Journal of Pharmaceutical and Biomedical Analysis 50 (2009) 563–569

Contents lists available at ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis

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

The molecular descriptor log Sum_{AA} and its alternatives in QSRR models to predict the retention of peptides

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

[a] Department of Analytical Chemistry and Pharmaceutical Technology, Vrije Universiteit Brussel-VUB, Laarbeeklaan 103, 1090 Brussels, Belgium
[b] Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk, Gen. J. Hallera 107, 80-416 Gdańsk, Poland
[c] Department of Medicinal Chemistry, Faculty of Pharmacy, Collegium Mediucum, Nicolaus Copernicus University, Bydgoszcz, Poland

ARTICLE INFO

Article history:
Received 26 June 2008
Received in revised form 2 September 2008
Accepted 3 September 2008
Available online 9 September 2008

Keywords:
HPLC retention
Peptides
Molecular descriptors
QSRR
log Sum_{AA}
Proteomics

ABSTRACT

The use of the experimental molecular descriptor log Sum_{AA} and some possible alternatives were evaluated in the QSRR analysis of peptides. To quantitatively characterize the structure of analytes in a previously proposed QSRR the following three structural descriptors were applied: 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*. Taking into consideration that most amino acids were hardly retained in the different RP-HPLC systems on which the peptides retention was measured, the contribution of most amino acids to the log Sum_{AA} descriptor is rather constant. Therefore, to enlarge the variability of the descriptor and the amino acids contributions for a given series of peptides, in a first instance, it was evaluated whether, by changing the chromatographic conditions, the retention differences between the amino acids could be increased, while maintaining their mutual selectivity. It was not evident to find such conditions. Secondly, it was also investigated whether the experimental descriptor log Sum_{AA} can be replaced by a theoretical, either based on a simple or on a weighted counting of the amino acids composing the peptide. The weighting factor for the retained amino acids was determined by their experimental gradient retention times measured on different systems. The predictive abilities of the new QSRR models (applying the alternative descriptors) were assessed using the leave-one-out cross-validation procedure and compared to that of the initial model. Finally, a descriptor was defined for which the retention measurement of only a limited number of amino acids is required. It resulted in QSRR models with similar predictive properties as those with log Sum_{AA}, but with a reduced workload.

1. Introduction

Over the last 30 years numerous reports have been published describing that the retention behaviour of peptides in reversed-phase high-performance liquid chromatography (RP-HPLC) provides useful information to predict the retention of related peptides, and consequently simplifies the identification of specific peptides present in mixtures [1–20].

Nowadays much interest in proteomic analysis goes to the practically useful processing of the information gained after the high-performance liquid chromatography (HPLC) separation of peptides (often with mass spectrometric detection).

A popular opinion among scientist is that properties of compounds are encoded in their structure and that a required property (e.g. pharmacological activity, boiling point) can be *a priori* estimated and designed. Moreover, it is well known that to predict a given physicochemical property, the relationship between the chemical structure and the desired property must be quantified. Such relationship between a chromatographic parameter (determined by physicochemical properties of both, mobile phase, stationary phase and eluted substance) and molecular descriptors, characterizing the molecular structure of the analytes, is known under the acronym QSRR: Quantitative Structure-Retention Relationship [21–24].

The use of QSRR models to predict the retention of peptides might be a valuable tool which could help identifying them during the proteomic research [25,26]. The building of a QSRR model needs a set of quantitatively comparable retention parameters for a large enough series of representative analytes and a set of their (theoretical) molecular descriptors [27–29].

* Corresponding author. Tel.: +48 58 349 32 60; fax: +48 58 349 32 62.
** Corresponding author. Tel.: +32 2 4774734; fax: +32 2 4774735.
E-mail addresses: roman.kaliszan@amg.gda.pl (R. Kaliszan), yvanvdh@vub.ac.be (Y. Vander Heyden).

0731-7085/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpba.2008.09.004

A theoretical descriptor can be defined as the final result of a mathematical procedure. It originates from translating the symbolic representation of the molecule into a useful numeric value, while an experimental descriptor is the result of a standardized experiment [30]. The number of structural descriptors which can be ascribed to an individual analyte is practically unlimited. For example, the Dragon software [31] which is frequently used for this purpose, calculates 3224 molecular descriptors. Thus, for proper QSRR building, descriptor selection is required.

With the aid of modeling techniques, the retention parameters are then modeled as a function of analyte descriptors. The most frequently used modeling approaches in the field of QSRR are multiple linear regression (MLR) and partial least square (PLS) regression.

Besides MLR and PLS, also advanced modeling techniques are applied [13,14,32,33], like classification and regression trees (CART), stochastic gradient boosting for tree-based models (Treeboost), random forests (RF), uninformative variable elimination partial least squares (UVE-PLS), genetic algorithms on multiple linear regression (GA-MLR) and multivariate adaptive regression splines (MARS), which either have a built-in variable-selection or -reduction feature.

The above techniques of descriptor selection for QSRR model building provide different sets of selected descriptors. One reason for this is the fact that the different techniques act differently, i.e. some are more local while others are more global [33]. A global model is based or gives information on the global domain of the data set, while a local one only on a part of this domain. For instance, the lower in a classification tree (CART) the more local the model becomes.

