Talanta 81 (2010) 1711–1718

Contents lists available at ScienceDirect

Talanta

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

Advanced QSRR modeling of peptides behavior in RPLC

K. Bodzioch[a,b], A. Durand[a], R. Kaliszan[b,c], T. Bączek[b,c], Y. Vander Heyden[a,*]

[a] Department of Analytical Chemistry and Pharmaceutical Technology, Center for Pharmaceutical Research (CePhaR), Vrije Universiteit Brussel-VUB, Laarbeeklaan 103, B-1090 Brussels, Belgium
[b] Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk, Gen. J. Hallera 107, 80-416 Gdańsk, Poland
[c] Department of Biopharmacy, Collegium Medicum, Nicolaus Copernicus University, Bydgoszcz, Poland

ARTICLE INFO

Article history:
Received 22 October 2009
Received in revised form 11 March 2010
Accepted 18 March 2010
Available online 25 March 2010

Keywords:
HPLC retention
Peptides
Molecular descriptors
QSRR
Proteomics

ABSTRACT

In QSRR the retention is modeled as a function of structural or molecular descriptors. Since the structural datasets can be very large a selection of informative variables is often required. But beside the question which subset of variables (descriptors) produces optimum predictions one should answer the question: can good prediction be used in the QSRR community even if the physical meaning of applied descriptors is hard to interpret?

The main focus in this paper is put on different modeling methodologies applied and molecular descriptors used in the QSRR approaches. Besides the widely used multiple linear regression (MLR), these methodologies include partial least squares (PLS), uninformative variable elimination partial least squares (UVE-PLS), genetic algorithms (GA) prior to MLR or PLS. The comparison will focus on the predictive performance but also on the descriptors found to be most important for the chromatographic retention prediction of peptides. The results of this study showed that stepwise-MLR and UVE-PLS are producing better predictions than the rest of the studied methodologies. From the variables selected by various methodologies one can see that the important information for the retention mechanism of RPLC was given by 2D-, 3D-descriptors and descriptors from the empirical QSRR equations, which bring the information about hydrogen-bonding properties, molecular size, and complexity. Overall, for the considered data set the empirical QSRR models were predicting the peptides retention best.

1. Introduction

Reversed-phase liquid chromatography (RPLC) probably still is one of the most frequently used techniques utilized to separate mixtures in pharmaceutical and biomedical analyses. Owing to the wide range of stationary phases and the great variety of possible chromatographic systems the appropriate selection of a suitable starting point for method development has become more difficult. Usually screening or trial-and-error approaches are applied which are time consuming and cost demanding. In order to overcome this problem the mathematical models that are able to predict chromatographic retention from chemical structure has been extensively investigated over the last two decades. If one was capable to predict the retention of analytes and/or the separation of a mixture on chromatographic systems relatively well, then the theoretical approach could, in some part, replace the time consuming experimental. More recently also models were built to predict the retention of peptides based on their structural descriptors. They should increase the confidence in the identification of peptides in the context of proteomic analyses [1–4]. Among all prediction methods, quantitative structure–retention relationships (QSRR), which are statistically derived relationships between chromatographic parameters and descriptors characterizing the molecular structure of analytes, are the most popular [5–10].

To undertake a QSRR study one needs a set of quantitatively comparable retention parameters for a sufficiently large series of analytes and a set of their structural descriptors. The main problem in this area is that the number of structural descriptors which can be ascribed to an individual analyte is practically unlimited. Generally, they are classified as physicochemical, quantum chemical, theoretical (topological, geometrical) etc. The advantage of physicochemical descriptors is that they are generally clearly related to the retention, but they are often either unavailable or contain large errors. The advantage of quantum chemical descriptors is that they provide insights into the mechanism of chromatographic retention at the molecular or even submolecular level. Unfortunately, their correlation with retention is rather weak and they also are not easy to calculate. The third type of descriptors, the theoretical, are easily calculated, but they are not always evidently related to the specific retention phenomena [8,11].

Analyte descriptors calculated from the molecular formula or a molecular graph appear attractive, but the question is whether

* Corresponding author. Tel.: +32 2 4774734; fax: +32 2 4774735.
E-mail address: yvanvdh@vub.ac.be (Y. Vander Heyden).

0039-9140/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2010.03.028

they bear actual information on the property of an analyte or only on its symbolic representation. Therefore, two main approaches of choosing the molecular descriptors to include in a QSRR model exist. In the first, QSRR models are built from an *a priori* chosen small set of molecular descriptors, which are selected based on their physicochemical properties, well known to the chemists. In the second, QSRR models are derived starting from a large set of molecular descriptors (hundreds or thousands), from which the best are selected by means of variable selection methods or by the modeling technique itself.

From a chemical point of view, only those mathematical descriptors of a chemical structure are valuable, which could next be related to physicochemical properties of the analyte. Numerous nonempirical structural descriptors were reported to contribute to various multivariate QSRR models [12,13]. On the other hand, one might invent thousands of new descriptors without improving their statistical quality and physical meaning. The question thus arises whether a good prediction, which proves the validity of a given QSRR equation also if the physical meaning of the applied descriptors is vague is to be applied in a QSRR context (in the QSRR society), or whether it is better to build a QSRR model based on the descriptors with a well known physicochemical meaning even if the prediction ability might be less good than in the previous case? Of course the question only is to be answered for situation where the former models are clearly better.

Since the success of the QSRR depends mainly on the selection of the most informative descriptors of the analytes from a large sets of often mutually correlated descriptors, a suitable descriptor selection method is a key for proper model building.

Multiple linear regression (MLR) is without doubt the most frequently applied technique in building QSRR models, but also several other methodologies have more recently been applied, like partial least squares (PLS), uninformative variable elimination PLS (UVE-PLS), MLR or PLS combined with genetic algorithms (GA) for feature selection [1,14], classification and regression trees (CART) [15], multivariate adaptive regression splines (MARS) [16] and artificial neural networks (ANN) [17].

