REGULAR ARTICLE

Prediction of high-performance liquid chromatography retention of peptides with the use of quantitative structure-retention relationships

Roman Kaliszan[1], Tomasz Bączek[1], Anna Cimochowska[1], Paulina Juszczyk[2], Kornelia Wiśniewska[2] and Zbigniew Grzonka[2]

[1] Medical University of Gdańsk, Department of Biopharmaceutics and Pharmacodynamics, Gdańsk, Poland
[2] University of Gdańsk, Department of Organic Chemistry, Gdańsk, Poland

Received: April 4, 2004
Revised: June 21, 2004
Accepted: June 23, 2004

Quantitative structure retention relationships (QSRR) were derived allowing prediction of reversed-phase high-performance liquid chromatography (HPLC) retention of peptides. To quantitatively characterize the structure of a peptide, and then to predict its gradient retention time under given HPLC conditions, the following descriptors are employed: logarithm of the sum of retention times of the amino acids composing the peptide, log Sum_{AA}, logarithm of Van der Waals volume of the peptide, log VDW_{Vol}, and logarithm of its calculated *n*-octanol-water partition coefficient, clog P. The first descriptor is based on a set of empirical data for 20 natural amino acids. The next two descriptors are easily calculated from a structural formula. The predicted gradient retention times are in excellent agreement with the experimental data, determined for a structurally diversified series of 101 peptides. The QSRR equation obtained predicts in a convenient and reliable manner the retention times for any peptide in a once characterized HPLC system.

Keywords:
Peptides / Prediction of retention / Quantitative structure-retention relationships

1 Introduction

Proteins are the main catalysts of biological functions. A comprehensive analysis and characterization of all the expressed proteins (proteome) with the use of modern analytical and bioinformatic tools [1, 2] includes identification of the proteins expressed, their quantification and determination of their contribution to one or more biological functions. One of the initial steps of proteomic analysis is peptide separation. However, little information from LC, usually employed for the separation, is actually utilized in proteomics [3, 4]. On the other hand, it is well known that chromatographic retention time (t_R) is a chemical structure dependent parameter, which should be constant for a given set of separation conditions (mobile phase composition, stationary phase, temperature, pH). In conjunction with MS/MS data, prediction of the t_R for a given peptide structure could help to improve the confidence of peptide identifications and to increase the number of correct identifications.

A number of reports have already been published describing the chromatographic behavior of peptides in RP-LC on the basis of amino acid composition [5–10]. In a paper by Meek [5] the derivation of specific values (retention coefficients) that represent the contribution to retention of each of the common amino acids and end groups was demonstrated. It was shown that retention coefficients could be derived directly from HPLC data for all amino acids and end groups such that the t_R of a peptide could be predicted from the sum of retention coefficients for each amino acid and

Correspondence: Dr. Roman Kaliszan, Medical University of Gdańsk, Department of Biopharmaceutics and Pharmacodynamics, Hallera 107, 80-416 Gdańsk, Poland
E-mail: roman.kaliszan@amg.gda.pl
Fax: +48-58-3493262

Abbreviations: QSRR, quantitative structure-retention relationships; **t_R**, retention time

end group. A similar strategy, but with different numbers of retention coefficients was proposed by Browne *et al.* [6], Casal *et al.* [7], and Guo *et al.* [8, 9]. In addition to the contribution of amino acids to the retention of peptides, [10] also considered Mant *et al.* the polypeptide chain length. Houghten and DeGraw [11] studied the influence of different amino acid sequences on peptide retention. Zhou *et al.* [12] observed that the presence of a preferred binding domain in an amphipathic α-helical peptide produced greater retention than might be predicted based on amino acid composition.

Recently, Palmblad *et al.* [3, 4] reported prediction of t_R for tryptic peptides for proteomic purposes. The applied algorithm was tested using tryptic digests of well characterized proteins and its accuracy was established on the basis of the differences between predicted and experimental retention for peptides that were identified by MS. The accuracy of predictions was promising in terms of distinguishing between true and false protein matches. Very recently an approach based on artificial neural networks (ANNs) was proposed for the prediction of peptide elution times by Petritis *et al.* [13]. The development of the initial ANN model was based on the assumption that peptide elution times should substantially depend on amino acid compositions. The predictive capability of ANN was tested by using large sets of confidently identified peptides and their t_R for the proteomes of two microorganisms. The model's predicted t_R was shown to increase the confidence of peptide identifications.

