was approximately calculated and utilized as the criteria for selection subset. The advantage of MCVS is that it allows a clear and simple statistical interpretation of the results, and it is equally compatible with the MLR-based or non-MLR-based quantitative structure-activity relationship (QSAR) models. The detailed description of the MCVS method can be found elsewhere [26]. In this article, the QSARBENCH v2 program was used for our study [38].

Table 2
Descriptors computed by Hyperchem.

Et: total energy; Hf: heat of formation;
μ: dipole moment; V: molecular volume; R: Refractivity
HOMO: energy of the highest occupied molecular orbital
LUMO: energy of the lowest unoccupied molecular orbital
MP: Mean Polarizability
P_{XY}, P_{XZ}, P_{XX}, P_{YZ}, P_{YY}, P_{ZZ}: polarizability components which measure the induced dipole moment in the i direction by a field in the j direction
SA_P: solvent accessible surface area calculated by a faster more approximate method
SA_G: solvent accessible surface area calculated using a grid method

Before the feature selection, some pretreatments were performed for the training set. The E-DRAGON error codes were replaced with zero, the constant and duplicate columns were deleted. With these operations, about 1300 descriptors were left for further analysis. Then the compounds (in rows) were sorted by their retention values, and the descriptors (in columns) were sorted by their correlation to the retention values column. After sorting, the retention values and descriptor columns were normalized to the [0,1] range.

Subsequently, Monte Carlo variable selection (MCVS) was performed to select suitable descriptor subsets using genetic algorithm which can avoid getting trapped in local minima [26]. The number of iterations for leave-group (n_v)-out Monte Carlo cross validation was 100,000. The size of validation set n_v was assigned to be half of the training set, i.e. 12 [39]. The number of descriptors k was examined from 1 to 10. From parsimonious consideration, the following variable selection schedule was proposed: $k = 1$ was considered at first, then it was consecutively increased until an "acceptable" P-value (normally $P \leq 0.05$) associated with current k was obtained. If subset with an acceptable P did not exist, the data set was trimmed to eliminate the descriptors cluster, and then the MCVS repeated. The simplest trimming method was adopted: first, the descriptors were sorted in descending order based on their correlation $|r_{jj'}|$ to retention value column, then starting from the second descriptor, a descriptor will be removed if it is correlated more strongly than some chosen threshold (r_{max}) to the remaining descriptors, i.e. $|r_{jj'}| > r_{max}$. To find the most suitable subsets, r_{max} was examined from 0.9 to 0.5.

3.4. Model building

The retention behavior of compounds is complex, which involves in several kinds of inter and intra molecular interactions. Therefore, besides the MLR, nonlinear ANN was adopted to explore the relationship between those descriptors and retention times.

3.4.1. Artificial neural network (ANN)

ANN has been widely used in QSRR modeling field, its principles, functioning and applications have been adequately described elsewhere [40,41]. In this article, some fully connected three-layer error back propagation neural networks with sigmoid transfer function were constructed. The number of neurons in input layer was set to be the number of descriptors chosen by MCVS. Levenberg–Marquardt algorithm was adopted to optimize weights and biases because it was significantly faster than other algorithms based on gradient descent [42]. For ANN modeling, the dataset was separated into three groups: training, validation, and test sets. It is noteworthy that the training data set was the same as that of MLR model, and the molecules in validation and test sets were just identical with those selected as prediction set in MLR model. Three compounds in prediction set were randomly selected as validation set to take care of the overfitting problems [40,43]. The test set, consisting of the remaining three molecules, was used to evaluate the generated models. The early-stopping method was adopted to avoid overtraining and overfitting. An internally developed C language program was used for modeling.

Table 3
Results of Monte Carlo variable selection and statistical results of Monte Carlo cross validation[a].

