ACTA CHROMATOGRAPHICA, NO. 18, 2007

IDENTIFICATION OF PEPTIDES IN PROTEOMICS SUPPORTED BY PREDICTION OF PEPTIDE RETENTION BY MEANS OF QUANTITATIVE STRUCTURE–RETENTION RELATIONSHIPS

T. Bączek[1,*], *C. Temporini*[2], *E. Perani*[3], *G. Massolini*[2], and *R. Kaliszan*[1]

[1]Medical University of Gdańsk, Department of Biopharmaceutics and Pharmacodynamics, Gdańsk, Poland
[2]University of Pavia, Department of Pharmaceutical Chemistry, Pavia, Italy
[3]University of Pavia, Centro Grandi Strumenti, Pavia, Italy

SUMMARY

Quantitative structure–retention relationships (QSRR) have been derived for prediction of RP-HPLC retention of peptides obtained by on-line digestion of myoglobin. To characterize the structure of a peptide quantitatively, and then to predict its retention time under gradient HPLC conditions, the structural descriptors used were: the logarithm of the sum of retention times of the amino acids of the peptide, log Sum_{AA}; the logarithm of the Van der Waals volume of the peptide, log VDW_{Vol}; and the logarithm of its calculated *n*-octanol–water partition coefficient, *c*log *P*. The predictive power of the QSRR model was checked by use of a myoglobin digest, after separation and identification of the peptides by LC–ESI-MS–MS. On-line protein digestion was performed by use of trypsin immobilized on an epoxy-modified silica monolithic support coupled on-line to LC–ESI-MS–MS. The predicted gradient retention times of the peptides were related to the experimental retention times obtained after on-line digestion of myoglobin. Identification of the components of the protein digest was supported by QSRR analysis. The QSRR approach was used as an additional constraint in proteomic research to verify results from MS–MS ion search, and to confirm both correctness of peptide identifications and indications of potential false positive and false negative results. The results suggest that because of the use of QSRR for prediction of peptide retention, information derived from standard liquid chromatographic separation in proteomics research could also be useful for eventual identification of the peptides.

INTRODUCTION

Previous studies [1,2] have revealed the high utility of quantitative structure–retention relationships (QSRR) for predicting reversed-phase high performance liquid chromatographic (RP-HPLC) retention of peptides on a variety of stationary phases under different HPLC conditions. To obtain these QSRR the structural descriptors used were:

- the logarithm of the sum of the gradient retention times of the amino acids of the individual peptide, log Sum_{AA};
- the logarithm of the peptide Van der Waals volume, log VDW_{Vol}; and
- the logarithm of its calculated *n*-octanol–water partition coefficient, *c*log *P*.

The general QSRR equation has the form:

$$t_R = k_1 + k_2 \log Sum_{AA} + k_3 \log VDW_{Vol} + k_4\, c\log P \qquad (1)$$

where t_R is the gradient HPLC retention time of a peptide and k_1–k_4 are regression coefficients.

The effect of amino acid composition on the chromatographic behaviour of peptides in reversed-phase high-performance liquid chromatography has been described in several reports [3–6]. So-called "retention coefficients" representing the contribution to peptide retention of the individual amino acids have usually been used [3]. These "retention coefficients" were derived by regression analysis and use of a set of peptide retention data. The values of the retention times of each amino acid were successively changed by 0.2 min until a good correlation between actual and predicted retention times was achieved. Similar strategies based on retention coefficients have been proposed by Browne et al. [4], Casal et al. [5], Guo et al. [6], and Palmblad et al. [7]. Palmblad et al. [7] attempted to use the approach to predict retention times of tryptic digest peptides in proteomic analysis.

To facilitate the approach based on retention coefficients, Petritis et al. [8] used artificial neural networks (ANNs). The predictive capability of an ANN was tested by use of large sets of well-identified peptides of microorganism proteomes. The objective of this approach was to increase the confidence of peptide identification [7,8]. Very recently, Strittmatter et al. [9] and Kawakami et al. [10] have used more sophisticated approaches to demonstrate the usefulness of peptide retention time predictions in proteomics. Strittmatter et al. [9] incorporated peptide retention time prediction into a discriminant function for use with tandem mass spectrometry

data analyzed with Sequest software. A database containing a set of known proteins and the proteome of *Drosophila melanogaster* was searched for false positive evaluation. A similar approach was also used by Kawakami et al. [10].