The above approaches, where QSRR models are build, starting from a large matrix of theoretical descriptors, are not much liked by a part of the QSRR community. A first reason is that, depending on the modeling techniques different descriptors are selected, as was already indicated. A second is the fact that those theoretical descriptors cannot always be linked easily to physico-chemical properties of the molecule. This makes that the models built, though having very good predictive properties, often are considered as a kind of "black boxes".

An alternative is that QSRR models are build, based on a limited number of very well understood descriptors, linked to known physico-chemical properties. Lately, in the latter context a QSRR model has been proposed by Kaliszan et al. [15,25] to predict the gradient retention times of peptides under given HPLC separation 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, $\log \mathrm{Sum_{AA}}$; the logarithm of the peptide's van der Waals volume, $\log \mathrm{VDW_{Vol}}$; and the logarithm of its theoretically calculated *n*-octanol–water partition coefficient, *c*log *P*. This QSRR equation has the following form:

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

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

Previous studies [15–17] demonstrated good predictive properties of the gradient retention times of peptides by means of the above QSRR. In these studies it was also observed that most amino acids were hardly retained at the applied HPLC conditions, i.e. they elute close to the dead time or even sooner. This can either be explained by the fact that exclusion phenomena may appear or that a too high elution strength of the mobile phase is used. It means that $\log \mathrm{Sum_{AA}}$ is determined by a large number of non-retained amino acids. In a situation were none of the amino acids would be retained the $\log \mathrm{Sum_{AA}}$ descriptor becomes in fact directly equivalent to the number of amino acids occurring in the peptide and experimental measurements become redundant. It then can be replaced by a theoretical descriptor, which is directly related to the number of amino acids in the peptide.

$\log \mathrm{Sum_{AA}}$ would become more meaningful as experimental descriptor when all amino acids would be retained on the chromatographic system and retention differences between all or most amino acids would occur. A first aim of this study was to try increasing the retention of the amino acids and their retention differences. The goal was to find experimental conditions where the retention of the amino acids would be increased and diversified but in such a way that no selectivity differences relative to the initial gradient conditions would be seen, i.e. the elution sequence of the amino acids remains the same.

The experiments were performed on chemically bonded silica RP stationary phases with different functional groups, phases packed with cross-linked polystyrene, and on highly porous monolithic silica rods. Seven columns were used, but 18 chromatographic systems were created by varying gradient profiles and column temperatures. A total of 98 peptides and 20 amino acids were analyzed. The peptides were selected to cover a wide range of structural diversity and were composed by 2 up to 24 amino acids. The data set concerned, is described in Ref. [15].

A second goal of this study concerned the replacement of the experimental $\log \mathrm{Sum_{AA}}$ with a calculated descriptor. The use of a theoretical descriptor instead of an experimental would reduce costs and time, and improve the availability of the descriptor. In a first instance, the theoretical descriptor was based on a simple counting of the amino acids in a peptide. In a second instance, variations to that descriptor were introduced and examined.

2. Materials and methods

2.1. Chemicals

Acetonitrile (ACN, HPLC grade) from Merck (Darmstadt, Germany) and trifluoroacetic acid (TFA) from Fluka (Buchs, Switzerland) were used. Water used during analyses 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): AA, AG, AF, YL, DD, ML, WW, GM, GL, WF and GHG. All other were synthesized at the Department of Organic Chemistry, University of Gdańsk, Poland. The peptides studied were selected to assure a wide range of structural diversity, including posttranslational modifications of peptides (e.g. acetylation and amidation)

2.2. Equipment and conditions for peptides retention measurements [15]

Chromatographic measurements were made on different instruments. The first HPLC apparatus was from Waters (Milford, MA, USA) and was equipped with a pump, variable wavelength UV/vis detector, autosampler, column oven and the Waters Millennium 2.15 software for data collection and instrument control. Measurements on that equipment were performed with an XTerra MS C18 column, 15.0–0.46 cm I.D., particle size 5 μm (Waters), packed with octadecyl-bonded silica.