The approaches listed above are generally based on simple, additive, amino acid composition of peptide based relationships. Here, we propose a new quantitative structure-retention relationships (QSRR) approach to the prediction of gradient HPLC t_R of peptides. QSRR are statistically derived relationships between the chromatographic parameters and the quantities (descriptors) characterizing molecular structure of analytes [14]. Previous studies in our laboratory [15–17] demonstrated a good retention prediction performance of QSRR for small M_r analytes. The aim of the present work was to find a proper QSRR model allowing reliable, even if approximate, prediction of retention of peptides of a defined amino acid composition.

2 Materials and methods

2.1 Equipment

Chromatographic measurements were performed with an HPLC apparatus LC Module I plus (Waters, Milford, MA, USA) equipped with a pump, variable wavelength UV/VIS detector, autosampler and thermostat (Model Code LCH). Data were collected using Waters Millennium 2.15 software. A XTerra MS C18 column (15.0 × 0.46 cm id, particle size 5 μm; Waters, packed with octadecyl-bonded silica, was used in the study. Gradient HPLC elution was carried out with solvent A (water with 0.12% TFA) and solvent B (ACN with 0,10% TFA). The mobile phase used was filtered through a GF/F glass microfibre filter (Whatman, Maidstone, UK) and degassed with helium during the analysis. The gradient was formed from 0% to 60% B within 20 min. All the chromatographic measurements were done at 40°C with an eluent flow rate of 1 mL/min. The experiments were performed at a detection wavelength of 223 nm and dead time (2.30 min) was determined by a signal of solvent B. Peptide samples were dissolved in water with 0.10% of TFA. The injected sample volume was 20 μL.

2.2 Chemicals

ACN (HPLC grade) was from P.C. Odczynniki (Gliwice, Poland) and TFA was from Fluka (Buchs, Switzerland). Water was prepared with a Milli-Q Water Purification System (Millipore, Bedford, MA, USA). The amino acids listed in Table 1 were used to determine the peptide structure descriptor, Sum_{AA}, used in QSRR analysis. The peptides studied are listed in Tables 2 and 3. The peptides from Table 2 were used to derive the QSRR model. These peptides were randomly selected by the Kennard-Stone design method within MATLAB 6.5 software (The MathWorks, Natick, MA, USA) from the total set of 101 peptides available. The remaining 66 peptides given in Table 3 served to test the reliability of the QSRR model derived. The following peptides were purchased from Sigma (St. Louis, MO, USA): AA, AG, AF, TL, DD, ML, WW, GM, GH, GL, WF and GHG. Amino acids and angiotensin II (DRVYIHPF) were from Fluka. Other peptides used were synthesized at the Department of Organic Chemistry, University of Gdańsk according to a general procedure reported elsewhere [18, 19].

Table 1. t_R of natural amino acids used to derive the sum of gradient t_R of the amino acids comprising the individual peptide, Sum_{AA}.

No.	Amino acid	Amino acid letter code	$t_{R\,exp}$ (min)
1	Alanine	A	2.10
2	Arginine	R	2.47
3	Asparagine	N	1.92
4	Aspartic acid	D	1.97
5	Cysteine	C	2.12
6	Glutamic acid	E	2.13
7	Glutamine	Q	2.00
8	Glycine	G	1.87
9	Histidine	H	2.02
10	Isoleucine	I	8.98
11	Leucine	L	9.40
12	Lysine	K	2.02
13	Methionine	M	4.97
14	Phenylalanine	F	11.60
15	Proline	P	2.60
16	Serine	S	1.85
17	Threonine	T	1.90
18	Tryptophan	W	12.02
19	Tyrosine	Y	8.63
20	Valine	V	4.17

Table 2. Structural descriptors and t_R of a subset of 35 peptides used to derive QSRR