All these approaches were based on a simple amino-acid-composition-of-the-peptide-based dependence. A few reports also consider effects other than those of the amino-acid-composition-of-the-peptide (e.g. Mant et al. [11], in addition to the contribution of the amino acids to the retention of peptides, took into consideration, to some extent, polypeptide chain length).

A-priori prediction of the properties (biological or physicochemical) of chemical substances from their structural formulae is a fundamental yet still difficult problem in chemistry. Quantitative structure–property relationships may enable reliable property predictions. To derive such relationships, accurate, reproducible property measurement, and unambiguously defined structural features of the chemical entities under consideration encoding specific information on their individual property aspects are necessary. Chromatography might be a good source of comparable measures of properties that could conveniently be collected for a series of analytes representative in terms of structure. Since their introduction in the late 1970s, quantitative structure–retention relationships (QSRR) have been regarded as a method of choice for testing the performance of a variety of chemometric data-processing methods and the property predictive power of numerous structural descriptors, especially those provided by computational chemistry [12–14]. Previous studies in our laboratory [15–18] have revealed the good retention prediction performance of a general QSRR model using the structural analyte descriptors:

1. total dipole moment, μ;
2. electron excess charge of the most negatively charged atom, δ_{Min}; and
3. water-accessible molecular surface area, A_{WAS}.

The model worked well for low-molecular-mass analytes chromatographed on a variety of HPLC columns. Attempts to apply this QSRR model to peptide analytes were unsuccessful, however. Instead, we obtained good predictions of gradient HPLC retention times of peptides by using the QSRR model described by eq. (1).

An important issue in proteomics is to find algorithms enabling unambiguous protein identification on the basis of a search of bioinformatics databases and mass spectrometry data. In one approach, molecular mass data theoretically obtained for peptides from enzymatic digestion of

a protein are compared with experimental data (the so-called peptide mass fingerprinting approach) [19]. Another possibility involves use of MS–MS data of peptides to confirm the identification of the protein (the so-called MS–MS ions-search approach). The experimental data are usually compared with the calculated peptide mass or fragment ion mass values obtained by applying appropriate cleavage rules to the entries in a sequence database. Corresponding mass values are then counted and scored so that the peptide or protein to be identified is the best match to the data from the database [19].

In 1994 Yates and co-workers [20–23] developed the correlation algorithm Sequest for identification of proteins. It matches the actual peptide tandem mass spectrometry data with appropriate data from protein databases. A collection of statistics is presented, which helps to classify each match. Initially, the difference between the normalized cross-correlation functions for the first and second ranked results (ΔC_n) is used to indicate a correctly selected peptide sequence. Additional criteria are then added, including the cross-correlation score between the observed peptide fragment mass spectrum and the theoretically predicted spectrum (X_{corr}), the preliminary score based on the number of ions in the MS–MS spectrum that match the experimental data (S_{p}), the rank of the particular match during the preliminary scoring (RS_{p}) and the ions value (I), which describes how many of the detected (observed) ions match the theoretical ions for the peptide listed. Current interest is focused on filtering criteria based on X_{corr} and ΔC_n only, applied by different researchers to develop their approaches [24–28]. One of the first considerations of the filtering criteria can be found in a paper by Washburn et al. [24]. Peptides in the +1 charge state were accepted if they were fully tryptic and X_{corr} was at least 1.9. Peptides in the +2 charge state were accepted if they were fully or partially tryptic and X_{corr} was between 2.2 and 3.0. Finally, peptides in the +3 charge state were accepted if they were fully or partially tryptic and X_{corr} was >3.75. For all the spectra analyzed, ΔC_n values were >0.08. On the basis of studies by Peng et al. [25], new criteria were estimated with the goal of an overall estimated false-positive rate of less than 1% – with a ΔC_n score of >0.08 and X_{corr} greater than 2.0, 1.5, or 3.3 for the charge states +1, +2, and +3, respectively, for fully tryptic peptides, and an X_{corr} score >3.0 (+2 charge state) or >4.0 (+3 charge state) for partially tryptic peptides. Very detailed considerations on the application of different filtering criteria producing different degrees of false-positive identification have recently been reported by Qian et al. [28]. According to those authors, all previously de-