Model	Descriptors[b] subset	F[c]	RMSE	P value	R^2_{cv}[d]	R^2[e]
1	ALOGPS_log P	218.909	0.1025	0.1916	0.873	0.913
2	ALOGP	153.400	0.1222	0.9205	0.823	0.907
3	ALOGPS_log P ($\|r_{ij'}\| > 0.9$)[f]	218.909	0.1024	0.0344	0.873	0.913
5	[ALOGPS_log P] HATS0m	189.832	0.082	0.7181	0.917	0.950
6	ALOGPS_log P HATS0m RDF070m	218.144	0.0699	0.697	0.940	0.972
7	ALOGPS_log P HATS0m P_{XY} ($\|r_{ij'}\| > 0.9$)	172.464	0.0744	0.4974	0.932	0.965
8	ALOGPS_log P HATS0m X1A P_{XY}	240.203	0.0622	0.7514	0.950	0.982
9	[ALOGPS_log P HATS0m P_{XY}] R6e+ ($\|r_{ij'}\| > 0.9$)	199.144	0.069	0.7477	0.942	0.978
10	[ALOGPS_log P] BEHp7 X1AHATS0m P_{XY}	350.818	0.052	0.7686	0.967	0.988
11	[ALOGPS_log P] PJI2 PJI3 DISPm G3u Ds ($\|r_{ij'}\| > 0.9$)	583.316	0.0443	0.5804	0.98	0.995
12	[ALOGPS_log P] RDF125v R1u R7u+ MATS5v G3e R7e	658.704	0.0425	0.969	0.997	0.997
13	[ALOGPS_log P] R3e+ZZMor01m Gm L1p P_{XY} Mor09e	909.393	0.0483	0.7802	0.97	0.998
14	[ALOGPS_log P] RDF065u ESpm13r E1p G3u Mor27m Mor28v Mor29m Mor27u	2100.658	0.0446	0.7808	0.94	0.999

[a] In this table, brackets denote preselected or fixed descriptors.
[b] The explanation of descriptors except P_{XY} was in Ref. [36].
[c] F statistic was calculated on the training data set.
[d] R^2 of Monte Carlo cross validation ($N = 100000$).
[e] R^2 of the MLR models built using the selected descriptors subsets.
[f] The data set was trimmed by $|r_{ij'}| \leq 0.9$.

3.5. Model validation

The following parameters were calculated to evaluate the performance of models: Q^2_{cv} (cross-validation square correlation coefficient), RMSEP (root-mean-square error of prediction), R^2 (square correlation coefficient for regression line of the experimental vs. predicted activity), R^2_0 (square correlation coefficients for regression line through the origin), and K (the slope of regression line through the origin). The residuals between predicted and experimentally derived activities were also calculated. The propositional criteria necessary for high predictive ability of a model were high Q^2_{cv} (at least > 0.5), high R^2 for external test set (at least > 0.6), $(R^2 - R^2_0)/R^2 < 0.1$, and $0.85 \leq K \leq 1.15$ [44,45]. Moreover, the leave-12-out cross validation was randomly performed 100 times to evaluate the robustness of ANN model.

4. Results

4.1. Principal component analysis (PCA)

The PCA has been performed to explore the correlation among 1682 molecular descriptors and evaluate data splitting method. The first two PCs can explain 73.02% of the total data variation (65.91 and 7.11 respectively), which suggested a high correlation among these molecular descriptors. Eight PCs were needed to explain 90% of data variance (Fig. 2a). Fig. 2b described the loading plot of the first two PCs, which also showed high correlation of those descriptors. Thus, feature selection was quite necessary before QSRR modeling. The scores plot of first three PCs was depicted to show the space locations of 29 taxanes (Fig. 2c). It can be seen that all the samples in the training and prediction sets were well scattered over the whole space, which demonstrated that the data splitting method was reliable for evaluation of predictive ability of QSRR models.

4.2. Monte Carlo variable selection

The MCVS was operated a lot of times to look for subsets with acceptable P-value (≤0.05) and smallest RMSEP. For $k = 1$, descriptors ALOGPS_log P and ALOGP were chosen as the best descriptors, and ALOGPS_log P was much better according to the statistical results. After trimming the data set by $|r_{ij'}| > 0.9$, descriptor "ALOGPS_log P" obtained acceptable P-value 0.0344 (Table 3). This result proved that descriptor "ALOGPS_log P" was the best single descriptor. For $k = 2$–10, even trimming the data set by $|r_{ij'}| > 0.5$, subset with P-value ≤0.05 was still not to be found. So for each "k", we recorded the subsets with the smallest P-value and/or the smallest RMSEP. The MCVS results were listed in Table 3, statistical results of Monte Carlo cross validation were also recorded. Without P-value criterion, we adopted the following method to determine optimum number of descriptors [46]: if the inclusion of new descriptor could not improve the statistic result of the model, it was deemed that the optimum number of descriptor subset has been achieved. The increase of "$R^{2\prime\prime}$ less than 0.01 was chosen as the breakpoint criterion. Based on Table 3, it can be concluded that four descriptors were already enough for modeling. The correlation matrix of these descriptors was shown in Table 4. From this table, it could be clearly seen that descriptor "X1A" had a moderate correlation with another two descriptors. So we fixed the other three descriptors, and performed MCVS again. Descriptor "R6e+" was selected as the best addition (trimmed by $|r_{ij'}| > 0.9$). Table 5 showed the correlation matrix of these new descriptors and no significant correlation was observed.