No.	Peptide sequence	log Sum_{AA}	log VDW_{Vol}	clog P	$t_{R\,exp}$ (min)
1	GHG	0.7604	2.3574	−2.63	2.72
2	LPQIENVKGTEDSGTT-NH2	1.6867	3.1736	−9.45	13.00
3	Ac-CEQDGDPE-NH2	1.2251	2.8836	−5.93	10.48
4	YKIEAVKSEPVEPPLPSQ-NH2	1.8060	3.2575	−1.94	13.95
5	LPPGPAVVDLTEKLEGQGG-NH2	1.8200	3.2262	−3.74	16.45
6	DRVYIHPF	1.6278	2.9741	1.97	15.15
7	SKPKTNMKHMAGAAAAG-NH2	1.6078	3.1931	−10.29	11.38
8	Ac-HNPGYPHNPGYPHNPGYP-NH2	1.7703	3.2501	−5.68	12.97
9	Ac-HNPGYPHNPGYPHNPGYPHNPGYP-NH2	1.9067	3.3717	−7.28	13.23
10	EVHHQKLVFFAEDVGSNK-NH2	1.8399	3.2699	−4.28	14.63
11	EVHHQKLVFFGEDVGSNK-NH2	1.8384	3.2662	−4.82	14.48
12	DAEFGHDSG-NH2	1.4374	2.8930	−5.27	10.93
13	LVFF-NH2	1.5655	2.7059	3.59	17.15
14	KTKEGVLY-NH2	1.5070	2.9363	−0.94	12.67
15	KEGVLY-NH2	1.4506	2.8140	0.07	12.78
16	EGVLY-NH2	1.4183	2.7220	0.51	13.22
17	MAGASELGTGPGA-NH2	1.5638	3.0030	−6.46	11.52
18	HT	0.5933	2.3468	−1.72	2.48
19	WHT	1.2025	2.5890	−0.47	11.62
20	HWHT	1.2543	2.7040	−1.29	11.63
21	SETHLHWHT	1.5473	2.9985	−3.26	12.92
22	EVRHQK	1.1706	2.8525	−3.36	8.82
23	Ac-DAEFGH	1.3363	2.7853	−1.85	12.25
24	AA	0.6232	2.1603	−0.74	2.75
25	AG	0.5988	2.1047	−1.28	2.38
26	AF	1.1367	2.3402	0.95	11.87
27	YL	1.2560	2.4394	1.86	12.65
28	GL	1.0519	2.2505	−0.08	10.95
29	WF	1.3733	2.5058	2.41	15.60
30	VAKETS	1.1514	2.7517	−2.45	8.70
31	HTVAKETS	1.2574	2.8846	−3.85	9.50
32	WHTVAKETS	1.4787	2.9702	−2.59	11.78
33	EVHHQK-NH2	1.1572	2.8413	−4.27	8.22
34	Ac-EVHHQKLVFF-NH2	1.7087	3.0841	0.51	15.92
35	EVRHQKLVFF	1.7125	3.0699	1.04	16.00

Table 3. Structural descriptors, experimental t_R, calculated t_R and their difference for the testing set of peptides not used to derive the QSRR equation

No.	Peptide sequence	log Sum_{AA}	log VDW_{Vol}	clog P	$t_{R\,exp}$ (min)	$t_{R\,pred}$ (min)	$\triangle t_R$ (min)
1	VKGTEDSGTT-NH2	1.3341	2.9326	−6.41	9.27	9.15	0.12
2	EHADLLAVVAASQKK-NH2	1.6950	3.1563	−3.89	15.15	13.92	1.23
3	VVAASQKK-NH2	1.3103	2.8874	−3.24	9.52	9.94	0.42
4	LAQAVRSS-NH2	1.4140	2.8750	−3.36	10.82	11.52	0.70
5	SFSMIKEGDYN-NH2	1.6793	3.0645	−4.20	13.90	14.14	0.24
6	VVDLTEKLEGQGG-NH2	1.6522	3.0789	−4.20	13.83	13.65	0.18
7	MAGAAAAG-NH2	1.2835	2.7642	−4.37	10.10	9.94	0.16
8	Ac-HNPGYPHNPGYP-NH2	1.6168	3.0812	−4.07	12.23	13.15	0.92
9	HSDGIFTDS	1.5316	2.9207	−3.56	13.40	12.95	0.45
10	HSEGTFTSD	1.4328	2.9162	−4.95	11.30	11.12	0.18
11	YKIEAVQSETVEPPPPAQ-NH2	1.7537	3.2457	−3.55	13.42	14.37	0.95
12	TLSYPLVSVVSESLTPER-NH2	1.8602	3.2251	−2.12	17.72	16.47	1.25
13	PYPLRDVRGEPLEPPEPS-NH2	1.8077	3.2583	−1.79	13.97	15.59	1.62