The scope of this work is to compare the two QSRR building strategies and to try to find an answer to the above posed question. Several modeling methodologies will be compared.

2. Theory

2.1. Stepwise multiple linear regression method for QSRR model building

As already mentioned, multiple linear regression (MLR) is very popular as technique to build the models in QSRR studies. In MLR, a regression analysis is carried out in order to obtain statistically significant models, where the retention in a given chromatographic system is presented as a function of a limited number of molecular descriptors.

QSRR models might be built from an *a priori* chosen small set of descriptors of known physicochemical properties or starting from a large set of potentially useful molecular descriptors. Normally, MLR as such cannot be used in the latter situation, because in most cases the number of available descriptive variables (descriptors) exceeds the number of objects (chromatographed substances). Therefore, prior to the building of a QSRR model a variable selection is needed. For that purpose a stepwise procedure can be used, where a forward descriptor selection iterates with a backward elimination. With stepwise-MLR the variables are selected step by step, from the original data matrix $\mathbf{X}$ (matrix of descriptors), taking into account their correlation with $\boldsymbol{y}$ (retention values). First, the variable that has the highest correlation with $\boldsymbol{y}$ is selected, and then a regression coefficient is obtained from the univariate regression model and its significance is tested using the *F*-test [18]. If this coefficient is significant the variable is included into the model. This step of including new variables into the model is called forward selection. After each inclusion a partial *F*-test is performed to test the significance of the variables that were already in the model. If variables are found that do not contribute significantly to the regression anymore, they are eliminated from the model. This step is called backward elimination. The entire process is repeated until no improvements of the model are achieved anymore by adding or removing variables. One of the drawbacks of the method is that the stepwise-MLR procedure is based on data fitting, i.e. the obtained model might be overfitted, which is disadvantageous for its predictive properties. To avoid this, cross-validation is applied, which tests the predictive capabilities of the models.

2.2. Partial least squares (PLS)

The partial least squares method (PLS) is based on indentifying a linear relationship between a dependent variable, $\boldsymbol{y}$, and a set of latent explanatory variables. The PLS approach helps to deal with the multicollinearity problem by replacing the original variables with a few latent orthogonal, so-called PLS factors. These factors are linear combinations of the original variables ($\mathbf{X}$) and maximize the covariance between $\mathbf{X}$ (the matrix of molecular descriptors) and $\boldsymbol{y}$ (the RPLC retention).

The optimal model complexity, i.e. the optimal number of the latent factors in the PLS model, can be determined by leave-one-out cross-validation (LOO-CV). Optimal complexity of the PLS model corresponds to the number of factors, for which the (nearly) minimal CV error appears.

2.3. Uninformative variable elimination partial least squares (UVE-PLS)

To eliminate irrelevant variables in QSRR model building the UVE-PLS is applied. The main idea of this approach is that the original data set $\mathbf{X}$ is augmented with a matrix of artificial random variables with normal distribution and very small amplitudes (absolute values of the order of 10^{-10}) to prevent their influence on the regression model. In our study different numbers of noise variables were simulated: 500, 250 and 125. The calculation showed that with different numbers of noise variables the optimal complexity of the model remains the same while the number of chosen descriptors and the root mean squared error of prediction (RMSEP) decreases with a decreasing number of noise variables. Moreover, to avoid overfitting and to validate the variable selection step, the elimination procedure was repeated 100 times. The final set of relevant variables was determined based on the frequency of choosing the individual variables in the different retained sets of variables. A retained variable was considered as relevant if it was retained in more than 70% of runs [19].

2.4. Genetic algorithms for variable selection (GA-MLR/GA-PLS)

For some techniques, such as MLR, data sets with more variables than objects are problematic. Genetic algorithms are adaptive heuristic search algorithms that are intended for feature selection on larger data sets. GA used in a combination with the MRL or PLS methods provide a subset of variables that are most suited for use within these techniques.

GA are global optimization procedures based on evolutionary computation and survival of the fittest. GA proceed first by randomly generating an initial population of objects, described by given variables, that through mutation, crossover and selection after a number of generations provides an optimal or near opti-

mal solutions. The size of this population remains then constant along the procedure. Each object is represented by a finite string of symbols (the genome) encoding a possible solution in the data space. At every iteration step (generation), the objects in the current population are tested according to some quality criterion (fitness function). To form a new population of objects (the next generation), they are selected according to their fitness. Selection alone cannot introduce new variables/objects into the population, which is necessary in order to make the solution as independent of the initial population as possible. New points (objects) in the search space are thus generated by two operations: crossover and mutation. In practice, new points in the search space are sampled randomly making possible to escape from local minima [20,21].

The strategy implemented for GA in MLR or in PLS regression can be described through the following steps. Objects (chromosomes) are defined as n-element (n-gene) vectors containing as many elements as descriptors were calculated. Each of these elements receives a binary code, 1 if the corresponding variable is selected, and 0 otherwise. The calculation starts with randomly coded objects, in this study 518, composing the initial population. The response of the population is evaluated numerically by cross-validation of the MLR or PLS models. The fittest objects are then selected. Afterwards, these objects undergo the reproduction step, which generates new objects. The reproduction step is composed of crossover and mutations. Crossover concerns two selected objects (parents) that randomly exchange parts of their genomes to form two new objects (children). Crossover can be done at one or several points (single crossover, double crossover, etc.) of the chromosomes, here the double crossover was performed. Mutations are random inversions of genes in chromosomes, that happen with a low probability, e.g. 0.005. The purpose of mutations is to give a chance to a variable that was not included in the initial random distribution, to appear in the coming generation. These different steps are repeated iteratively until the termination conditions of the algorithm are fulfilled, here 100 was the predefined maximum number of generations. Finally, as results might be dependent on the initial variable population constructed randomly, their validity should be checked by performing different runs starting from different initial distributions, in this study 5 replicate runs were performed [22,23]. In this study, the variables that were finally kept for PLS model building are those included in 80% or more of the GA selections.