4.3. MLR models

For comparison purpose, the MLR models were built using descriptor subsets: "ALOGPS_log P", "ALOGPS_log P + HATS0$_m$", "ALOGPS_log P + HATS0m + P_{XY}" and "ALOGPS_log P + HATS0m + P_{XY} + R6e+". The built models were used to predict the external prediction set (results shown in supporting information). The statistical characteristics of MLR model using four descriptors

Table 4
The correlation matrix of the selected four descriptors.

	ALOGPS_log P	HATS0$_m$	P_{XY}	X1A
ALOGPS_log P	1.0000			
HATS0m	−0.7035	1.0000		
P_{XY}	0.3494	−0.2177	1.0000	
X1A	0.8548	−0.8668	0.2526	1.0000

Table 5
The correlation matrix of the selected descriptors after trimming data set.

	ALOGPS_log P	HATS0$_m$	P_{XY}	R6e+
ALOGPS_log P	1.0000			
HATS0$_m$	−0.7035	1.0000		
P_{XY}	0.3494	−0.2177	1.0000	
R6e+	−0.7066	0.7946	−0.1434	1.0000

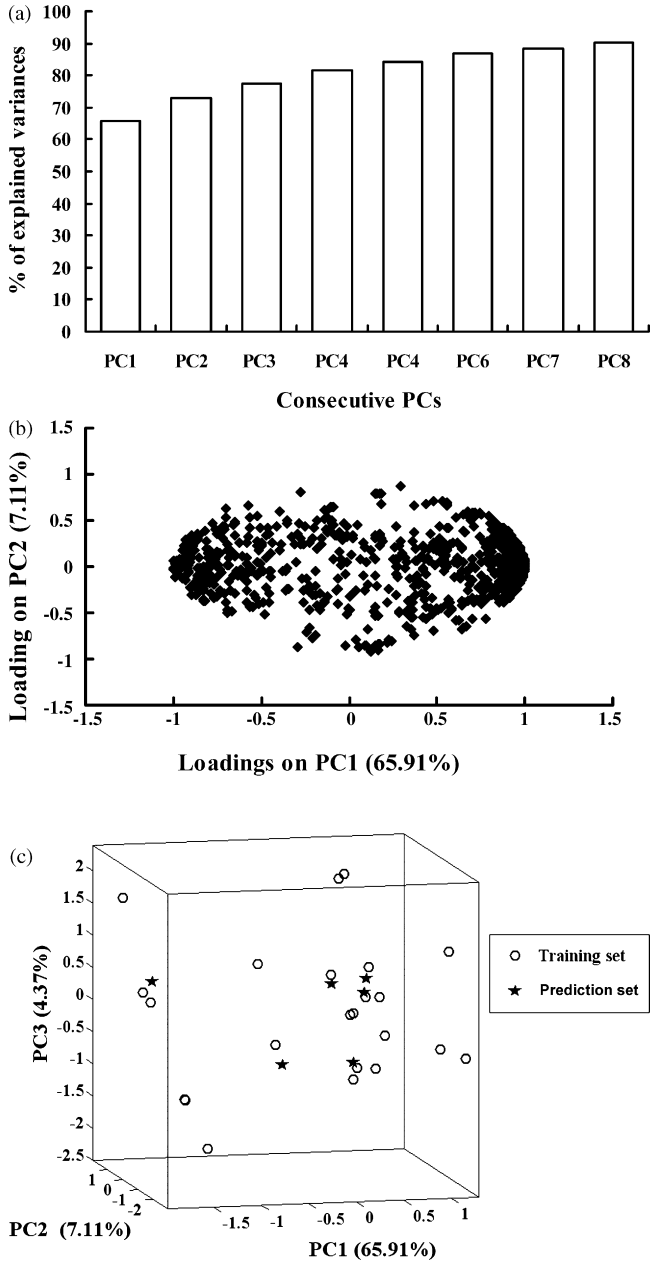


Fig. 2. (a) Cumulative percent of explained data variance by consecutive PCs; (b) PC1–PC2 loading plot; (c) PC1–PC2–PC3 scores plot.

were listed in Table 6 and the predicted values for all the taxanes were given in Table 1. According to the criteria for a good model mentioned above, the MLR model using four descriptor chosen by MCVS method had satisfactory predictive ability. The plot of experimental vs. predicted retention times (normalized) by MLR using four descriptors were shown in Fig. 3. The identification of isomers can be chosen as a judgment about if a QSRR model can reproduce the "translation" function between structures and retention times realized by the chromatography system. The predicted results for all pairs of isomers were depicted as shown in Fig. 4.