veloped filtering criteria were evaluated using either standard proteins or relatively simple proteomes (e.g. the yeast proteome). Hence the number of false positive results generated when these criteria are extended to the significantly more complex human proteome and other mammalian proteomes has not yet been characterized. In general, however, it was suggested that the probability of a random match increases with the size of the protein database. Two new sets of filtering criteria were therefore developed independently for human plasma and human cell-line samples. For human plasma samples the new criteria include: for the +1 charge state, $X_{corr} \geq 2.0$ for fully tryptic peptides and $X_{corr} \geq 3.0$ for partially tryptic peptides; for the +2 charge state, $X_{corr} \geq 2.4$ for fully tryptic peptides and $X_{corr} \geq 3.5$ for partially tryptic peptides; and for the +3 charge state, $X_{corr} \geq 3.7$ for fully tryptic peptides and $X_{corr} \geq 4.5$ for partially tryptic peptides. All the criteria had a ΔC_n value of ≥ 0.1. For human cell line samples the new criteria were: for the +1 charge state, $X_{corr} \geq 1.5$ for fully tryptic peptides and $X_{corr} \geq 3.1$ for partially tryptic peptides; for the +2 charge state, $X_{corr} \geq 1.9$ for fully tryptic peptides and $X_{corr} \geq 3.8$ for partially tryptic peptides; and for the +3 charge state, $X_{corr} \geq 2.9$ for fully tryptic peptides and $X_{corr} \geq 4.5$ for partially tryptic peptides. Again, all the criteria had a ΔC_n value of ≥ 0.1.

The objective of the current project was to test whether the sensitivity and reliability of the derived QSRR model were sufficient to enable prediction of RP-HPLC retention of peptides originating from protein digest obtained on-line by use of bioreactor based on trypsin immobilized on a monolithic support [29]. Interpretation of peptide retention predictions is proposed in respect of the correctness of their identifications and the possibility of false positives and false negatives.

EXPERIMENTAL

Equipment

On-column digestion and peptide analysis were performed with the column-switching equipment depicted in Fig. 1. Chromatographic experiments were performed with two HPLC systems. System 1 consisted of an HP 1050 isocratic pump, a thermostatted column oven, a Surveyor autosampler (Thermo Finnigan, San Jose, CA, USA) set at 37.0 ± 0.1°C, a trypsin bioreactor (25 mm × 4.6 mm i.d.) and a C_{18} trapping column (C-18 Kromasil 100, 10 mm × 4.6 mm i.d.). System 2 consisted of a quaternary gradient pump, a Surveyor LC system equipped with a diode-array detector,

and an LCQ DECA ion-trap mass spectrometer (MS) with an electrospray ionization (ESI) ion source controlled by Xcalibur software 1.3 (Thermo Finnigan). The analytical column was a 100 mm × 2.1 mm i.d., particle size 3.5 μm, Symmetry 300 C_{18} (Waters, Milford, MA, USA).