Table 1
Peptides considered in the study

No.	Amino acid sequence	No.	Amino acid sequence	No.	Amino acid sequence
1	AA	34	SKPKTNMKHMAGAAAAG-CONH2	67	LVFF-CONH2
2	AG	35	Ac-HNPGYPHNPGYP-CONH2	68	GSNKGAIIGLM-CONH2
3	AF	36	Ac-HNPGYPHNPGYPHNPGYPHNPGYP-CONH2	69	GKTKEGVLY-CONH2
4	YL	37	HSDGIFTDS	70	KTKEGVLY-CONH2
5	DD	38	HSEGTFTSD	71	TKEGVLY-CONH2
6	ML	39	YKIEAVQSETVEPPPPAQ-CONH2	72	KEGVLY-CONH2
7	WW	40	TLSYPLVSVVSESLTPER-CONH2	73	EGVLY-CONH2
8	GM	41	PYPLRDVRGEPLEPPEPS-CONH2	74	GVLY-CONH2
9	GH	42	EVHHQKLVFFAEDVGSNK-CONH2	75	MAGASELGTGPGA-CONH2
10	GL	43	EVHHQKLVFFAKDVGSNK-CONH2	76	AGGYKPFNLETA-CONH2
11	WF	44	EVHHQKLVFFAQDVGSNK-CONH2	77	GAPGGPAFPGQTQDPLYG-CONH2
12	GHG	45	EVHHQKLVFFAGDVGSNK-CONH2	78	Ac-ETHLHWHTVAK-CONH2
13	LPQIENVKGTEDSGTT-CONH2	46	EVHHQKLVFFAENVGSNK-CONH2	79	Ac-ETHLHWHTVAKET-CONH2
14	VKGTEDSGTT-CONH2	47	EVHHQKLVFFGEDVGSNK-CONH2	80	HT
15	EHADLLAVVAASQKK-CONH2	48	pEADPNKFYGLM-CONH2	81	WHT
16	VVAASQKK-CONH2	49	DAEFRH-CONH2	82	HWHT
17	LAQAVRSS-CONH2	50	Ac-DAEFRH-CONH2	83	LHWHT
18	SFSMIKEGDYN-CONH2	51	DAEFGH-CONH2	84	HLHWHT
19	Ac-NH-CEQDGDPE-CONH2	52	Ac-DAEFGH-CONH2	85	THLHWHT
20	YKIEAVKSEPVEPPLPSQ-CONH2	53	DAEFRHDSG-CONH2	86	ETHLHWHT
21	LPPGPAVVDLTEKLEGQGG-CONH2	54	DAEFGHDSG-CONH2	87	SETHLHWHT
22	VVDLTEKLEGQGG-CONH2	55	DAEFRHDSGY-CONH2	88	Ac-EVHHQK
23	DRVYIHPF	56	Ac-DAEFRHDSGY-CONH2	89	EVHHQK
24	ETS	57	DAEFGHDSGF-CONH2	90	EVRHQKLVFF
25	KETS	58	Ac-DAEFGHDSGF-CONH2	91	EVRHQK
26	AKETS	59	EVHHQK-CONH2	92	Ac-EVRHQK
27	VAKETS	60	Ac-EVHHQK-CONH2	93	Ac-EVHHQKLVFF
28	TVAKETS	61	EVRHQK-CONH2	94	EVHHQKLVFF
29	HTVAKETS	62	Ac-EVRHQK-CONH2	95	Ac-EVRHQKLVFF
30	WHTVAKETS	63	EVHHQKLVFF-CONH2	96	Ac-DAEFRH
31	HWHTVAKETS	64	Ac-EVHHQKLVFF-CONH2	97	DAEFGH
32	LHWHTVAKETS	65	EVRHQKLVFF-CONH2	98	Ac-DAEFGH
33	MAGAAAAG-CONH2	66	Ac-EVRHQKLVFF-CONH2		

Amino acid abbreviations: see text. Ac: acetylation; -CONH2: amidation.

Measurements on PLRP-S, 15.0–0.46 cm I.D. (Polymer Laboratories, Amherst, MA), made of cross-linked polystyrene(divinylbenzene), on Discovery HS F5-3, 15–0.46 cm I.D. (Supelco, Bellefonte, PA), packed with a pentafluorophenylpropyl-terminated reversed phase and on Chromolith columns, 10.0–0.46 cm I.D. (Merck), made of a highly porous monolithic rod of silica, were performed on an HPLC apparatus, comprising a detector HP series 1050, an autosampler HP series 1050, a pump Agilent 1100 series and a heater/chiller model 7956 (Jones Chromatography, Glamorgan, UK). The instrument ran on HP Chem Station for LC software. Measurements on LiChrospher RP-18, 25.0–0.46 cm I.D. (Merck), packed with octadecyl-bonded silica, LiChrospher CN, 10.0–0.46 cm I.D. (Merck), packed with silica terminated with cyanopropyl ligands and Discovery RP-Amide C16 columns, 15–0.46 cm I.D. (Supelco), packed with silica terminated with amide groups, were done on a Merck-Hitachi LaChrom HPLC system (Merck-Hitachi, Frankfurt-Tokyo, Germany-Japan), equipped with a UV/vis detector (L-7400), autosampler (L-7200), column oven (L-7360) and the software D-7000 HPLC System Manager, version 4.1.

The injected sample volume was 20 μl. The eluent flow rate was 1 ml/min and detection wavelengths 214 and 223 nm. Gradient elution was carried out with solvent A (water with 0.12% 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. The dead time was determined by injection of solvent B. All samples were dissolved in water containing 0.10% (v/v) trifluoroacetic acid.

On the XTerra MS C18 column the gradient was from 0% B to 60% B within 20 min (t_G) at a temperature, T, of 40 °C. For all other columns it was from 4% B to 60% B. On the LiChrospher RP-18 column experiments were carried out with t_G equal to 20 min and at temperatures of 40, 60 and 80 °C, as well as with t_G of 60 and 120 min and T of 40 °C (Table 2). On the PLRP-S column t_G of 20 and 60 min, and T of 40, 60 and 80 °C were examined. For the Discovery RP-Amide C16 column experiments were with t_G = 20 min and T = 40, 60 and 80 °C. For the LiChrospher CN, Discovery HS F5-3 and Chromolith columns the experiments were with t_G = 20 min and at T = 40 °C.

2.3. Determination of chromatographic retention parameters

The retention times, t_R, of the peptides and amino acids were measured in Ref. [15] on the eighteen chromatographic systems specified higher. In Table 2 the retention times of the amino acids on these systems are given. These t_R were used to derive the log Sum_{AA} of the peptides.

2.4. Conditions for the alternative amino acids retention measurements

All experiments for amino acids were performed with solvents A and B, described above. At first, instead of in gradient mode the amino acids were analyzed under isocratic conditions with 3, 4 or 5% (v/v) of solvent B. Two temperatures, 25 and 40 °C, and two flow rates, 0.5 and 1 ml/min, were tested.