4.4. ANN model

In order to explore the nonlinear relationship between retention times and the selected descriptors, ANN technique was adopted to build models. A 4-3-1 back propagation artificial neural network model was developed. The parameters such as the number of nodes for hidden layer, learning rate, and momentum were optimized using the validation set. The ability to generalize the model was evaluated by an external test set. The predicted retention times values (normalized) for all the data set were listed in Table 1. The scatter plot of experimental retention times vs. predicted ones was

Table 6
The molecular descriptors and statistical characteristics of the MLR model.

Descriptors	Chemical meaning	Coefficient	Sig.
Constant	Intercept	0.133 ± 0.06	
ALOGPS_log*P*	Computed logarithm of octanol-water partition coefficient	1.009 ± 0.06	0.038
$HATS0_m$	GATAWAY descriptor (leverage-weighted Autocorrelation of lag 0 / weighted by atomic masses)	−0.444 ± 0.07	0.000
P_{XY}	Mean polarizability component	−0.154 ± 0.04	0.004
R6e+	GATAWAY (R maximal autocorrelation of lag 6 /Weighted by atomic Sanderson electronegativities)	0.219 ± 0.07	0.000

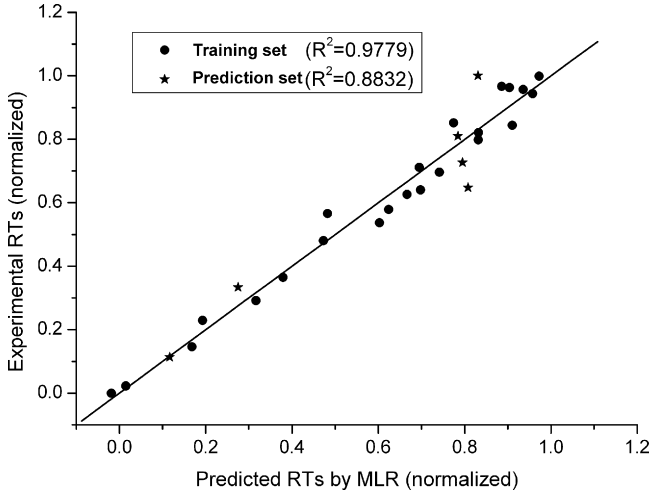


Fig. 3. Plot of experimental vs. predicted retention times (normalized) by MLR.

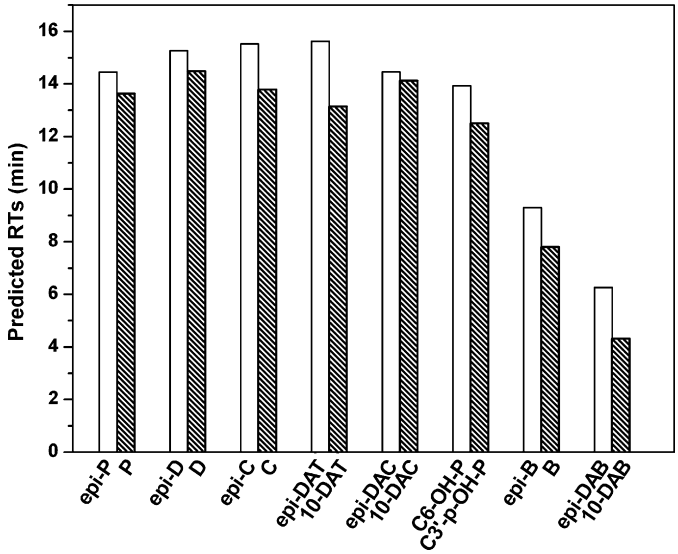


Fig. 4. Histogram of the predicted retention times for all pairs of isomers by MLR using descriptors ALOGPS_log*P* + $HATS0_m$ + P_{XY} + R6e+.

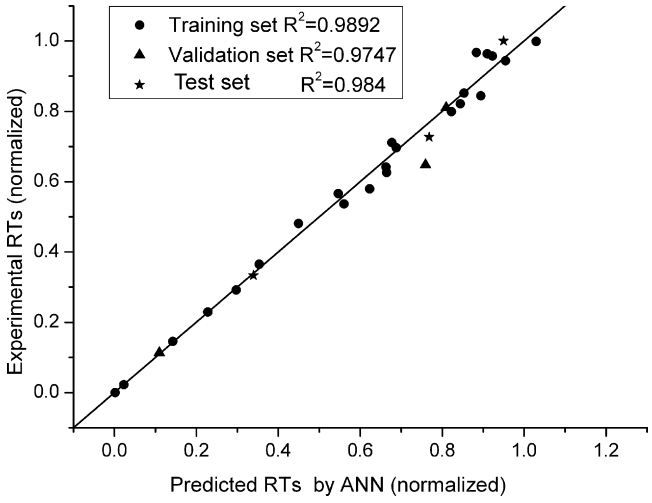


Fig. 5. Plot of experimental vs. predicted retention times (normalized) by ANN.

shown in Fig. 5. It was evident that the predicted values agreed well with experimental values. The statistical results of ANN model were listed in Table 7, and all the results were in accord with the criteria for a good predictive model. In order to compare the MLR model with ANN, the validation and test set in ANN were evaluated together. The better results of ANN than MLR model as shown in Table 7 demonstrated the complexity of chromatography retention process.