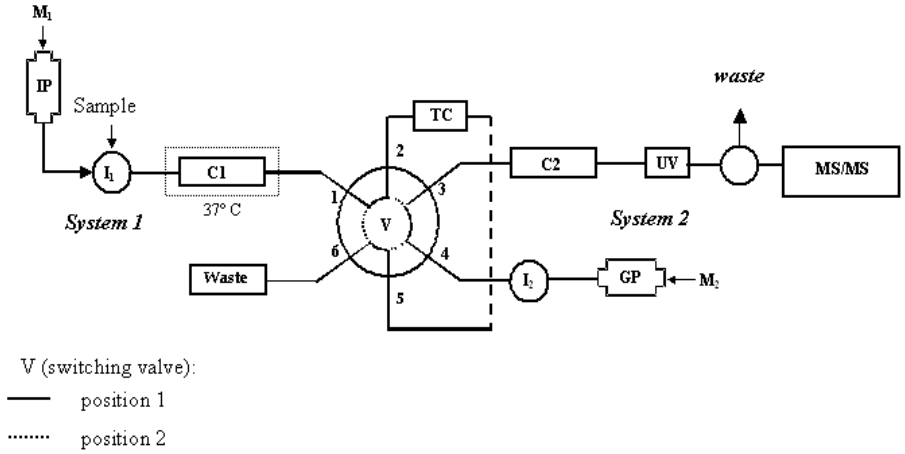

Fig. 1

Schematic diagram of the chromatographic equipment used for on-column digestion and LC–ESI-MS–MS analysis of peptide mixtures from a trypsin bioreactor. IP, isocratic pump; I_1, thermostatted column oven autosampler set at 37.0 ± 0.1°C; M_1, 100 mM phosphate buffer (pH 7.0) delivered at 1.0 mL min^{-1}; C1, trypsin bioreactor; TC, C_{18} trapping column; GP, quaternary gradient pump; M_2, gradient mobile phase delivered at a flow rate of 0.3 mL min^{-1} (the gradient conditions are given in Table I); C2, analytical column; UV, diode-array detector; MS–MS, ion-trap mass spectrometer with electrospray ionization ion source; V, automatically controlled six-port Rheodyne sample valve

The experiments were performed in the positive-ion mode under constant instrumental conditions: source potential 4.5 kV, capillary potential −20 V, sheet gas flow 70 (arbitrary units), auxiliary gas flow 20 (arbitrary units), capillary temperature 200°C, tube lens potential –5 V. MS–MS spectra were obtained by collision-induced dissociation (CID). Studies in the ion trap were performed with an isolation width of 3 Th (*m*/*z*), the activation amplitude was approximately 35% of the ejection radio frequency (RF) amplitude, which corresponded to 1.85 V.

The MS–MS spectra acquired were automatically searched against the protein database for equine proteins using Sequest software (Bioworks 3.0 package, Thermo Finnigan). Initially no specific searching criteria we-

re used. Subsequently, during interpretation of the results obtained after correlation analysis of experimental and predicted retention times of peptides, typical filtering criteria used in our studies were the same as those discussed previously, proposed by Washburn et al. [24]. The spectra for singly-charged peptides with a cross-correlation score to a tryptic peptide (X_{corr}) greater than 1.9, the spectra for doubly-charged tryptic peptides with X_{corr} of at least 2.2, and the spectra for triply-charged tryptic peptides with $X_{corr} > 3.75$ were accepted as correctly identified by the Sequest software. For all the spectra analyzed, ΔC_n values were >0.08.

Chromatographic measurements required to develop and test the primary QSRR model were performed with LC Module I plus (Waters) HPLC equipment comprising a pump, a variable-wavelength UV–visible detector, an autosampler, and a thermostat. Data were collected using Waters Millennium 2.15 software. In these measurements, as in LC–MS–MS experiments, a 100 mm × 2.1 mm i.d., particle size 3.5 μm, Symmetry 300 C_{18} column (Waters) was used.

Chromatographic Conditions

Systems 1 and 2 could be used independently or the eluent from System 1 could be automatically directed to System 2 through a six-port Rheodyne sample valve, controlled with a Kontron (Bletchley, UK) valve interface 492 (V in Fig. 1).

In step 1 (valve in position 1) the sample was loaded on the enzymatic column (C1); 100 mM phosphate buffer (pH 7.0) (M1) was used as mobile phase, delivered by IP at 1.0 mL min^{-1}. A C_{18} column trap (TC) was inserted for concentration and desalting of the tryptic digest.

In step 2 (valve in position 2) peptides were flushed back from the column trap by means of the quaternary gradient pump (GP). To elute peptides from the trapping column on to the analytical column, the GP started a mobile phase gradient with water containing 0.1% TFA (component A) and acetonitrile containing 0.1% TFA (component B) at a flow rate of 0.3 mL min^{-1}. Before coupling to the mass spectrometer the effluent from the analytical column (C2) was diverted to waste for the first 10 min.

In step 3 (valve in position 1) the valve was switched back to the original position to condition the trapping column. The chromatographic gradient conditions used in the separation experiments are summarized in Table I.