In the next step, the addition of 1 mM SDS (sodium dodecyl sulfate), as ion-pairing reagent in the mobile phase, was tested. The experiments in ion-pair HPLC mode were performed under isocratic conditions with different fractions of solvent B, i.e. 4, 15, 20, 22 and 25% (v/v).

Table 2
Retention times (min) of the amino acids in the examined HPLC systems [15]

Column	Gradient time, t_G (min)	Temp., T (°C)	Amino acids																			
			A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V
XTerra MS C18	20	40	2.10	2.47	1.92	1.97	2.12	2.13	2.00	1.87	2.02	8.98	9.40	2.02	4.97	11.6	2.60	1.85	1.90	12.02	8.63	4.17
LiChrospher RP-18	20	40	2.51	2.69	2.26	2.32	2.63	2.53	3.35	2.29	2.50	7.70	8.11	2.54	5.67	9.42	1.87	2.28	2.44	11.17	6.92	4.68
	20	60	2.43	2.54	2.20	2.27	2.53	2.39	3.02	2.28	2.38	7.08	7.46	3.42	4.67	8.67	1.87	2.22	2.40	10.05	5.77	4.17
	20	80	2.40	2.42	2.17	2.20	2.47	2.29	2.83	2.23	2.30	6.38	6.73	2.34	4.05	7.84	1.85	2.18	2.34	9.12	4.39	3.77
	60	40	2.51	2.70	2.27	2.34	2.64	2.51	3.34	2.29	2.52	9.02	9.73	2.54	5.70	12.77	1.85	2.27	2.46	17.36	7.72	4.64
	120	40	2.47	2.75	2.26	2.31	2.33	2.46	3.23	2.29	2.48	8.20	9.01	2.49	5.07	13.37	1.88	2.27	2.93	20.96	7.21	4.26
PLRP-S	20	40	3.19	3.22	3.14	3.16	3.24	3.20	3.16	3.15	3.17	4.41	4.58	3.18	4.03	6.21	3.27	3.14	3.18	8.08	4.60	3.57
	20	60	3.17	3.19	3.13	3.14	3.17	3.18	3.53	3.13	3.15	4.21	4.35	3.16	3.83	5.79	3.25	3.13	3.16	7.43	4.18	3.53
	20	80	3.13	3.14	3.09	3.10	3.10	3.12	3.44	3.09	3.10	4.01	4.10	3.11	3.62	5.27	3.19	3.09	3.12	6.13	3.84	3.44
	60	40	3.18	3.22	3.14	3.16	3.24	3.20	3.16	3.14	3.17	4.38	4.56	3.17	4.00	6.68	3.26	3.13	3.17	10.69	4.55	3.56
	60	60	3.19	3.21	3.14	3.16	3.23	3.20	3.56	3.15	3.17	4.23	4.37	3.18	3.88	6.12	3.27	3.14	3.18	9.44	4.22	3.55
	60	80	3.13	3.13	3.08	3.10	3.12	3.13	3.43	3.09	3.10	4.01	4.09	3.11	3.62	5.32	3.19	3.09	3.12	7.63	3.85	3.45
Discovery RP-Amide C16	20	40	1.98	2.00	2.06	2.00	2.07	1.99	2.19	1.99	2.19	2.02	2.19	1.98	2.14	2.27	2.21	1.97	1.95	2.52	2.16	2.16
	20	60	1.95	1.96	1.98	1.96	2.05	1.96	2.14	1.93	1.99	2.18	2.17	1.95	2.09	2.20	2.18	1.95	1.95	2.39	2.11	2.18
	20	80	1.95	1.96	1.96	1.95	1.99	1.96	2.14	1.94	1.95	2.27	2.32	1.92	2.10	2.32	2.21	1.97	1.95	2.69	2.14	2.24
LiChrospher CN	20	40	1.19	1.55	1.51	1.53	1.44	1.51	1.48	1.53	1.54	1.81	1.81	1.52	1.70	2.26	1.53	1.19	1.19	3.63	1.92	1.60
Discovery HS F5-3	20	40	2.39	2.98	2.11	2.21	2.22	2.42	2.92	2.16	2.59	6.13	6.56	2.78	4.78	8.03	2.84	2.13	2.32	10.07	5.92	4.20
Chromolith	20	40	1.61	1.67	1.54	1.57	1.55	1.61	1.89	1.56	1.60	3.47	3.72	1.62	2.54	4.78	1.80	1.55	1.60	6.26	3.28	2.20

Amino acids abbreviations: see text.

Fig. 1. Retention times (min) of the 20 amino acids measured on the 18 RP-HPLC systems (●). The median (■) for each amino acid also is indicated.

2.5. Structural descriptors of the peptides

The theoretical descriptors employed in the QSRR, i.e. $\log VDW_{Vol}$, *clog P* were calculated by the molecular modeling program HyperChem for personal computers with the extension ChemPlus (HyperCube, Gainesville, FL, USA). The software 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. The experimental descriptor, $\log Sum_{AA}$, was calculated as defined higher.

In the second part of this paper $\log Sum_{AA}$ was replaced by: (1) the logarithm of the number of amino acids composing the individual peptide, $\log No_{AA}$; (2) the logarithm of the sum of the median retention times of the amino acids composing the peptide, $\log SumM_{AA}$. The median was derived from the measurements on all chromatographic systems and was rounded to the closest half of an integer (Fig. 1); (3) the logarithm of the summed $k+1$ values of the amino acids, $\log Sum(k+1)_{AA}$ with k the apparent retention factor. 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 was $k=0$, after rounding. For the 7 retained amino acids, their individual k on a given system was used. Not k but $k+1$ was used to avoid zero values in the calculation of the logarithm.