Fig. 6 showed a plot of the residuals vs. experimental retention times for ANN model. The residuals were equally distributed on both sides of zero line which indicates that no symmetric error exists in the development of our ANN model. 10-DAC was found with the largest predicted error. The reason may be that few cephalomannine derivatives were included in training set due to lack of commercial available standards. The predicted retention times for all pairs of isomers of ANN model were visualized in Fig. 7. It can be seen that all of the isomers have been successfully discriminated using this nonlinear model, which testified the effectiveness of this model again.

Table 7
Statistical results of the MLR and ANN models.

Model	Data set	Q^2_{LOO}	RMSE	R^2	R^2_0	$(R^2 - R^2_0)/R^2$	K
MLR	Training	0.9641	0.0450	0.9779	0.9779	0	1
	Prediction		0.1023	0.8832	0.8817	0.0017	0.9971
ANN	Training	0.9748	0.0315	0.9892	0.9892	0	1.0005
	Validation		0.0645	0.9747	0.9746	0.0001	0.9324
	Test		0.0379	0.9840	0.9813	0.0003	1.0086
	Prediction[a]		0.0529	0.972	0.9717	0.0003	0.9753

[a] To compare with MLR results, the validation and test set of ANN model were also analyzed together.

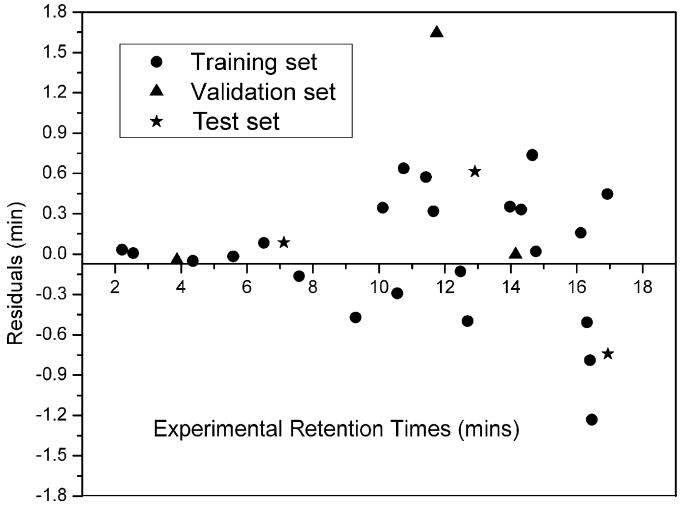


Fig. 6. Plot of residuals vs. experimental retention times for the ANN model.

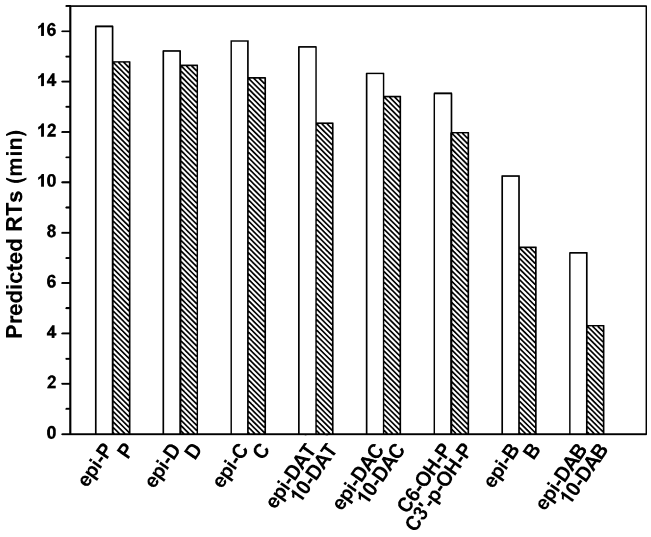


Fig. 7. Histogram of the predicted retention times for all pairs of isomers by ANN using descriptors ALOGPS_log P + HATS0$_m$ + P_{XY} + R6e+.

4.5. Leave-12-out cross validation

In order to examine the stability of this ANN model, leave-12-out cross validation was performed for 100 times, which produced 100 models, by randomly deleting 12 objects from the training set each time as test set. Calculation of R^2 and RMSEP was done using the results from these 100 models. The statistical results were considered as a benchmark of predictive ability of the model. The cross validation R^2 was 0.942, and RMSEP was 0.0736, which proved the robustness of our ANN model.