Table I

Chromatographic conditions for on-line LC–MS–MS

Time (min)	%A (water + 0.1% trifluoroacetic acid)	%B (acetonitrile + 0.1% trifluoroacetic acid)
0	100	0
10	100	0
75	45	55
78	0	100
80	0	100
85	100	0

For QSRR studies, gradient HPLC was again performed with mobile phase components A and B. The mobile phase was filtered through a GF/F glass microfibre filter (Whatman, Maidstone, UK) and degassed with helium during the analysis. The gradient was 0, 55, and 100% B at 0, 65, and 68 min, respectively, by analogy with the HPLC conditions listed in Table I for the on-line LC–MS–MS experiments. All chromatography was performed at 25°C with a mobile phase flow rate of 0.3 mL min^{-1}. The detection wavelength was always 223 nm. The dead time (1.47 min) was determined from the signal for mobile phase component B. Peptide samples were dissolved in water with addition of 0.10% of trifluoroacetic acid. The volume of sample injected was 20 μL.

Chemicals

Trypsin from the bovine pancreas (EC 3.4.21.4) was purchased from Sigma (St Louis, MO, USA). Potassium dihydrogen phosphate, dipotassium hydrogen phosphate for on-line digestion, trifluoroacetic acid (TFA), and acetonitrile (HPLC grade) for on-line and off-line analysis of myoglobin digest were purchased from Merck (Darmstadt, Germany). Horse heart myoglobin (Mb) was kindly provided by Professor L. Casella (University of Pavia, Italy). The apo form of Mb was prepared by the standard acid-2-butanone procedure [30].

The epoxy-modified silica Chromolith Flash (25 mm × 4.6 mm i.d.) support was prepared as a research sample at Merck by a procedure reported elsewhere [29]. Trypsin immobilization and column characterization were also performed as described elsewhere [29].

For HPLC analysis for QSRR studies, acetonitrile (HPLC grade) was purchased from P.C. Odczynniki (Gliwice, Poland) and TFA from Flu-

ka (Buchs, Switzerland). Water was prepared with a Milli-Q water-purification system (Millipore, Bedford, MA, USA). The amino acids listed in Table II were used to determine the peptide structure descriptor, Sum_{AA}, used in QSRR analysis. The peptides investigated are listed in Tables III–V.

Table II

Retention times, t_R, of natural amino acids used to derive the sum, Sum_{AA}, of gradient retention times of the amino acids comprising the individual peptide

No.	Amino acid	Letter code	t_{Rexp} (min)
1	Alanine	A	1.28
2	Arginine	R	1.33
3	Asparagine	N	1.25
4	Aspartic acid	D	1.28
5	Cysteine	C	1.33
6	Glutamic acid	E	1.31
7	Glutamine	Q	1.28
8	Glycine	G	1.25
9	Histidine	H	1.28
10	Isoleucine	I	5.07
11	Leucine	L	5.09
12	Lysine	K	1.28
13	Methionine	M	2.48
14	Phenylalanine	F	11.81
15	Proline	P	1.49
16	Serine	S	1.25
17	Threonine	T	1.28
18	Tryptophan	W	21.17
19	Tyrosine	Y	5.04
20	Valine	V	2.16

The peptides in Table III were used to derive the QSRR model. These peptides were randomly selected from a total set of 101 available peptides [1] by use of the Kennard–Stone design method in Matlab 6.5 software (The MathWorks, Natick, MA, USA). The peptides listed in Tables IV and V were used to test the predictive ability of the QSRR models derived. These were structurally diverse peptides which had not previously been used to derive the QSRR model (Table IV), or were obtained after on-line digestion of myoglobin and were identified by use of the LC–ESI-MS–MS system (Table V) [29]. Within the model and the test set of peptides (Tables III and IV) there was a fraction of acetylated and post-translationally

Table III

Structural descriptors, experimental retention times, t_{Rexp}, and calculated retention times, t_{Rpred}, and their difference, Δt_R, for a subset of 30 peptides used to derive the model QSRR equation. Individual symbols are explained in the text