2.5. Molecular descriptors

Molecular descriptors are currently of much interest in chemistry, pharmaceutical sciences, health research, etc. They can be defined as numbers, being the final result of a mathematical procedure, which are derived from translating the symbolic representation of the molecule into a useful numeric value (theoretical descriptor). They can also be the result of a standardized experiment (experimental descriptors). In this context the term "useful" means that the resulting number can contribute to a better understanding of molecular properties and/or can be used in a model to predict properties of chemical compounds.

The number of structural descriptors, which can be ascribed to an individual analyte, is practically unlimited. For example, the Dragon software which is frequently used in this context, calculates 3224 molecular descriptors [24]. Among the theoretical descriptors, depending on the initial representation of the molecule, one can distinguish zero- (0D), one- (1D), two- (2D), three- (3D) and four-dimensional (4D) descriptors. Since the 0D-descriptor is derived from the molecular formula, it is independent of the molecular structure. Examples are the number and type of atoms, the molecular weight and any function of the atomic properties (e.g. sum of van der Waals volumes).

If the molecule is represented considering its functional groups or substituents, the derived molecular descriptors are called 1D-descriptors, e.g. atom-centered fragments, functional group counts. When the topological representation of the molecule is considered, the resulting descriptors are called 2D-descriptors. They describe how atoms are connected in a molecule, their type of bonding and the interaction of particular atoms, e.g. total path counts. A 3D-representation of a molecule is called a geometrical representation and allows describing not only a representation of the nature and connectivity of the atoms, but also the overall spatial configuration of the molecule. The 4D-descriptors are derived from the stereoelectronic representation (or lattice representation) of the molecule. These descriptors result from the molecular representation, which is related to the molecular properties arising from the electron distribution–interaction of the molecule with probes characterizing the space surrounding them [11].

2.6. QSRR model statistics

The quantitative structure–retention relationship (QSRR) equations were derived by means of different modeling methodologies, employing the Matlab 7.0.1 software (The Mathworks, Natick, MA, USA), to describe the retention times (t_R) of the peptides based on the given set of molecular descriptors. Before QSRR model building, the descriptor values were autoscaled in order to remove undesired scale differences.

The obtained QSRR models require an estimation of their predictive performances, which was done using an independent test set. For all modeling methodologies the same training (50 peptides) and test (19 peptides) sets were used. The splitting of the data into training and test sets was done by random selection.

Selection of the optimal number of factors is crucial when constructing a model. Usually, model complexity is estimated using the cross-validation procedure, in its simplest variant, leave-one-out cross-validation (LOO-CV). This standard LOO procedure is based on taking out one sample from the entire data set as the hold-out case. Then a model is built on the remaining samples. The resulting model is used to predict the hold-out case. This entire process is repeated until each sample once became the hold-out case. The root mean squared error of cross-validation (RMSECV) is calculated as:

$$\mathrm{RMSECV}(f) = \sqrt{\frac{\sum_{i-1}^{m}(y_i - \hat{y}_i)^2}{m}}$$

where y_i is the experimental value of the dependent variable for the ith object, $\hat{y}_i$ is the predicted value for the ith object based on the model built with f factors, and m is the number of objects.

In order to characterize model fit and its prediction abilities, usually root mean squared error of prediction (RMSEP) is calculated as:

$$\mathrm{RMSEP}(f) = \sqrt{\frac{\sum_{i=1}^{m_t}(y_i^{\mathrm{test}} - \hat{y}_i^{\mathrm{test}})^2}{m_t}}$$

where m_t is the number of samples in the test set, whereas y_i^{test} and $\hat{y}_i^{\mathrm{test}}$ denote the experimental value and the predicted value from the model with f factors for the ith object from test set.

Both root mean squared error of cross-validation (RMSECV) with the leave-one-out (LOO) procedure and root mean squared error of prediction (RMSEP) were calculated with the use of Matlab 7.0.1 software.

Table 1
The set of 69 peptides studied. For amino acid abbreviations see text. Ac-NH: Acetylation; -$CONH_2$: amidation.

No.	Code	Amino acid sequence	No.	Code	Amino acid sequence
1	1d	AA	36	65p	EGVLY-$CONH_2$
2	2d	AG	37	67p	GLSPMIETIDQVR
3	3d	AF	38	69p	AGGYKPFNLETA-$CONH_2$
4	4d	YL	39	70p	GAPGGPAFPGQTQDPLYG-$CONH_2$
5	5d	DD	40	71p	Ac-NH-ETHLHWHTVAK-$CONH_2$
6	6d	ML	41	76p	LHWHT
7	7d	WW	42	77p	HLHWHT
8	8d	GM	43	79p	ETHLHWHT
9	9d	GH	44	82p	EVHHQK
10	10d	GL	45	83p	EVHHQKLVFF
11	11d	WF	46	86p	Ac-NH-EVHHQKLVFF
12	11t	GHG	47	87p	EVRHQKLVFF
13	11p	DRVYIHPF	48	88p	Ac-NH-EVRHQKLVFF
14	18p	HTVAKETS	49	89p	DAEFRH
15	20p	HWHTVAKETS	50	91p	DAEFGH
16	21p	LHWHTVAKETS	51	126p	GLFDVIKKVASVIGGL-$CONH_2$
17	27p	Ac-NH-HNPGYPHNPGYPHNPGY PHNPGYP-$CONH_2$	52	127p	DVIKKVASVIGGL-$CONH_2$
18	35p	EVHHQKLVFFAKDVGSNK-NH2	53	128p	IKKVASVIGGL-$CONH_2$
19	41p	DAEFRH-$CONH_2$	54	129p	KVASVIGGL-$CONH_2$
20	43p	DAEFGH-$CONH_2$	55	130p	GLFDVIKKVASVIGG-$CONH_2$
21	45p	DAEFRHDSG-$CONH_2$	56	131p	GLFDVIKKVASVIG-$CONH_2$
22	46p	DAEFGHDSG-$CONH_2$	57	132p	GLFAVIKKVASVIGG-$CONH_2$
23	47p	DAEFRHDSGY-$CONH_2$	58	133p	GLFAVIKKVASVIG-$CONH_2$
24	48p	Ac-NH-DAEFRHDSGY-$CONH_2$	59	134p	GLFDVIKKVASVI-$CONH_2$
25	49p	DAEFGHDSGF-$CONH_2$	60	135p	GLFDVIKKVAS-$CONH_2$
26	50p	Ac-NH-DAEFGHDSGF-$CONH_2$	61	136p	GLFDVIKKV-$CONH_2$
27	55p	EVHHQKLVFF-$CONH_2$	62	137p	GLFDVIK-$CONH_2$
28	57p	EVRHQKLVFF-$CONH_2$	63	138p	DVIKKVASVIG-$CONH_2$
29	58p	Ac-NH-EVRHQKLVFF-$CONH_2$	64	139p	IKKVASV-$CONH_2$
30	59p	LVFF-$CONH_2$	65	140p	GLFDVIKASVIGGL-$CONH_2$
31	60p	GSNKGAIIGLM-$CONH_2$	66	141p	GLFDVVIGGL-$CONH_2$
32	61p	GKTKEGVLY-$CONH_2$	67	142p	GLFAVIKKVASVI-$CONH_2$
33	62p	KTKEGVLY-$CONH_2$	68	143p	GLFDVIKKVASV-$CONH_2$
34	63p	TKEGVLY-$CONH_2$	69	144p	GLFAVIKKVASV-$CONH_2$
35	64p	KEGVLY-$CONH_2$			