2.6. Data analysis

The QSRR models were derived by means of multiple linear regression (MLR) using Matlab 7.0.1 software (The Mathworks, Natick, MA). Prior to the QSRR model building, the descriptor's values were autoscaled in order to remove undesired scale differences, and also because by autoscaling the direct comparison of the coefficients from the model is possible.

To quantify the predictive power of the constructed QSRR models the cross-validation root-mean-squared errors (RMSECV) were calculated with Matlab 7.0.1. The leave-one-out (LOO) method was applied. The RMSECV value determined with the LOO procedure originates from taking out one case (peptide) from the entire data set as the hold-out case. Then a model is built on the remaining cases. The resulting model is used to predict the hold-out case. This entire process is repeated until each case once became the hold-out case. When different models are compared, the one with the smallest RMSECV is considered the best predictive model.

In this study only the RMSECV values of different models were calculated and no RMSE values from a test set, because the main

interest was to see whether or not the new QSRR models possess equally good prediction abilities as the initial. We were not directly interested in the absolute value of the errors, a situation where RMSE from a test set will show more realistic values than the usually over-optimistic RMSECV.

3. Results and discussion

3.1. Attempts to increase amino acids retention (variability)

The first goal of this study was to try improving retention differences between the (gradient) retention times of the amino acids. Most amino acids have almost no retention in the initial conditions (Table 2). This can be due to a too high elution strength of the mobile phase, because most amino acids are small, charged, polar compounds that have a much higher affinity for the mobile than for the stationary phase in RP-HPLC. Elution before the dead time marker is either a consequence of the experimental variability on the measured times or of exclusion phenomena. The latter occurs when the small molecule used to mark the dead time penetrates the stationary phase pores more thoroughly than do the usually larger analyte molecules [34]. However, from a practical point of view the exclusion mode possibilities in analytical RP-HPLC are very limited because of a too small elution window, i.e. the elution time interval between complete pore penetration and complete exclusion. Therefore, our main interest goes to the retention of compounds.

Other chromatographic conditions than for the peptides were tested to improve the differences in the retention of the amino acids and to make log Sum_{AA} more meaningful and variable for different peptides. The most important requirement when changing the chromatographic conditions for the amino acids analyses was that their retention variability increases but their retention sequence remains similar as in the initial systems (i.e. as in Table 2) in order to keep the correlation with the initial data set of [15].

First, the amino acids were analyzed isocratically. Their retention then is not affected by an increasing organic modifier concentration. The amino acids were analyzed with 3, 4 and 5% (v/v) of solvent B in the mobile phase. Unfortunately, these isocratic conditions did not provide much better discrimination between the retention of the amino acids. A reduced temperature of 25 °C and a reduced flow rate of 0.5 ml/min at 5% of solvent B neither increased the amino acids retention variability.

As it is known that poorly retained, ionizable solutes can be retained by the addition of submicellar quantities of ionic surfactants acting as ion-pairing agents [35], the amino acids were analyzed with the addition of 1 mM sodium dodecyl sulfate (SDS) as ion-pairing reagent. Several isocratic elutions in ion-pair mode were considered with different concentrations of solvent B, i.e. 4, 15, 20, 22 and 25% (v/v). Higher concentrations of solvent B were needed to elute the amino acids, because of the interaction between the ionized amino acids and the SDS ions. The sequence of the amino acids observed on a system containing 1 mM SDS was different from that on a system without (Fig. 2). The shifts in retention were seen for three amino acids—arginine, histidine and lysine, i.e. the basic amino acids. These amino acids have two interaction sides with SDS. This could explain their relative longer retention. Because the elution sequence of the amino acids changes, the ion-pair chromatography conditions cannot be used as alternative to estimate log Sum_{AA}.

Thus, the attempts to find alternative conditions resulting in more retention variability of the amino acids were disappointing. Therefore, the next part of this study discusses the possibility to replace the experimental descriptor log Sum_{AA} with a theoretical.

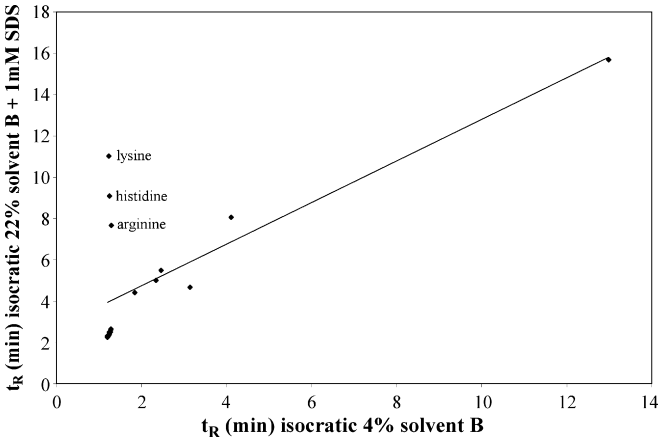

Fig. 2. The retention times of the amino acids analyzed on XTerra RP-18 in isocratic conditions with 4% of solvent B versus those under isocratic conditions in ion-pairing mode with 22% of solvent B and 1 mM SDS; flow rate 1 ml/min.