Table 3. Continued

No.	Peptide sequence	log Sum_{AA}	log VDW_{Vol}	clog P	$t_{R\ exp}$ (min)	$t_{R\ pred}$ (min)	$\triangle t_R$ (min)
14	EVHHQKLVFFAENVGSNK-NH2	1.8395	3.2707	−4.98	14.52	15.11	0.59
15	pEADPNKFYGLM-NH2	1.6921	3.0634	−2.03	15.78	14.93	0.85
16	DAEFRH-NH2	1.3481	2.8287	−2.57	10.62	11.03	0.41
17	Ac-DAEFRH-NH2	1.3481	2.8503	−2.41	11.68	10.95	0.73
18	DAEFGH-NH2	1.3363	2.8067	−3.23	10.93	10.79	0.14
19	Ac-DAEFGH-NH2	1.3363	2.8306	−3.07	11.95	10.70	1.25
20	DAEFRHDSG-NH2	1.4468	2.9427	−5.13	10.60	11.12	0.52
21	DAEFRHDSGY-NH2	1.5636	3.0089	−3.93	11.60	12.81	1.21
22	Ac-DAEFRHDSGY-NH2	1.5636	3.0231	−3.77	12.50	12.77	0.27
23	DAEFGHDSGF-NH2	1.5908	2.9632	−3.79	13.13	13.52	0.39
24	Ac-DAEFGHDSGF-NH2	1.5908	2.9794	−3.63	14.02	13.47	0.55
25	EVRHQKLVFF-NH2	1.7125	3.0796	0.56	15.53	15.86	0.33
26	Ac-EVRHQKLVFF-NH2	1.7125	3.0920	0.72	16.00	15.84	0.16
27	GSNKGAIIGLM-NH2	1.6611	3.0014	−4.12	15.47	14.25	1.22
28	GKTKEGVLY-NH2	1.5316	2.9592	−1.69	12.65	13.24	0.59
29	TKEGVLY-NH2	1.4789	2.8685	−0.50	12.90	13.31	0.41
30	GVLY-NH2	1.3815	2.6266	1.09	13.08	13.70	0.62
31	GLSPMIETIDQVR-NH2	1.7266	3.1280	−3.52	15.85	14.66	1.19
32	AGGYKPFNLETA-NH2	1.6825	3.0531	−2.22	14.28	14.80	0.52
33	GAPGGPAFPGQTQDPLYG-NH2	1.7885	3.1807	−4.86	14.57	14.90	0.33
34	Ac-ETHLHWHTVAK-NH2	1.6201	3.0975	−2.78	13.78	13.46	0.32
35	Ac-ETHLHWHTVAKET-NH2	1.6602	3.1590	−3.93	13.20	13.38	0.18
36	LHWHT	1.4371	2.7919	−0.30	13.08	13.19	0.11
37	HLHWHT	1.4681	2.8669	−1.11	13.10	12.99	0.11
38	THLHWHT	1.4953	2.9153	−1.69	13.02	12.96	0.06
39	ETHLHWHT	1.5239	2.9677	−2.27	12.87	12.92	0.05
40	Ac-EVRHQKLVFF	1.7125	3.0906	1.41	16.47	16.04	0.43
41	DAEFRH	1.3481	2.8267	−1.87	10.90	11.23	0.33
42	Ac-DAEFRH	1.3481	2.8485	−1.71	11.90	11.15	0.75
43	DAEFGH	1.3363	2.7585	−2.01	11.18	11.41	0.23
44	DD	0.5955	2.2847	−1.98	2.32	3.14	0.82
45	ML	1.1575	2.4508	−0.17	12.78	11.04	1.74
46	WW	1.3809	2.5420	2.18	15.87	14.49	1.38
47	GM	0.8351	2.3318	−1.90	8.82	6.45	2.37
48	GH	0.5899	2.2600	−1.89	2.42	3.23	0.81
49	ETS	0.7694	2.4458	−2.47	2.88	4.66	1.78
50	KETS	0.8976	2.6063	−2.91	4.20	5.52	1.32
51	AKETS	1.0000	2.6695	−3.12	5.02	6.62	1.60
52	TVAKETS	1.2060	2.8135	−3.03	9.43	8.87	0.56
53	HWHTVAKETS	1.5069	3.0222	−4.10	11.77	11.84	0.07
54	LHWHTVAKETS	1.6184	3.0656	−2.42	12.95	13.72	0.77
55	EVHHQKLVFFAKDVGSNK-NH2	1.8392	3.2748	−4.15	13.93	15.31	1.38
56	EVHHQKLVFFAQDVGSNK-NH2	1.8390	3.2709	−4.98	14.45	15.10	0.65
57	EVHHQKLVFFAGDVGSNK-NH2	1.8382	3.2561	−4.45	14.48	15.32	0.84
58	Ac-EVHHQK-NH2	1.1572	2.8621	−4.11	9.33	7.57	1.76
59	EVRHQK-NH2	1.1706	2.8556	−4.06	8.53	7.82	0.71
60	Ac-EVRHQK-NH2	1.1706	2.8766	−3.90	9.43	7.74	1.69
61	EVHHQKLVFF-NH2	1.7087	3.0711	0.35	15.52	15.80	0.28
62	Ac-EVHHQK	1.1572	2.8602	−3.42	9.50	7.77	1.73
63	EVHHQK	1.1572	2.8384	−3.58	8.52	7.85	0.67
64	Ac-EVRHQK	1.1706	2.8735	−3.20	9.73	7.95	1.78
65	Ac-EVHHQKLVFF	1.7087	3.0825	1.20	16.42	15.97	0.45
66	EVHHQKLVFF	1.7087	3.0782	1.25	16.02	16.01	0.01