3. Material and methods

3.1. Chemicals

Acetonitrile (ACN, HPLC grade) from Merck (Darmstadt, Germany) and trifluoroacetic acid (TFA) from Fluka (Buchs, Switzerland) were used. Water was prepared with a Milli-Q Water Purification System (Millipore Corporation, Bedford, MA, USA). All analyzed peptides are presented in Table 1. Angiotensin II and the 20 natural amino acids: alanine (A), arginine (R), asparagine (N), aspartic acid (D), cysteine (C), glutamic acid (E), glutamine (Q), glycine (G), histidine (H), isoleucine (I), leucine (L), lysine (K), methionine (M), phenylalanine (F), proline (P), serine (S), threonine (T), tryptophan (W), tyrosine (Y) and valine (V) were purchased from Fluka. Sodium dodecyl sulfate (SDS) and the following peptides were from Sigma–Aldrich (St. Louis, MO, USA): AA, AG, AF, YL, DD, ML, WW, GM, GH, GL, WF and GHG. All other peptides were synthesized at the Department of Organic Chemistry, University of Gdańsk, Poland [3]. The peptides studied were selected to assure a wide range of structural diversity, including posttranslational modifications of peptides (e.g. acetylation and amidation).

3.2. Equipment and chromatographic conditions

The chromatographic measurements were performed on four columns: (1) XTerra MS (XT), 15.0 cm × 0.46 cm i.d. (Waters, Milford, MA), packed with octadecyl-bonded silica; (2) Acclaim 300 C18, 3 μm, 5.0 cm × 0.46 cm i.d. (Dionex Corporation, Sunnyvale, CA, USA), packed with spherical particles of octadecyl-bonded silica; (3) Cadenza 5CD-C18, 5 μm, 7.5 cm × 0.46 cm i.d. (Imtakt Corporation, Kyoto, Japan), packed with 5 μm spherical particles of porous, octadecyl-bonded silica, and (4) Presto FT-C18, 3.0 cm × 0.46 cm i.d. (Imtakt Corporation), packed with 5 μm spherical particles of non-porous, octadecyl-bonded, end-capped silica.Chromatographic measurements were performed on a Merck-Hitachi LaChrom system (Merck-Hitachi, Frankfurt-Tokyo, Germany-Japan), equipped with a UV/vis detector (L-7400), autosampler (L-7200), thermostat (L-7360), pump (L-7100) and the software D-7000 HPLC System Manager, version 4.1. The ChemStation program was used for data collection.

Gradient HPLC elution was carried out with solvent A (water containing 0.10% trifluoroacetic acid) and solvent B (acetonitrile with 0.10% trifluoroacetic acid). The mobile phase used was filtered through a GF/F glass microfibre filter (Whatman, Maidstone, UK) and degassed with helium during the analysis. On the XTerra MS column the gradient during the analysis was formed from 0% to 60% B within 20 min. For the other columns the gradient was formed from 4% to 60% B, with a gradient time t_G of 20 min. All columns were thermostated at 40 °C. The injected sample volume was 20 μl. The chromatographic measurements were performed at an eluent flow rate of 1 ml/min. The eluent was monitored at a detection wavelength of 223 nm. Peptide samples were dissolved in 0.10% of aqueous trifluoroacetic acid (TFA).

3.3. Structural descriptors of the peptides

The experimental descriptor, $\log \mathrm{Sum}_{AA}$, was calculated as the logarithm of the summed individual amino acid gradient retention times measured under the same HPLC conditions as the peptides. The descriptor, $\log \mathrm{Sum}(k+1)_{AA}$ [2], was calculated as the logarithm of the sum of the $k+1$ values of the amino acids. For 13 non-retained amino acids (A, R, N, D, C, E, Q, G, H, K, P, S, T) the median k over the different systems studied in [2] was taken and rounded, resulting in $k=0$ for each of those amino acids. For the 7 retained amino acids,

their individual k on a given system was used. Not k but $k+1$ were used to avoid zero values for the non-retained factors.