No.	Peptide sequence	log Sum_{AA}	log VDW_{Vol}	*c*log *P*	t_{Rexp} (min)	t_{Rpred} (min)	$\lvert\Delta t_R\rvert$ (min)
1	GHG	0.58	2.3574	−2.63	1.52	3.79	2.27
2	LPQIENVKGTEDSGTT-NH_2	1.46	3.1736	−9.45	25.79	20.64	5.15
3	Ac-CEQDGDPE-NH_2	1.02	2.8836	−5.93	17.33	13.49	3.84
4	YKIEAVKSEPVEPPLPSQ-NH_2	1.57	3.2575	−1.94	29.41	32.28	2.87
5	LPPGPAVVDLTEKLEGQGG-NH_2	1.58	3.2262	−3.74	37.52	30.07	7.45
6	DRVYIHPF	1.47	2.9741	1.97	31.73	32.14	0.41
7	Ac-HNPGYPHNPGYPHNPGYP-NH_2	1.55	3.2501	−5.68	25.65	27.48	1.83
8	Ac-HNPGYPHNPGYPHNPGYPHNPGYP-NH_2	1.67	3.3717	−7.28	27.33	29.10	1.77
9	EVHHQKLVFFAEDVGSNK-NH_2	1.70	3.2699	−4.28	32.83	32.17	0.66
10	EVHHQKLVFFGEDVGSNK-NH_2	1.70	3.2662	−4.82	32.29	31.50	0.79
11	DAEFGHDSG-NH_2	1.34	2.8930	−5.27	18.37	20.46	2.09
12	LVFF-NH_2	1.49	2.7059	3.59	35.17	31.79	3.38
13	KTKEGVLY-NH_2	1.27	2.9363	−0.94	23.25	24.58	1.33
14	KEGVLY-NH_2	1.21	2.8140	0.07	22.91	23.41	0.50
15	EGVLY-NH_2	1.17	2.7220	0.51	22.93	22.26	0.67
16	MAGASELGTGPGA-NH_2	1.34	3.0030	−6.46	20.05	20.16	0.11
17	WHT	1.38	2.5890	−0.47	19.65	23.83	4.18
18	HWHT	1.40	2.7040	−1.29	20.19	24.38	4.19
19	SETHLHWHT	1.55	2.9985	−3.26	26.19	27.84	1.65
20	EVRHQK	0.94	2.8525	−3.36	14.72	14.65	0.07
21	Ac-DAEFRH	1.26	2.7853	−1.85	20.91	21.85	0.94
22	AA	0.41	2.1603	−0.74	1.52	0.81	0.71
23	AG	0.40	2.1047	−1.28	1.39	−0.55	1.94
24	AF	1.12	2.3402	0.95	17.01	18.09	1.08
25	YL	1.01	2.4394	1.86	19.81	18.01	1.80
26	GL	0.80	2.2505	−0.08	14.08	9.90	4.18
27	WF	1.52	2.5058	2.41	29.15	29.04	0.11
28	EVHHQK-NH_2	0.93	2.8413	−4.27	5.92	13.29	7.37
29	Ac-EVHHQKLVFF-NH_2	1.60	3.0841	0.51	34.85	34.00	0.85
30	EVRHQKLVFF	1.60	3.0699	1.04	35.47	34.48	0.99

modified peptides. The peptides AA, AG, AF, TL, DD, ML, WW, GM, GH, GL, WF, and GHG were purchased from Sigma–Aldrich (St Louis, MO, USA). Twenty naturally occurring amino acids (A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V) and angiotensin II (DRVYIHPF) used in the study were from Fluka. Other peptides were synthesized at the Department of Organic Chemistry, University of Gdańsk, Poland, by general procedures reported elsewhere [1].