$t_{R\ exp}$, experimental retention time; $t_{R\ pred}$, calculated retention time; $\triangle t_R$, the difference between the experimental and calculated retention time

2.3 Determination of chromatographic retention parameters

t_R of the set of natural amino acids and test peptides were measured with a linear gradient of 0–60% ACN with the addition of TFA, developed within a gradient time, t_G, of 20 min. The t_R values for amino acids are listed in Table 1 and for peptides in Tables 2 and 3. The analytes were chromatographed individually.

2.4 Structural descriptors of peptides

The QSRR peptide descriptor Sum_{AA} was calculated by simple addition of individual amino acid retention data from Table 1. Molecular structural descriptors of the peptides, logarithm of Van der Waals volume, log VDW_{Vol}, and logarithm of calculated *n*-octanol-water partition coefficient, clog P, were calculated by the molecular modeling program HyperChem for personal computers with the extension ChemPlus (HyperCube, Waterloo, Canada). The software performed geometry optimization by the molecular mechanics MM+ force field method. The descriptors log VDW_{Vol} and clog P were selected by means of stepwise multiple regression from a set of more than 40 calculation chemistry descriptors provided by HyperChem. The requirements of significant analysis were observed. Moreover, only those descriptors whose physical sense is more or less obvious were taken into consideration. Hence, some descriptors of obscure physical meaning, like so-called topological indices, were excluded from QSRR analysis. The structural descriptors employed are listed in Tables 2 and 3.

2.5 Statistical analysis

QSRR equations were derived by means of multiple regression analysis employing the Statistica computer program (StatSoft, Tulsa, OK, USA) run on a personal computer. Regression coefficients (± standard errors), multiple correlation coefficients, R, standard errors of estimate, s, significance levels of each term and of the whole equation, p, and the values of the *F*-test of significance, F, were calculated. To randomly select a subseries of peptides to form the training set for deriving QSRR the Kennard-Stone design method was applied within MATLAB 6.5 software.