Some theoretical descriptors for the peptides, i.e. the logarithm of the peptide's van der Waals volume, $\log VDW_{Vol}$, and the logarithm of its theoretically calculated n-octanol–water partition coefficient, $c\log P$, were calculated by the standard molecular modeling program HyperChem for personal computers with the extension ChemPlus (HyperCube, Gainesville, FL, USA). This software also performed geometry optimization of the peptide's structures using the molecular mechanics force field method (MM+) with the Polak–Ribière conjugate gradient algorithm and with an RMS gradient of 0.05 kcal/(Å mol) as stopping criterion. Dragon Professional 5.0 software version, 2004 (Milano Chemometrics and QSAR Research Group, Talete, Milano, Italy) was then used to calculate 1630 molecular descriptors, belonging to 20 classes, from the geometrically optimized peptide structures. Correlated descriptors ($r > 0.99$) and constant values were deleted in Dragon. As a consequence 514 descriptors were retained as possible predictor variables. Finally, with the addition of the four descriptors described above, 518 descriptors were used for the modeling.

4. Results

4.1. QSRR models built from an a priori chosen small set of molecular descriptors

In this case, QSRR models are build, based on a limited number of well understood descriptors, that can easily be linked to known physicochemical properties. Lately, a QSRR model has been proposed by Kaliszan et al. [25,26] to predict the gradient retention times of peptides under given HPLC conditions. This model employs the following structural descriptors: the logarithm of the sum of gradient retention times of the amino acids composing the individual peptide, $\log Sum_{AA}$; the logarithm of the peptide's van der Waals volume, $\log VDW_{Vol}$; and the logarithm of its theoretically calculated n-octanol–water partition coefficient, $c\log P$. This QSRR equation has the following form:

$$t_R = b_0 + b_1\ \log Sum_{AA} + b_2 \log VDW_{Vol} + b_3 c\log P \quad (1)$$

where t_R is the peptides gradient RPLC retention time and b_0–b_3 are regression coefficients estimated by MLR. To estimate $\log Sum_{AA}$, one needs the retention times of the 20 natural amino acids determined at the same HPLC conditions as the peptides.

Since, in previous studies [1,3,4,17,26,27], beside good predictive abilities of the above QSRR model, it was also observed that most amino acids were hardly retained at the applied RPLC conditions, i.e. they eluted close to the dead time, we proposed an alternative $\log Sum(k+1)_{AA}$ descriptor [2]:

$$t_R = b_0 + b_1\ \log Sum(k+1)_{AA} + b_2 \log VDW_{Vol} + b_3 c\log P \quad (2)$$

The predictive abilities of both QSRR models containing either $\log Sum_{AA}$ or $\log Sum(k+1)_{AA}$ are presented in Table 2. It can be noticed that the QSRR model containing the $\log Sum(k+1)_{AA}$ descriptor has similar predictive abilities, even slightly better than the QSRR model containing $\log Sum_{AA}$.

4.2. QSRR models derived starting from the large set of molecular descriptors

In this approach, to model the retention times of 69 peptides, six different methodologies were used: stepwise-MLR, PLS and PLS performed on only the 30% of descriptors that are the best correlated with $\boldsymbol{y}$ (in this case, t_R), UVE-PLS, GA-MLR and GA-PLS.

Table 2
Predictive abilities of the QSRR models with three structural descriptors (Eqs. (1) and (2)) of which one is either $\log Sum_{AA}$ or $\log Sum(k+1)_{AA}$. The RMSECV is the root mean squared error of cross-validation and was calculated by the leave-one-out procedure; RMSEP is the root mean squared error of prediction and R^2 is the determination coefficient (square of correlation coefficient).

Column		$\log Sum_{AA}$	$\log Sum(k+1)_{AA}$
XTerra MS	RMSECV	1.52	1.48
	RMSEP	1.99	1.95
	R^2	0.86	0.87
Acclaim 300 C18	RMSECV	1.98	1.96
	RMSEP	2.15	2.14
	R^2	0.80	0.80
	R^2	0.86	0.87
Cadenza 5CD-C18	RMSECV	1.51	1.49
	RMSEP	1.79	1.76
	R^2	0.87	0.87
	R^2	0.86	0.87
Presto FT-C18	RMSECV	2.60	2.60
	RMSEP	2.99	2.94
	R^2	0.64	0.64

4.2.1. Stepwise-multiple linear regression (stepwise-MLR)

Stepwise-MLR models, using autoscaled data and evaluated by LOO-CV, were built. The models obtained are constituted by four descriptors for XTerra MS and Cadenza 5CD-C18 columns, by seven for Acclaim 300 C18 and by five for Presto FT-C18 column (Table 3). Those models are characterized by RMSECV values between 1.03 and 1.86, RMSEP values between 2.11 and 2.95 and predictive R^2 values between 0.58 and 0.76 (Table 4). Two remarkable observations can be noticed. First of all, the $\log Sum(k+1)_{AA}$ descriptor is selected for all models out of 518 potential candidates. It demonstrates its high relevance and good correlation with the retention time. Secondly, the models derived from the full descriptor set have considerably smaller RMSECV values than the models described by Eqs. (1) and (2). However, their predictive properties do not seem to be better because their RMSEP are similar, or even slightly worse than those from Eqs. (1) and (2).

To be able to relate the selected descriptors to properties of molecules a short characterization of the descriptors is necessary. The description is done by groups, e.g. 2D autocorrelations, constitutional, geometrical, WHIM descriptors. Individual interpretation of given descriptors is often rather difficult, since for the theoretical molecular descriptors it is not always evident to find a link with physicochemical properties.

The descriptors chosen by stepwise-MLR belong to several different groups. *Constitutional descriptors*, like nRO5, are 0D-descriptors, which are the most simple and commonly used descriptors, independent from molecular connectivity and conformations, accounting for molecular composition, like atom and bond counts, molecular weight, sum of atomic properties, etc. [11]. Parameter nRO5 gives information on the number of 5-membered rings. *Information indices*, calculated as information content of molecules, based on the calculation of equivalence classes from the molecular graph. Among them, the most important are *topological information indices* (IVDE), which are graph-theoretical invariants, and can be considered as a quantitative measure of the lack of

Table 3
Descriptors selected by stepwise-MLR at QSRR model building.