3.2. The alternative descriptors for log Sum_{AA}

An experimental descriptor, such as log Sum_{AA}, is the result of a standardized experiment, which in this case means that the retention of 20 amino acids was measured at given conditions. If the experimental descriptor can be reflected in a theoretical, which allows describing the retention behaviour of the peptides equally good and which results in a QSRR model with a good predictive ability, then the measurements of the retention of the amino acids become redundant.

From its definition it is seen that log Sum_{AA} contains information about the size of the peptide and about the side groups of the amino acids composing the individual peptide. The groups are responsible for the retention of both amino acids and peptides. The two other descriptors involved in the QSRR model, log VDW_{Vol} and *c*log *P*, also carry information about size of the peptide and side groups. The first, log VDW_{Vol}, is the volume within the van der Waals molecular surface [30] which gives also information about the size of the molecule. The second, *c*log *P*, is the octanol–water partition coefficient which describe the hydrophobicity of the molecule, and is influenced by the presence of polar or apolar side groups.

Therefore, because all three descriptors relate to the size and the polarity of the peptide it was at first investigated whether the log Sum_{AA} descriptor provides significant information in the QSRR model. This was done by removing the descriptor from the QSRR equation and assessing the predictive property of the resulting equation from the LOO-RMSECV procedure. Evaluation of the significance of the log Sum_{AA} coefficient [36], as an alternative was not considered, because as noticed higher the RMSECV gives us values which immediately can be compared for different models. The RMSECV values calculated for the QSRR equations with just two descriptors: log VDW_{Vol} and *c*log *P*, were in the range 1.50–2.04 for t_G of 20 min, between 4.01–4.70 for t_G = 60 min, and 8.94 for t_G = 120 min (column (2) in Table 3). For the QSRR equation containing also log Sum_{AA} the RMSECV values were remarkably lower, i.e. in the range 0.93–1.82 for t_G = 20 min, between 2.91–3.36 for t_G = 60 min, and 7.06 for t_G = 120 min (column (1) in Table 3). This indicates that log Sum_{AA} provides important information in predicting the retention times of peptides.

When, most amino acids are hardly retained the log Sum_{AA} descriptor tends to become directly related to the number of amino acids composing the peptide and measurement of the amino acids retention would become unnecessary. Therefore, it was evaluated if log Sum_{AA} can be replaced by a theoretical descriptor, log No_{AA},

Table 3
Predictive abilities of the QSRR models expressed by the root-mean-squared cross-validation error calculated from the leave-one-out procedure

Column	Gradient time, t_G (min)	Temp., T (°C)	RMSECV				
			log Sum_{AA} (1)	without log Sum_{AA} (2)	log No_{AA} (3)	log $SumM_{AA}$ (4)	log $Sum(k+1)_{AA}$ (5)
XTerra MS C18	20	40	1.01	1.92	1.94	1.42	1.00
LiChrospher RP-18	20	40	0.98	1.89	1.87	1.30	0.94
	20	60	1.09	1.91	1.85	1.27	0.95
	20	80	1.10	1.94	1.82	1.31	1.02
	60	40	2.91	4.70	4.44	3.31	2.92
	120	40	7.06	8.94	7.93	7.00	6.97
PLRP-S	20	40	1.28	1.64	1.52	1.12	1.27
	20	60	1.31	1.62	1.46	1.13	1.30
	20	80	1.34	1.59	1.38	1.12	1.33
	60	40	3.25	4.17	3.57	3.13	3.23
	60	60	3.29	4.09	3.38	3.15	3.26
	60	80	3.36	4.01	3.12	3.24	3.34
Discovery RP-Amide C16	20	40	1.70	1.79	1.34	1.66	1.71
	20	60	1.68	1.78	1.36	1.66	1.69
	20	80	1.64	1.78	1.35	1.54	1.64
LiChrospher CN	20	40	1.82	2.04	1.74	1.81	1.84
Discovery HS F5-3	20	40	1.29	1.70	1.71	1.38	1.27
Chromolith	20	40	0.93	1.50	1.39	1.01	0.91

based on a simple counting of the amino acids composing the peptide. Counting means that to each amino acid a weight one was ascribed, but since some of the amino acids have a considerably higher retention this weight did not describe their retention properly. The RMSECV values (Table 3) confirm that this log No_{AA} is less appropriate for the QSRR prediction of peptides retention.

For that reason, in a next step it was evaluated whether the joint experimentally measured retentions on the different systems could not lead to the definition of a generalized descriptor, applicable on all systems and which would not require future amino acids measurements anymore. For this purpose the retention times of each amino acid analyzed on all systems were characterized by their medians, rounded to the closest half of an integer (Fig. 1). These rounded values were then used as weights in the calculation of a descriptor, log $SumM_{AA}$, calculated similarly to log No_{AA}. The new models showed RMSECV values between 1.01–1.08 for t_G = 20 min, between 3.13–3.31 for t_G = 60 min, and of 6.99 for t_G = 120 min (column (4) in Table 3). It was seen that models with log $SumM_{AA}$ perform better than those with log No_{AA}, but the predictive properties were still worse than for the models based on log Sum_{AA}. The reason is probably the fact that the median, representing the weight for a given amino acid, on some systems is still far from its measured retention time and thus not representative.