5. Discussion

Besides prediction of retention times, our QSRR model was also adopted to probe the retention mechanism according to these selected descriptors. In the MCVS results, "ALOGPS_log P" was the best single descriptor with statistical meaning. It is the logarithm of octanol–water partition coefficient calculated by free software ALOGPS2.1. It is well known that log P is considered as the measure of lipophilic ability of molecules. The choice of it as best descriptor showed that interaction between analytes and hydrocarbon chain, i.e. hydrophobic interaction was the most important factor determining the retention times of taxanes in reversed phase UFLC system. The largest positive coefficient of this descriptor in MLR model shows that increase of molecules' lipophilic property will prolong their retention times. As for the isomeric metabolites, owing to the different hydroxyl sites, they were easy to be differentiated by using this descriptor only. However, this descriptor was of no use in the differentiation of epimers (supporting information). This is because "ALOGPS_log P" was calculated based on the electrical-topological state descriptors which did not take the three dimensional (3-D) geometrical information into consideration [47]. Therefore, the taxanes and corresponding *epi*-taxanes always hold the same values for this descriptor. Additional descriptors were needed for correct description of the

minor differences between pairs of epimers. According to former researches [27,48,49], there existed an intramolecular hydrogen bond between the C-7 α-hydroxyl and C-4 carbonyl of *epi*-taxanes, which made a big difference between the 3-D conformation of taxanes and their corresponding epimers. Thus, descriptors correlated with 3-D conformation will be helpful. Fortunately, the next selected descriptor $HATA0_m$ was an H-GETAWAY (GEometry, Topology and Atom-Weights AssemblY) descriptor [50]. It was based on a leverage matrix, which was called "Molecular Influence Matrix", and was a new molecular representation calculated from the spatial coordinates of atoms in a chosen conformation. As deriving from the geometry of molecule, this descriptor was sensitive to the conformational change and bond lengths that account for atom types and bond multiplicity [51]. This character was exactly what we need to differentiate epimers. Inclusion of this descriptor has improved the predictive ability of MLR model as shown in Table 3. The $RMSE_{CV}$ decreased from 0.1024 to 0.082, and the R^2 have increased from 0.913 to 0.950. And some of the epimers which have no side chains were correctly differentiated although it was not good enough (supporting information). Compounds with side chain may be involved in some other interactions. The next included descriptor was polarizability which is the induced dipole moment of a molecule by per unit of applied electric field. Polarizability was a tensor with components P_{ij}. Descriptor P_{XY} represents the induced dipole moment in the *X* direction by an electric field in the *Y* direction. Analytes' polarizability was mostly related with dipole-induced dipole interaction. The negative coefficient of P_{XY} means that increase of this descriptor value will favor the elution process. It is reasonable because the dipole-induced dipole attractions between an analyte and polar molecules of the eluents are obviously stronger than that between the same analyte and nonpolar ligands (mainly hydrocarbons) of stationary phase. The addition of this descriptor has improved the predictive ability of QSRR model remarkably as shown in Table 3, all the isomers have been differentiated (supporting information). This result testifies that the interaction of analytes with eluents is quite important for separation process in UFLC system. The last but not least important descriptor included was R6e+. It is an R-GETAWAY descriptor which combines the information provided by molecular influence matrix with geometric interatomic distances in the molecule. The remarkable character of R6e+ is that it is weighted by atomic Sanderson electronegativities. So it is a polarity descriptor which depicts the electronegativity distribution of a molecule in 3-D space. The positive coefficient of this descriptor indicates the possible presence of free silanols on the surface of the silica-based material, and the polar interaction between analytes and free silanols always increases the retention of taxanes in UFLC. The MLR model using four descriptors did not improve the differentiation results significantly as exhibited in Table 3. However, when the nonlinear ANN model was built, the ability of differentiation between epimers was improved and the accuracy of predicted retention times has been enhanced (Figs. 4 and 7). This reminds us that the retention process on UFLC was complicated and not simple summation of several interactions.

Based on above discussion, we can conclude that hydrophobic interaction between analytes and the hydrocarbon chain was the driving force for retention in UFLC. The isomeric metabolites were easy to be distinguished just by the log *P* descriptor. However, due to lack of 3-D conformational information, the computed log *P* could not embody the difference between epimers and their corresponding taxanes. This result reminded us that a new 3-D log *P* computation method was necessary. The discrimination of epimers was owing to the inclusion of descriptor related with geometry conformation. The dipole-induced dipole interactions between analytes and the mobile phase were usually in favor of elution process. Moreover, the polar interaction between analytes and free silanols was another important factor for retention of analytes. It can be seen that all the selected descriptors have physical and chemical meanings, and these descriptors can account for leading structural features responsible for the retention behavior of taxanes. The main interactions determining the retention of analytes in UFLC can be described by these descriptors. The robustness of our QSRR model and the exploration of retention mechanism demonstrated that this QSRR model can be used as an approximate surrogate of the UFLC system.