Table IV

Structural descriptors, experimental retention times, t_{Rexp}, and calculated retention times, t_{Rpred}, and their difference, Δt_R, for the testing set of peptides not used to derive the QSRR equation. Other symbols are explained in the text

No.	Peptide sequence	log Sum_{AA}	log VDW_{Vol}	$c$log P	t_{Rexp} (min)	t_{Rpred} (min)	$\lvert\Delta t_R\rvert$ (min)
1	VKGTEDSGTT-NH_2	1.13	2.9326	−6.41	13.95	15.60	1.65
2	EHADLLAVVAASQKK-NH_2	1.46	3.1563	−3.89	34.95	26.85	8.10
3	VVAASQKK-NH_2	1.08	2.8874	−3.24	11.82	17.76	5.94
4	EVRHQKLVFF	1.17	2.8750	−3.36	18.60	19.33	0.73
5	SFSMIKEGDYN-NH_2	1.52	3.0645	−4.20	30.78	26.86	3.92
6	VVDLTEKLEGQGG-NH_2	1.41	3.0789	−4.20	30.83	24.95	5.88
7	HWHTVAKETS	1.53	3.0222	−4.10	23.83	26.63	2.80
8	MAGAAAAG-NH_2	1.06	2.7642	−4.37	15.62	14.83	0.79
9	DAEFRH-NH_2	1.26	2.8287	−2.57	18.17	21.48	3.31
10	KEGVLY-NH_2	1.26	2.8067	−3.23	18.28	20.47	2.19
11	DAEFRHDSG-NH_2	1.34	2.9427	−5.13	18.97	21.18	2.22
12	DAEFRHDSGY-NH_2	1.43	3.0089	−3.93	23.20	24.93	1.73
13	Ac-DAEFRHDSGY-NH_2	1.43	3.0231	−3.77	25.93	25.25	0.68
14	DAEFGHDSGF-NH_2	1.53	2.9632	−3.79	27.72	26.47	1.25
15	Ac-DAEFGHDSGF-NH_2	1.53	2.9794	−3.63	30.17	26.82	3.35
16	EVHHQKLVFF-NH_2	1.60	3.0711	0.35	35.73	33.62	2.11
17	Ac-EVRHQKLVFF-NH_2	1.60	3.0920	0.72	36.97	34.26	2.71
18	GKTKEGVLY-NH_2	1.30	2.9592	−1.69	24.98	24.50	0.48
19	TKEGVLY-NH_2	1.24	2.8685	−0.50	24.67	23.87	0.80
20	AGGYKPFNLETA-NH_2	1.53	3.0531	−2.22	31.98	29.13	2.85
21	GAPGGPAFPGQTQDPLYG-NH_2	1.62	3.1807	−4.86	33.03	29.12	3.91
22	Ac-ETHLHWHTVAK-NH_2	1.59	3.0975	−2.78	32.20	30.08	2.12
23	Ac-ETHLHWHTVAKET-NH_2	1.62	3.1590	−3.93	29.65	29.88	0.23
24	LHWHT	1.48	2.7919	−0.30	27.28	27.89	0.61
25	HLHWHT	1.50	2.8669	−1.11	28.02	28.03	0.01
26	ETHLHWHT	1.53	2.9677	−2.27	28.02	28.32	0.30
27	Ac-EVHHQKLVFF	1.60	3.0825	1.20	38.12	34.71	3.41
28	EVHHQKLVFF	1.60	3.0782	1.25	37.28	34.73	2.55
29	Ac-EVRHQKLVFF	1.60	3.0906	1.41	38.77	35.05	3.72
30	DAEFGH	1.26	2.7585	−2.01	19.05	21.41	2.36

Structural Descriptors of Peptides and Statistical Analysis

The QSRR peptide descriptor log Sum_{AA} was obtained by adding the retention times of the amino acids of a given peptide measured individually by HPLC. The molecular structural descriptors of the peptides, the logarithm of Van der Waals volume, log VDW_{Vol}, and the logarithm of the calculated *n*-octanol–water partition coefficient, $c$log P, were calculated by use of HyperChem molecular modelling software with the ChemPlus extension (HyperCube, Waterloo, Canada). The software performed geometry optimization using the molecular mechanics MM+ force-field me-

thod. The structural descriptors used in this work are summarized in Tables III–V.