3 Results

Searching for a statistically significant and physically meaningful QSRR model applicable to peptides we arrived at the general regression equation employing the following analyte descriptors: logarithm of the sum of gradient t_R of the amino acids comprising the individual peptide, log Sum_{AA}, logarithm of the peptide Van der Waals volume, log VDW_{Vol}, and logarithm of its theoretically calculated *n*-octanol-water partition coefficient, clog P:

$$t_R = k_1 + k_2 \log Sum_{AA} + k_3 \log VDW_{Vol} + k_4 \text{ clog P} \quad (1)$$

In Eq. 1 t_R is a specific peptide gradient HPLC t_R and k_1–k_4 are regression coefficients. Whereas the Sum_{AA} parameter denotes additive inputs to t_R due to individual component amino acids, the parameters log VDW_{Vol} and clog P must be treated as correction terms, accounting for the resulting peptide structure, which certainly is not a simple sum of the component amino acids. The descriptors VDV_{Vol} and clog P additionally account for all the modifications of the amino acids involved. The quality of the QSRR equations obtained proves the reliability of the two correction terms applied.

A subseries of 35 peptides was chosen with the use of the Kennard-Stone design method from the total number of 101 peptides chromatographed (Table 2). The subset of peptides suffices to derive a reliable QSRR equation, which can next be used to predict t_R under given HPLC conditions for any other structurally defined peptide. The model QSRR equation has the form:

$$t_R = 7.52\ (\pm 3.12) + 15.24\ (\pm 1.54) \log Sum_{AA} - 5.83\ (\pm 1.84) \log VDW_{Vol} +$$

$$p = 0.022 \qquad p = 4 \times 10^{-11} \qquad p = 0.003$$

$$+ 0.26\ (\pm 0.08) \text{ clog P} \quad (2)$$

$$p = 0.004$$

$$n = 35;\ R = 0.966;\ F = 144;\ s = 1.06;\ p < 3 \times 10^{-18}$$

The following QSRR equation is obtained with gradient t_R data for all 101 peptides:

$$t_R = 8.02\ (\pm 2.04) + 14.86\ (\pm 0.93) \log Sum_{AA} - 5.77\ (\pm 1.16) \log VDW_{Vol} +$$

$$p = 1 \times 10^{-4} \qquad p = 6 \times 10^{-29} \qquad p = 3 \times 10^{-6}$$

$$+ 0.28\ (\pm 0.06) \text{ clog P} \quad (3)$$

$$p = 3 \times 10^{-6}$$

$$n = 101;\ R = 0.963;\ F = 411;\ s = 0.97;\ p < 5 \times 10^{-55}$$

The description of t_R by Eq. 3 is very good as documented by the following criteria of statistical quality. All the regression coefficients are highly statistically significant ($p < 3 \times 10^{-6}$) as is the whole equation ($p < 5 \times 10^{-55}$). Multiple correlation coefficient, R, standard error of estimate, s, and the value of the *F*-test of significance, F, all are also excellent. The experimental gradient t_R, t_{exp}, and those calculated with the use of Eq. 3, $t_{R\,pred}$, are given in Table 3. Prediction potency of QSRR is illustrated in Fig. 1. The predicted gradient t_R are in excellent agreement with experimental data determined

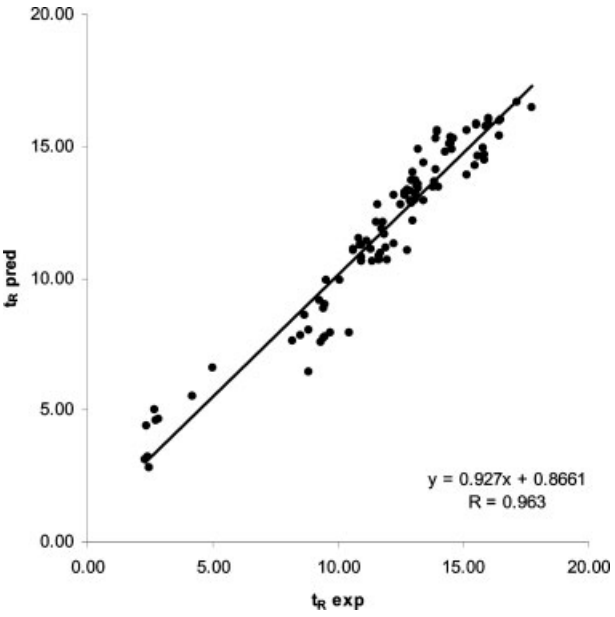

Figure 1. Correlation between the t_R calculated by QSRR (Eq. 3) and experimental t_R for a set of 101 peptides studied.

for a series of 101 structurally diversified peptides. With t_R ranging from 2.32 min to 17.72 min the mean difference between the calculated and experimental gradient t_R of peptides is less than 1 min and the mean error is 0.76 min.