6. Conclusion

In the present study, pairs of isomers with similar structures but different retention behaviors were included to build QSRR models. Four descriptors were chosen as the best subset from a large pool of descriptors by using MCVS method. With these four descriptors, MLR and ANN have been used to build QSRR models for prediction retention times of taxanes in UFLC. Both models have shown good predictive ability, ANN model was found to be better. All the isomers can be differentiated by ANN model, and molecules' 3-D conformational information was critical for the differentiation of epimers. As for the retention mechanism, hydrophobic interaction of analytes with hydrocarbon chain was found to be the driving force for the retention in UFLC. Analytes' dipole-induced dipole interaction with eluents, polar interactions with free silanols were also considered as important factors influencing chromatography retention behaviors. In conclusion, MCVS method is effective for variable selection, and combining with ANN, a good QSRR model can be built. This model has grasped the primary essence for retention in UFLC. It will be useful for isolation and purification of taxanes, and also helpful for the identification of epimers and metabolites from complex biological samples combining with mass spectra.

Acknowledgments

The authors are grateful to the 973 program (2007CB707802) of the Ministry of Science and Technology of China, the Key Direction Project of the Chinese Academy of Sciences (KSCX2-YW-G-050) and the DICP Innovation Fund of Chinese Academy of Sciences for financial supports.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.chroma.2009.08.079.

References

[1] M.C. Wani, H.L. Taylor, M.E. Wall, P. Coggon, A.T. McPhail, J. Am. Chem. Soc. 93 (1971) 2325.
[2] M.L. Miller, I. Ojima, Chem. Rev. 1 (2001) 195.
[3] A. Montero, F. Fossella, G. Hortobagyi, V. Valero, Lancet Oncol. 6 (2005) 229.
[4] V.S. Parmar, A. Jha, K.S. Bisht, P. Taneja, S.K. Singh, A. Kumar, Poonam, R. Jain, C.E. Olsen, Phytochemistry 50 (1999) 1267.
[5] I. Ojima, M. Das, J. Nat. Prod. 72 (2009) 554.
[6] J.N. Denis, A.E. Greene, D. Guenard, F. Guerittevoegelein, L. Mangatal, P. Potier, J. Am. Chem. Soc. 110 (1988) 5917.
[7] K.V. Rao, J. Heterocycl. Chem. 34 (1997) 675.
[8] A.A.L. Gunatilaka, M.D. Chordia, D.G.I. Kingston, J. Org. Chem. 62 (1997) 3775.
[9] K.C. Nicolaou, Z. Yang, J.J. Liu, H. Ueno, P.G. Nantermet, R.K. Guy, C.F. Claiborne, J. Renaud, E.A. Couladouros, K. Paulvannan, E.J. Sorensen, Nature 367 (1994) 630.
[10] G. Theodoridis, G. Laskaris, C.F. de Jong, A.J.P. Hofte, R. Verpoorte, J. Chromatogr. A 802 (1998) 297.
[11] J.P. Liu, K.J. Volk, M.J. Mata, E.H. Kerns, M.S. Lee, J. Pharm. Biomed. Anal. 15 (1997) 1729.
[12] Y.G. Zu, Y.J. Fu, S.M. Li, R. Sun, Q.Y. Li, G. Schwarz, J. Sep. Sci. 29 (2006) 1237.
[13] S.M. Li, Y.J. Fu, Y.G. Zu, R. Sun, Y. Wang, L. Zhang, H. Luo, C.B. Gu, T. Efferth, J. Pharm. Biomed. Anal. 49 (2009) 81.
[14] D.G.I. Kingston, G. Samaranayake, C.A. Ivey, J. Nat. Prod. 53 (1990) 1.