In a next step the dead time was taken into account and log $SumM(k+1)_{AA}$ was examined, which was based on the apparent retention factor, k. This is another measure of retention and is calculated as the retention factor in isocratic conditions:

$$k = \frac{t_R - t_0}{t_0} \quad (2)$$

where t_R is the retention time and t_0 the time needed to detect an unretained compound. The retention factor was expected to be more similar for different related systems than t_R, because it compensates for some physical differences between columns. The predictive abilities of the QSRR models with log $SumM(k+1)_{AA}$ were calculated, but the RMSECV values were not better than for the models with the log $SumM_{AA}$ descriptor (results not shown in Table 3).

The above attempts to replace the experimental descriptor, log Sum_{AA}, with a theoretical one or with at least one not requiring future measurements for the amino acids, were not successful. Therefore in the next approach, a descriptor requiring only the measurement of a limited number of amino acids was considered. As it was already mentioned, during measurements of the amino acids retention it was noticed that some (I, L, M, F, W, Y, V) have retention which the others hardly have. Related to this fact the next descriptor was proposed. This descriptor, log $Sum(k+1)_{AA}$, was also based on the apparent retention factor, k, but only for the 7 retained amino acids the experimental retention values on the individual systems are used. For the 13 hardly retained amino acids (A, R, N, D, C, E, Q, G, H, K, P, S, T) fixed values were ascribed (k = 0). The use of this log $Sum(k+1)_{AA}$ descriptor in the QSRR model resulted in RMSECV values between 0.91–1.84 for the t_G = 20 min, between 2.92–3.34 for t_G = 60 min, and 6.97 for t_G = 120 min (column (5) in Table 3). It indicates that the QSRR model containing this log $Sum(k+1)_{AA}$ descriptor has similar or in some cases even better predictive abilities than the QSRR model containing log Sum_{AA}.

The graphical interpretation of the predictive potencies of the newly derived QSRR models based on alternatives for log Sum_{AA} is presented in Fig. 3. This plot shows the values of RMSECV for some selected systems. The plots for the other systems have a

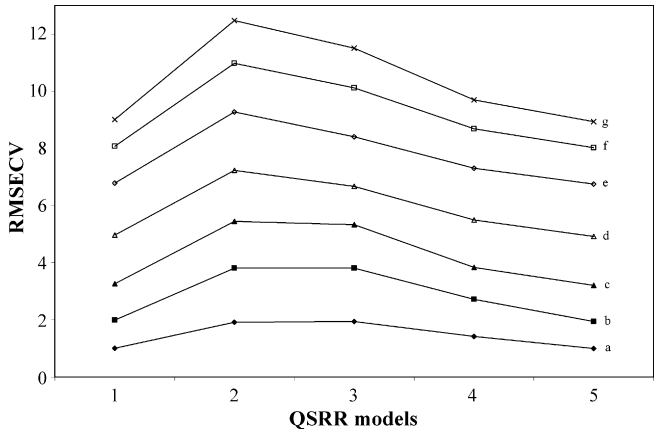

Fig. 3. Graphical interpretation of RMSECV values calculated for the five QSRR models considered in Table 3. QSRR model: (1) with log Sum_{AA}, (2) without log Sum_{AA}, (3) with log No_{AA}, (4) with log $SumM_{AA}$, (5) with log $Sum(k+1)_{AA}$; chromatographic systems: (a) XTerra MS C18, (b) LiChrospher RP-18, (c) PLRP-S, (d) Discovery RP-Amide C16, (e) LiChrospher CN, (f) Discovery HS F5-3, (g) Chromolith (all t_G = 20 min, T = 40 °C).

similar behaviour. The QSRR model containing the $\log \mathrm{Sum}(k+1)_{AA}$ descriptor has similar predictive abilities as the initial QSRR model using $\log \mathrm{Sum}_{AA}$.

Since the predictions of the retention of peptides using either $\log \mathrm{Sum}_{AA}$ or $\log \mathrm{Sum}(k+1)_{AA}$ are quite similar it seems logic to apply the latter because it requires the retention measurement for only seven (I, L, M, F, W, Y, V) instead of 20 amino acids. However, the retention of the amino acids on more gradient RP-HPLC systems for peptides still is to be considered to generalize the approach.

4. Conclusions

This study focused in a first instance on the improvement of retention differences between amino acids by changing the chromatographic conditions to determine $\log \mathrm{Sum}_{AA}$. This research was motivated by the fact that most amino acids almost have no retention in the examined conditions. An enlarged difference between the retention times of the amino acids seems recommended to improve the variability of the $\log \mathrm{Sum}_{AA}$ descriptor for different peptides. However, the attempts to find conditions for the analysis of amino acids resulting in more variability but maintaining the selectivity were disappointing.

Therefore, in the second part of our study the possibility to replace the experimental descriptor $\log \mathrm{Sum}_{AA}$ by a theoretical was evaluated. The latter was based on either a simple or on a weighted counting of the amino acids composing the individual peptide. Four molecular descriptors were derived and used as alternatives for $\log \mathrm{Sum}_{AA}$ in the QSRR analysis to predict peptides gradient retention. The most valuable alternative information about the gradient retention of peptides was provided by the $\log \mathrm{Sum}(k+1)_{AA}$ descriptor. The QSRR model containing this descriptor possesses a similar predictive property as the model with $\log \mathrm{Sum}_{AA}$. The new descriptor has the advantage that only for seven of the 20 amino acids the retention times are needed.

The QSRR model with the $\log \mathrm{Sum}(k+1)_{AA}$ descriptor should be applied to more RP-HPLC systems and to different peptides sets to confirm its usefulness in a more generalized sense.