The statistical significance of Eq. 3 and its individual terms is exceptionally high as documented by the statistical quality parameters R, s, F and p. It could not be attained by chance. A cross-validation procedure was performed with the use of the leave-one-out strategy within MATLAB 6.5 software to further confirm the statistical significance of Eq. 3. The calculated cross-validated root mean square error of prediction CRMSECV for Eq. 3 equals 1.00 min. The parameter < log Sum_{AA} alone gives 1.58 min for RMSECV (correlation coefficient, R, for the descriptor alone and t_R is 0.893). Adding log VDW_{Vol} to log Sum_{AA} decreases the RMSECV to 1.11 min (R increases to 0.949). After the third descriptor, clog P, is included RMSECV drops to 1.00 min (R increases to 0.961). Using log VDW_{Vol} alone gives 2.65 min for RMSECV (R = 0.661) whereas clog P alone produces RMSECV = 3.52 min (R = 0.101). If both log VDW_{Vol} and clog P are present (without log Sum_{AA}) the discussed values are: RMSECV = 1.90 min, R = 0.847. Thus, the meaningful role of log VDW_{Vol} and clog P for retention prediction cannot be questioned.

The here derived QSRR comprised peptides of up to 24 amino acid residues. Certainly, the performance of our QSRR model with longer peptides might change, but not necessarily for the worse. In the case of longer peptides the correction to log Sum_{AA} (additivity of individual amino acid t_R) by log VDW_{Vol} and clog P seems to be more pronounced, at least within our present data. However, for now, we must admit that we are uncertain how well our QSRR model will perform with longer peptides. In this work the gradient time was 20 min. Comparative experiments on another column (LiChrospher RP-18, 25.0 × 0.46 cm id; Merck, Darmstadt, Germany) with longer gradients confirmed that RMSECV would tend to increase (data not shown). It increases proportionally to the gradient time. In the case of a 20 min gradient time the RMSECV was about 1 min, for a gradient time of 60 min it was about 3 min. It must be emphasized here, however, that the correlation between the experimental and the predicted gradient t_R decreased only marginally: from R = 0.964 in the case of a 20 min gradient to R = 0.951 in the case of a 60 min gradient. We admit that much longer gradients worsen the predictions of retention significantly, *e.g.*, a 120 min gradient gives 7.1 min and R = 0.913 for RMSECV.

4 Discussion

The QSRR equation obtained predicts in a convenient and reliable manner t_R for peptides on a characterized HPLC system. In order to characterize the chromatographic system to be employed for peptide separation initial retention measurements are done for individual naturally occurring amino acids followed by measurements for a series of peptides representative enough to derive a statistically meaningful QSRR equation, *i.e.*, 15–20 diverse peptides. Next, having the structural descriptors for any peptide to be chromatographed in a characterized HPLC system one calculates its t_R. The approach proposed here applies well to gradient HPLC on standard RP columns with eluent composed of ACN and water and containing small amounts of TFA. There was a fraction of acetylated and PTM peptides within our series of analytes. Obviously, the modifications affect retention. Evidently, these modifications also correspondingly affect the structural descriptors log VDW_{Vol} and clog P employed in the QSRR analysis.

The proposed QSRR approach provides approximate, however useful, prediction of gradient retention of peptides, based solely on their chemical formulas and the contribution by individual amino acids. Thus, a rational basis for a systematic optimization of chromatographic separations of peptides instead of the trial-and-error method normally applied at present, has been elaborated. The approach consists of a gradient experiment carried out for a specific gradient time for a series of natural amino acids and test peptides. The data obtained was used to derive a QSRR equation valid for a given column/eluent system. The equation, once established, can next be used to evaluate the gradient t_R for any peptide of a defined molecular structure which might be chromatographed in the given HPLC system. Consequently, chromatographic conditions can be predetermined for any structurally defined peptides which may help to optimize their separation.