[15] I. Royer, P. Alvinerie, J.P. Armand, L.K. Ho, M. Wright, B. Monsarrat, Rapid Commun. Mass Spectrom. 9 (1995) 495.
[16] N. Sun, G. Lu, M. Lin, G. Fan, Y. Wu, Talanta 78 (2009) 506.
[17] S.I. Kawano, Y. Inohana, K. Arakawa, H. Mikami, S.I. Yamaguchi, J. Liq. Chromatogr. Related Technol. 31 (2008) 23.
[18] M. Takahashi, M. Nishimura, W. Hedgepeth, Proceedings of the HPLC 2006, San Francisco, California June 17–22, 2006.
[19] R. Kaliszan, J. Chromatogr. A 656 (1993) 417.
[20] R. Kaliszan, Chem. Rev. 107 (2007) 3212.
[21] K. Heberger, J. Chromatogr. A 1158 (2007) 273.
[22] F. Ruggieri, A.A. D'Archivio, G. Carlucci, P. Mazzeo, J. Chromatogr. A 1076 (2005) 163.
[23] B. Hemmateenejad, M. Shamsipur, A. Safavi, H. Sharghi, A.A. Amiri, Talanta 77 (2008) 351.
[24] H.R. Dong, L.N. Luo, P.Y. Bi, Y. Zheng, J.C. Zhao, Anal. Lett. 38 (2005) 929.
[25] I. Gut, I. Ojima, R. Vaclavikova, P. Simek, S. Horsky, I. Linhart, P. Soucek, E. Kondrova, L.V. Kuznetsova, J. Chen, Xenobiotica 36 (2006) 772.
[26] D.A. Konovalov, N. Sim, E. Deconinck, Y.V. Heyden, D. Coomans, J. Chem. Inf. Model. 48 (2008) 370.
[27] Q.Y. Zheng, L.G. Darbie, X.Q. Cheng, C.K. Murray, Tetrahedron Lett. 36 (1995) 2001.
[28] S.H. Chen, J.F. Kadow, V. Farina, C.R. Fairchild, K.A. Johnston, J. Org. Chem. 59 (1994) 6156.
[29] M.Z. Hoemann, D. Vandervelde, G.I. Georg, L.R. Jayasinghe, J. Org. Chem. 60 (1995) 2918.
[30] J.W. Zhang, G.B. Ge, Y. Liu, L.M. Wang, X.B. Liu, Y.Y. Zhang, W. Li, Y.Q. He, Z.T. Wang, J. Sun, H.B. Xiao, L. Yang, Drug Metab. Dispos. 36 (2008) 418.
[31] J.T. Leonard, K. Roy, QSAR Comb. Sci. 25 (2006) 235.
[32] D. Mastropaolo, A. Camerman, Y.G. Luo, G.D. Brayer, N. Camerman, Proc. Natl. Acad. Sci. U.S.A. 92 (1995) 6920.
[33] S.F. Braga, D.S. Galvao, J. Mol. Graphics Modell. 21 (2002) 57.
[34] I.V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. Palyulin, E. Radchenko, N.S. Zefirov, A.S. Makarenko, V.Y. Tanchuk, V.V. Prokopenko, J. Comput. Aided Mol. Des. 19 (2005) 453.
[35] http://www.vcclab.org/lab/edragon/.
[36] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-Vch, Weinheim, 2000.
[37] Q.S. Xu, Y.Z. Liang, Y.P. Du, J. Chemom. 18 (2004) 112.
[38] www.dmitrykonovalov.org.
[39] J. Shao, J. Am. Stat. Assoc. 88 (1993) 486.
[40] J. Zupan, J. Gasteiger, Neural Networks in Chemistry and Drug Design, 2nd ed., Wiley-VCH, New York, 1999.
[41] Y.L. Loukas, J. Chromatogr. A 904 (2000) 119.
[42] M.T. Hagan, M.B. Menhaj, IEEE Trans. Neural Network 5 (1994) 989.
[43] K. Asadpour-Zeynall, N. Jalili-Jahani, J. Sep. Sci. 31 (2008) 3788.
[44] A. Golbraikh, A. Tropsha, J. Mol. Graphics 20 (2002) 269.
[45] A. Tropsha, P. Gramatica, V.K. Gombar, QSAR Comb. Sci. 22 (2003) 69.
[46] A.R. Katritzky, V.S. Lobanov, M. Karelson, CODESSA: TrainingManual, University of Florida, Gainesville, Florida, 1995.
[47] I.V. Tetko, V.Y. Tanchuk, A.E.P. Villa, J. Chem. Inf. Comput. Sci. 41 (2001) 1407.
[48] G.N. Chmurny, B.D. Hilton, S. Brobst, S.A. Look, K.M. Witherup, J.A. Beutler, J. Nat. Prod. 55 (1992) 414.
[49] Q. Gao, W.L. Parker, Tetrahedron 52 (1996) 2291.
[50] A.A. D'Archiuio, M.A. Maggi, P. Mazzeo, F. Ruggieri, Anal. Chim. Acta 628 (2008) 162.
[51] V. Consonni, R. Todeschini, M. Pavan, J. Chem. Inf. Comput. Sci. 42 (2002) 682.