Acknowledgements

The work was supported by the BWS Flemish – Polish bilateral project BWS (BIL) 05-03 and by the DWTC bilateral projects BL/03/C36 and BL/03/V09, between Belgium and China and Vietnam, respectively. This work was also partially carried out under Scientific Research Projects no. N N405 063634 from the Polish State Committee.

References

[1] J.L. Meek, Proc. Natl. Acad. Sci. U.S.A. 77 (1980) 1632–1636.
[2] C.A. Browne, H.P.J. Bennett, S. Solomon, Anal. Biochem. 124 (1982) 201–208.
[3] D. Guo, C.T. Mant, A.K. Taneja, J.M.R. Parker, R.S. Hodges, J. Chromatogr. 359 (1986) 499–518.
[4] D. Guo, C.T. Mant, A.K. Taneja, R.S. Hodges, J. Chromatogr. 359 (1986) 519–532.
[5] V. Casal, P.J. Martin-Alvarez, T. Herraiz, Anal. Chim. Acta 326 (1996) 77–84.
[6] C.T. Mant, N.E. Zhou, R.S. Hodges, J. Chromatogr. 476 (1989) 363–375.
[7] R.A. Houghten, S.T. DeGraw, J. Chromatogr. 386 (1987) 223–228.
[8] N.E. Zhou, C.T. Mant, R.S. Hodges, Pept. Res. 3 (1990) 8–20.
[9] S. Rothemund, E. Krause, M. Beyermann, M. Dathe, H. Engelhardt, M. Bienert, J. Chromatogr. A 689 (1995) 219–226.
[10] T. Wieprecht, S. Rothemund, M. Bienert, E. Krause, J. Chromatogr. A 912 (2001) 1–12.
[11] M. Palmblad, M. Ramström, K.E. Markides, P. Håkansson, J. Bergquist, Anal. Chem. 74 (2002) 5826–5830.
[12] M. Palmblad, M. Ramström, C.G. Bailey, S.L. McCutchen-Maloney, J. Bergquist, L.C. Zeller, J. Chromatogr. B 803 (2004) 131–135.
[13] R. Put, M. Daszykowski, T. Baczek, Y. Vander Heyden, J. Prot. Res. 5 (2006) 1618–1625.
[14] R. Put, Q.S. Xu, D.L. Massart, Y. Vander Heyden, J. Chromatogr. A 1055 (2004) 11–19.
[15] T. Bączek, P. Wiczling, M.P. Marszałł, Y. Vander Heyden, R. Kaliszan, J. Prot. Res. 4 (2005) 555–563.
[16] T. Bączek, J. Sep. Sci. 29 (2006) 547–554.
[17] T. Bączek, Curr. Pharm. Anal. 1 (2005) 31–40.
[18] O.V. Krokhin, R. Craig, V. Spicer, W. Ens, K.G. Standing, R.C. Beavis, J.A. Wilkins, Mol. Cell. Proteomics 3 (2004) 908–919.
[19] J.N. Adkins, S.P. Varnum, K.J. Auberry, R.J. Moore, N.H. Angell, R.D. Smith, D.L. Springer, J.G. Pounds, Mol. Cell. Proteomics 1 (2002) 947–955.
[20] K. Shinoda, M. Sugimoto, N. Yachie, N. Sugiyama, T. Masuda, M. Robert, T. Soga, M. Tomita, J. Prot. Res. 5 (2006) 3312–3317.
[21] R. Kaliszan, Quantitative Structure-Chromatographic Retention Relationships, Wiley, New York, 1987.
[22] R. Kaliszan, Structure and Retention in Chromatography. A Chemometric Approach, Harwood Academic Publishers, Amsterdam, 1997.
[23] L.R. Snyder, J.J. Kirkland, J.L. Glajch, Practical HPLC Method Development, John Wiley & Sons, New York, 1997.
[24] R. Kaliszan, T. Bączek, A. Buciński, B. Buszewski, M. Sztupecka, J. Sep. Sci. 26 (2003) 271–282.
[25] R. Kaliszan, T. Bączek, A. Cimochowska, P. Juszczyk, K. Wisniewska, Z. Grzonka, Proteomics 5 (2005) 409–415.
[26] R. Kaliszan, Chem. Rev. 107 (2007) 3212–3246.
[27] T. Bączek, R. Kaliszan, J. Chromatogr. A 987 (2003) 29–37.
[28] R. Kaliszan, M.A. Van Straten, M. Markuszewski, C.A. Cramers, H.A. Claessens, J. Chromatogr. A 855 (1999) 455–486.
[29] J. Jiskra, H.A. Claessens, C.A. Cramers, R. Kaliszan, J. Chromatogr. A 977 (2002) 193–206.
[30] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, 2000.
[31] www.talete.mi.it/dragon_exp.htm (accessed on 19.06.2008).
[32] T. Hancock, R. Put, D. Coomans, Y. Vander Heyden, Y. Everingham, Chemom. Intell. Lab. Syst. 76 (2005) 185–196.
[33] R. Put, Y. Vander Heyden, Proteomics 7 (2007) 1664–1677.
[34] J.H. Knox, R. Kaliszan, J. Chromatogr. 349 (1985) 211–234.
[35] A.H. Rodgers, J. Chromatogr. 636 (1993) 203–212.
[36] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. DeJong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, Part A, Elsevier, Amsterdam, 1997.