Column	Selected descriptors
XTerra MS	$\log Sum(k+1)_{AA}$, $c\log P$, nRO5, HOMA
Acclaim 300 C18	$\log Sum(k+1)_{AA}$, $\log Sum_{AA}$, IVDE, MLOGP2, $c\log P$, E2u, MATS5v
Cadenza 5CD-C18	$\log Sum(k+1)_{AA}$, $c\log P$, nRO5, HOMA
Presto	$\log Sum(k+1)_{AA}$, $c\log P$, nRO5, RCI, MATS3e

Table 4
Predictive abilities of the QSRR models characterized by the root mean squared error of cross-validation (RMSECV) calculated by the leave-one-out procedure and by the root mean squared error of prediction (RMSEP) (n.a.: not applicable).

Column		Stepwise-MLR	PLS	PLS only 30% best correlated **X** with ***y***	UVE-PLS (100 times)	GA-PLS	GA-MLR	GA-MLR
XTerra MS	No. of chosen descriptors	4		160	10	58	1	7
	Cut-off value	Alpha 1%	n.a.	n.a.	n.a.	n.a.	Alpha 1%	Alpha 5%
	No. of latent variables	n.a.	9	6	6	4	n.a.	n.a.
	RMSECV	1.03	2.51	2.12	1.63	1.70	3.97	2.76
	RMSEP	2.11	3.49	3.12	2.73	3.51	4.05	4.92
	Predictive R^2	0.75	0.47	0.47	0.62	0.52	0.17	0.24
Acclaim 300 C18	No. of chosen descriptors	7	n.a.	167	9	68	3	4
	Cut-off value	Alpha 1%	n.a.	n.a.	n.a.	n.a.	Alpha 1%	Alpha 5%
	No. of latent variables	n.a.	9	4	6	3	n.a.	n.a.
	RMSECV	1.11	2.61	2.56	2.28	2.38	2.53	2.42
	RMSEP	2.19	3.42	2.99	2.65	3.48	3.57	3.56
	Predictive R^2	0.74	0.51	0.51	0.61	0.53	0.41	0.38
Cadenza 5CD-C18	No. of chosen descriptors	4	n.a.	158	8	48	1	5
	Cut-off value	Alpha 1%	n.a.	n.a.	n.a.	n.a.	Alpha 1%	Alpha 5%
	No. of latent variables	n.a.	9	6	6	4	n.a.	n.a.
	RMSECV	1.10	2.36	2.05	1.69	1.89	2.85	2.29
	RMSEP	1.95	3.27	2.88	2.28	2.85	3.96	4.29
	Predictive R^2	0.76	0.48	0.49	0.71	0.57	0.17	0.22
Presto FT-C18	No. of chosen descriptors	5	n.a.	177	17	44	1	4
	Cut-off value	Alpha 1%	n.a.	n.a.	n.a.	n.a.	Alpha 1%	Alpha 5%
	No. of latent variables	n.a.	4	4	2	5	n.a.	n.a.
	RMSECV	1.86	3.32	2.89	2.52	2.52	3.69	3.28
	RMSEP	2.95	3.52	3.31	3.21	4.81	4.07	3.67
	Predictive R^2	0.58	0.42	0.45	0.51	0.48	0.096	0.31

structural homogeneity or the diversity of the molecular graph and in this way they are related to symmetry associated with structure. The two descriptors, MATS5v and MATS3e, belong to *2D autocorrelations*, which are measures of the homogeneity of the molecular structure, i.e. of how the considered property is distributed along the topological structure. *2D autocorrelations* can be weighted by the atomic mass, polarizability, electronegativity, and van der Waals volumes (MATS5v – Moran autocorrelation, weighted by atomic van der Waals volumes, MATS5v – Moran autocorrelation, weighted by atomic Sanderson electronegatives). *2D autocorrelations* contain topological information able to capture structural complexity. In this study the *2D autocorrelation descriptors* describe correlations between the molecular structures and retention parameters.

Geometrical descriptors (3D descriptors), like HOMA (Harmonic Oscillator Model of Aromaticity index) and RCI (Jug RC index) view a molecule as a rigid geometrical object in a space and allow representation of not only the nature and connectivity of the atoms but also overall spatial configuration of the molecule. More precise, HOMA and RCI are aromaticity indices. HOMA encodes information of any conjugated system, and can be applied as a descriptor of both local and global aromaticity, while RCI are derived only from aromatic rings. *WHIM descriptors*, like E2u, are three-dimensional molecular indices calculated from the (x, y, z) atomic coordinates that represent different sources of chemical information. These descriptors try to capture relevant information about the whole 3D-molecular structure in terms of size, shape, symmetry and atom distribution, e.g. since *WHIM descriptors* are sensitive to any conformational change in the molecule they are able to distinguish different conformations of the same molecule.

MLOGP2 (*molecular properties* descriptor), which stands for squared Moriguchi octanol–water partition coefficient ($\log P^2$) gives information about the liphophilicity [11].

4.2.2. Partial least squares (PLS)

The partial least squares (PLS) models were constructed for each RPLC system separately. The retention predictions of those models are characterized in Table 4. Since the PLS methodology performed not so promising as it was expected (results were considerably worse than from MLR), the PLS analysis was performed again but only on the 30% of descriptors that are best correlated with y. It allowed to decrease the model complexity (number of PLS factors) and also improved the predictive performance of the method. These last models provide RMSECV values between 2.05 and 2.89, RMSEP values between 2.88 and 3.31 and predictive R^2 values between 0.45 and 0.51. Even though the latter models had improved results they are still worse than with MLR.