5 Concluding remarks

The advantage of our QSRR model is that it accounts for structural changes within peptides resulting from various sequences and slight structural modifications of the individual amino acids. These changes are quantitatively reflected by the descriptors which are easily calculated from the peptide's molecular formula. Therefore, the t_R we calculate from QSRR is not just a sum of retentions of individual amino acids. However, unlike in the models of Palmblad *et al.* [4] and Petritis *et al.* [13] our QSRR approach does not imply that in addition to a large number of peptides in the training set, each amino acid must be present in several peptides in several positions in this set. Finally, it must be emphasized that information from one of the initial stages of proteomic analyses, *i.e.*, LC separation of peptides, so far not utilized fully for protein identification purposes, can be exploited due to QSRR. Peptide retention predictions based on QSRR is expected to improve PMF and MS/MS ion searches. We believe that automation of the QSRR procedure and appropriate adjustment of standard bioinformatic software should cause no problem.

T. B. thanks the Foundation for Polish Science for support during the course of this work.

6 References

[1] Wilkins, M. R., Sanchez, J. C., Gooley, A. A., Appel, K. L. *et al.*, *Biotechnol. Genet. Eng. Rev.* 1996, *13*, 19–50.

[2] Bączek, T., Buciński, A., Ivanov, A. R., Kaliszan, R., *Anal. Chem.* 2004, *76*, 1726–1732.

[3] Palmblad, M., Ramström, M., Markides, K. E., Håkansson, P., Bergquist, J., *Anal. Chem.* 2002, *74*, 5826–5830.

[4] Palmblad, M., Ramström, M., Bailey, C. G., McCutchen-Maloney, S. L. *et al.*, *J. Chromatogr. B* 2004, *803*, 131–135.

[5] Meek, J. L., *Proc. Natl. Acad. Sci. USA* 1980, *77*, 1632–1636.

[6] Browne, C. A., Bennett, H. P. J., Solomon, S., *Anal. Biochem.* 1982, *124*, 201–208.

[7] Casal, V., Martin-Alvarez, P. J., Herraiz, T., *Anal. chim. Acta* 1996, *326*, 77–84.

[8] Guo, D., Mant, C. T., Taneja, A. K., Parker, J. M. R., Hodges, R. S., *J. Chromatogr.* 1986, *359*, 499–518.

[9] Guo, D., Mant, C. T., Taneja, A. K., Hodges, R. S., *J. Chromatogr.* 1986, *359*, 519–532.

[10] Mant, C. T., Zhou, N. E., Hodges, R. S., *J. Chromatogr.* 1989, *476*, 363–375.

[11] Houghten, R. A., DeGraw, S. T., *J. Chromatogr.* 1987, *386*, 223–228.

[12] Zhou, N. E., Mant, C. T., Hodges, R. S., *Pept. Res.* 1990, *3*, 8–20.

[13] Petritis, K., Kangas, L. J., Ferguson, P. L., Anderson, G. A. *et al.*, *Anal. Chem.* 2003, *75*, 1039–1048.

[14] Kaliszan, R., *Quantitative Structure-Chromatographic Retention Relationships*, Wiley, New York, USA 1987, pp. 1–4.

[15] Bączek, T., Kaliszan, R., *J. Chromatogr. A* 2002, *962*, 41–55.

[16] Bączek, T., Kaliszan, R., *J. Chromatogr. A* 2003, *987*, 29–37.

[17] Kaliszan, R., Bączek, T., Buciński, A., Buszewski, B., Sztupecka, M., *J. Sep. Sci.* 2003, *26*, 271–282.

[18] Atherton, E., Sheppard, R. C., *Solid Phase Peptide Synthesis: A Practical Approach (The Practical Approach Series)*, IRL Press at Oxford University Press, Oxford, England, 1989, pp. 112–117.

[19] Kowalik-Jankowska, T., Ruta-Dolejsz, M., Wiśniewska, K., Łankiewicz, L., Kozłowski, H., *J. Chem. Soc., Dalton Trans.* 2000, 4511–4519.