4.2.3. Uninformative variable elimination partial least squares (UVE-PLS)

The variable elimination procedure was applied to the complete data set of 518 descriptors. The UVE-PLS analysis was performed 100 times with the addition of 125 artificial variables. For the different RPLC columns different numbers of descriptors were retained during the elimination procedure (Table 4). Overall, the elimination procedure retained two constitutional descriptors, four 2D-descriptors, eight 3D-descriptors, three functional groups counts, two molecular properties and again $\log \mathrm{Sum}(k+1)_{AA}$, further $\log \mathrm{Sum}_{AA}$, and $\log \mathrm{VDW}_{vol}$.

The final PLS model was built using the training set of 50 samples and then tested with the remaining 19 samples. The optimal number of factors was determined using LOO-CV. The predictions quality obtained is presented in Table 4. Generally, the RMSECV values for those models are between 1.63 and 2.52, RMSEP values between 2.28 and 3.21 and predictive R^2 values between 0.51 and 0.71 (Table 4). These models are better than the previous PLS models but still worse than the MLR models.

The eight most important descriptors, i.e. the descriptors that present the highest weights over all four HPLC systems, are presented with a brief description in Table 5. 3D-MoRSE descriptors, like Mor08u, Mor08v, Mor08p, present information from the 3D atomic coordinates by using the same transform as in electron diffraction studies. They give an idea of how a weighting property is distributed. The GETAWAY descriptors (3D-descriptors) are sensitive to molecular branching and cyclicity and they also give information about the presence of significant substituents in the

Table 5
Descriptors selected by UVE-PLS at QSRR model building. Only these descriptors are listed which provide the highest weights over all four HPLC systems studied.

Selected descriptors	Class of descriptors	Description
Mor08u	3D-MORSE	3D-MoRSE-signal 08/unweighted
Mor08v	3D-MORSE	3D-MoRSE-signal 08/weighted by van der Waals volumes
Mor08p	3D-MORSE	3D-MoRSE-signal 08/weighted by atomic polarizability
R1u+	GETAWAY descriptors	R maximal autocorrelation of lag1/unweighted
nRSR	Functional groups counts	Number of sulfides
$\log \mathrm{Sum}(k+1)_{AA}$	Lately proposed in [2]	The logarithm of the sum of the $k+1$ values of the amino acids (see Section 3.3)
$\log \mathrm{Sum}_{AA}$	Proposed in [3,28]	The logarithm of the summed individual amino acid gradient retention times (see Section 3.3)
$\log \mathrm{VDW}_{vol}$	Proposed in [3,28]	The logarithm of the peptide's van der Waals volume

Table 6
Descriptors selected by GA at GA-PLS QSRR model building.

Column	0D descriptors	1D descriptors	2D descriptors	3D descriptors	Others	Empirical descriptors
XTerra MS			2: Topological descriptors	2: Randic molecular profiles	4: Functional group counts	$\log \mathrm{Sum}(k+1)_{AA}$
			3: Information indices	2: Geometrical descriptors	5: Atom-centered fragments	$\log \mathrm{Sum}_{AA}$
			8: 2D autocorrelcation	3: RDF descriptors		
			1: Edge adjacency indices	14: 3D-MoRSE descriptors		
			1: Topological charge indices	2: WHIM descriptors		
			1: Eigenvalue-based indices	8: GETAWAY descriptors		
Acclaim 300C18		1: Charge descriptors	2: Topological descriptors	1: Randic molecular profiles	5: Functional group counts	$\log \mathrm{Sum}(k+1)_{AA}$
		1: Molecular properties	5: Information indices	1: Geometrical descriptors	4: Atom-centered fragments	$\log \mathrm{Sum}_{AA}$
			12: 2D autocorrelcation	4: RDF descriptors		$\log \mathrm{VDW}_{vol}$
			1: Burden eigenvalues	12: 3D-MoRSE descriptors		
			1: Topological charge indices	4-WHIM descriptors		
				11: GETAWAY descriptors		
Cadenza 5CD-C18			1: Topological descriptors	1: Geometrical descriptors	3: Functional group counts	$\log \mathrm{Sum}(k+1)_{AA}$
			2: Information indices	2: RDF descriptors	2: Atom-centered fragments	$\log \mathrm{Sum}_{AA}$
			9: 2D autocorrelcation	5: 3D-MoRSE descriptors		$\log \mathrm{VDW}_{vol}$
			3: Topological charge indices	2: WHIM descriptors		
				15: GETAWAY descriptors		
Presto FT-C18	1: Constitutional descriptors	1: Charge descriptors	2: Topological descriptors	2: RDF descriptors	1: Functional group counts	$\log \mathrm{Sum}_{AA}$
			3: Walk and path counts	11: 3D-MoRSE descriptors		
			2: Information indices	2: WHIM descriptors		
			3: 2D autocorrelcation	10: GETAWAY descriptors		
			2: Edge adjacency indices			
			1: Topological charge indices			

molecule. From the descriptors with the highest weights the nRSR represent functional groups counts. This class of molecular descriptors is based on the counting of chemical functional groups, here sulfides [11].

4.2.4. Genetic algorithms for variable selection

Genetic algorithms as a variable selection step was used for both partial least squares (GA-PLS) and multiple linear regression (GA-MLR). As shown in Table 4, GA-MLR models present the worst predictive performances overall of all tested methodologies. Therefore, only the results of GA-PLS are discussed further.

For each chromatographic system, GA variable selection was carried out with the parameters described earlier (see Section 2.4). It resulted in different number of descriptors included in the models for different columns (Table 4). It can be observed that for all chromatographic systems mainly *2D autocorrelations*, 3D-MoRSE and GETAWAY descriptors were chosen (Table 6). The meaning of those groups of descriptors was given in Sections 4.2.1 and 4.2.3.

After the variable selection step, the partial least squares analysis was performed. The optimum number of latent variables to be included in the calibration model was determined from the RMSECV. The performance of the models was estimated by computing the RMSECV and RMSEP values which are presented in Table 4. The models provide RMSECV values between 1.70 and 2.52, RMSEP values between 2.85 and 4.81 and predictive R^2 values between 0.48 and 0.57 (Table 4). However, the predictive performance of these models is still worse than for the MLR models.

5. Discussion and conclusions

The comparison of the individual QSRR models, built starting from the large set of molecular descriptors, found that stepwise-MLR and UVE-PLS were both producing good predictions.

Table 4 shows that the GA-PLS models were performing similar to the PLS models, but in GA-PLS less complexity of the models was noted which goes also with a slight prediction improvement over the PLS models. Moreover, as it was expected, the PLS models with only the 30% of descriptors that are best correlated with ***y***

were performing better than the PLS models built starting from all descriptors (518 descriptors).

When one compares the prediction performance of the QSRR models built from an *a priori* chosen small set of descriptors and of the best QSRR models built starting from a large set of descriptors, it can be seen that the first models were performing slightly better. From the models built from large set of descriptors, the stepwise-MLR models also here were found to perform best. Moreover, the empirical QSRR equations (Eqs. (1) and (2)) employ well known descriptors which can easily be linked to physicochemical properties of the molecule, while methods like PLS, UVE-PLS, GA-PLS use combinations of the original variables (latent factors), which make understanding of the resulting equation rather impossible. Therefore, MLR or MLR with feature selection (stepwise-MLR) can be preferred in QSRR, especially when they also result in the best models.

It can be concluded from the variables selected by the several methodologies that important information for the retention mechanism of RPLC was given by 2D- and 3D-descriptors. Most of the important molecular descriptors selected account for hydrogen-bonding properties, molecular size, and -complexity. It must be stressed here that, in each methodology the descriptors from the empirical QSRR equations, especially $\log \mathrm{Sum}(k+1)_{\mathrm{AA}}$, also were retained.

For some of the theoretical descriptors selected, it is not evident to explain their meaning and relation to the chromatographic retention, but they are needed in order to obtain proper QSRR predictions.

It should be noted here that the worst predictive results overall were achieved on the Presto FT-C18 column. Apparently, this column is less suitable for a chromatographic characterization (modeling) of peptides and amino acids.

References

[1] R. Put, M. Daszykowski, T. Bączek, Y. Vander Heyden, J. Proteome Res. 5 (2006) 1618–1625.
[2] K. Bodzioch, T. Bączek, R. Kaliszan, Y. Vander Heyden, The molecular descriptor log SumAA and its alternatives in QSRR models to predict the retention of peptides, J. Pharm. Biomed. Anal. 50 (2009) 563–569.
[3] R. Kaliszan, T. Bączek, A. Cimochowska, P. Juszczyk, K. Wisniewska, Z. Grzonka, Proteomics 5 (2005) 409–415.
[4] T. Bączek, Curr. Pharm. Anal. 1 (2005) 31–40.
[5] R. Kaliszan, J. Chromatogr. B 715 (1998) 229–244.
[6] R. Kaliszan, Chem. Rev. 107 (2007) 3212–3246.
[7] A.A. D'Archivio, M.A. Maggi, P. Mazzeo, F. Ruggieri, Anal. Chim. Acta 628 (2008) 162–172.
[8] K. Heberger, J. Chromatogr. A 1158 (2007) 273–305.
[9] R. Kaliszan, Structure and Retention in Chromatography, A Chemometric Approach, Harwood Academic, Amsterdam, 1997.
[10] L.R. Snyder, J.J. Kirkland, J.L. Glajch, Practical HPLC Method Development, John Wiley & Sons, New York, 1997.
[11] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, 2000.
[12] O. Ivanciuc, T. Ivanciuc, D. Cabrol-Bass, A.T. Balaban, J. Chem. Inf. Comput. Sci. 40 (2000) 723–743.
[13] B.S. Junkes, R.D.M.C. Amboni, R.A. Yunes, V.E.F. Heinzen, Anal. Chim. Acta 477 (2003) 29–39.
[14] T. Hancock, R. Put, D. Coomans, Y. Vander Heyden, Y. Everingham, Chem. Intel. Lab. Sys. 76 (2005) 185–196.
[15] R. Put, C. Perrin, F. Questier, D. Coomans, D.L. Massart, Y. Vander Heyden, J. Chromatogr. A 988 (2003) 261–276.
[16] R. Put, Q.S. Xu, D.L. Massart, Y. Vander Heyden, J. Chromatogr. A 1055 (2004) 11–19.
[17] T. Bączek, A. Buciński, A.R. Ivanov, R. Kaliszan, Anal. Chem. 76 (2004) 1726–1732.
[18] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, Part A, Elsevier, Amsterdam, 1997.
[19] M. Daszykowski, W. Wu, A.W. Nicholls, R.J. Ball, T. Czekaj, B. Walczak, J. Chemom. 21 (2007) 292–302.
[20] C. Ruckebusch, F. Orhan, A. Durand, T. Boubellouta, J.P. Huvenne, Appl. Spectrosc. 60 (2006) 539–544.
[21] A. Durand, O. Devos, C. Ruckebusch, J.P. Huvenne, Anal. Chim. Acta 595 (2007) 72–79.
[22] R. Leardi, A.L. González, Chemom. Intell. Lab. Sys. 41 (1998) 195–207.
[23] D. Jouan-Rimbaud, D.L. Massart, O.E. de Noord, Chemom. Intell. Lab. Sys. 35 (1996) 213–220.
[24] www.talete.mi.it/dragon_exp.htm (accessed on 11.03.2009).
[25] R. Kaliszan, TRAC 18 (1999) 400–410.
[26] T. Bączek, P. Wiczling, M.P. Marszałł, Y. Vander Heyden, R. Kaliszan, J. Proteome Res. 4 (2005) 555–563.
[27] T. Bączek, R. Kaliszan, K. Novotná, P. Jandera, J. Chromatogr. A 1075 (2005) 109–115.
[28] T. Bączek, J. Sep. Sci. 29 (2006) 547–554.