Table V

Structural descriptors, experimental retention times, t_{Rexp}, and calculated retention times, t_{Rpred}, and their difference, Δt_R, for the myoglobin digest set of peptides not used to derive the QSRR equation. Other symbols are explained in the text

No.	Peptide sequence	log Sum_{AA}	log VDW_{Vol}	$c$log P	t_{Rexp} (min)	t_{Rpred} (min)	$\lvert\Delta t_R\rvert$ (min)
1	K.ELGFQG.-	1.34	2.7574	−1.95	25.82	23.04	2.78
2	K.HKIPIK.Y	1.19	2.8449	−0.44	18.11	22.73	4.62
3	K.YKELGFQG.-	1.45	2.9267	−3.14	26.91	25.40	1.51
4	K.HLKTEAEMK.A	1.22	2.9886	−3.86	18.74	20.74	2.00
5	K.ALELFRNDIAAK.Y	1.57	3.0923	−1.15	32.94	31.63	1.31
6	K.HGTVVLTALGGILK.K	1.54	3.1080	−0.49	39.92	31.96	7.96
7	K.HPGDFGADAQGAMTK.A	1.49	3.1116	−7.97	24.38	22.37	2.01
8	K.HGTVVLTALGGILKK.K	1.56	3.1485	−0.93	38.24	32.15	6.09
9	R.NDIAAKYKELGFQG.-	1.60	3.1406	−4.83	31.08	28.34	2.74
10	K.VEADIAGHGQEVLIR.L	1.51	3.1559	−3.12	28.86	28.78	0.08
11	K.ALELFRNDIAAKYK.E	1.64	3.1770	−2.11	33.72	32.63	1.09
12	K.HGTVVLTALGGILKK.K	1.56	3.1485	−0.93	37.99	32.15	5.84
13	K.HGTVVLTALGGILKK.K	1.56	3.1485	−0.93	38.61	32.15	6.46
14	K.GHHEAELKPLAQSHATK.H	1.47	3.2148	−8.92	23.74	21.87	1.87
15	K.HLKTEAEMKASEDLKK.H	1.47	3.2209	−6.94	24.70	24.16	0.54
16	K.YLEFISDAIIHVLHSK.H	1.74	3.2323	−1.10	49.37	36.20	13.17
17	K.KGHHEAELKPLAQSHATK.H	1.49	3.2463	−9.36	21.63	22.02	0.39
18	K.ALELFRNDIAAKYKELGFQG.-	1.82	3.3116	−3.51	40.54	35.71	4.83
19	R.LFTGHPETLEKFDKFKHLKTEAEMK.A	1.88	3.4293	−6.35	34.26	34.83	0.57

QSRR equations were derived by multiple regression analysis (MRA), by use of Statistica software (StatSoft, Tulsa, OK, USA). Regression coefficients (± standard errors), multiple correlation coefficients, R, standard errors of estimates, s, significance levels for each regression term and for the whole equation, p, and the values of the F-test of significance, F, were calculated. Usually the descriptive power of the model is evaluated as the root-mean-squared error (*RMSE*), computed for the calibration data (model set of peptides), which is defined as:

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

where y_i is the experimental retention of the ith calibration sample, $\hat{y}_i^{(f)}$ the predicted retention value for the ith calibration sample using the model (of complexity f), and c the number of calibration set objects.

The predictive ability of a model was characterised by the root-mean-squared error of prediction (*RMSEP*) of an independent external test set, defined as:

$$\mathrm{RMSEP}(f) = \sqrt{\frac{\sum_{i=1}^{t}(y_i - \hat{y}_i^{(f)})^2}{t}} \qquad (3)$$

where y_i is the experimental retention of the ith test set object, $\hat{y}_i^{(f)}$ the predicted retention of the ith object using the model (of complexity f), and t the number of test set objects.

RESULTS AND DISCUSSION

A subseries of thirty structurally diverse peptides summarized in Table III sufficed for derivation of a statistically reliable QSRR equation which could be used to predict the retention time of any other structurally defined peptide under appropriate HPLC conditions. The derived model QSRR equation has the form:

$$\underset{p = 0.002}{t_R = -27.10\ (\pm 7.76)} + \underset{p = 8 \times 10^{-6}}{19.17\ (\pm 3.46) \log Sum_{AA}} + \underset{p = 0.03}{9.68\ (\pm 4.16) \log VDW_{Vol}} + \underset{p = 0.0002}{1.17\ (\pm 0.27)\ c\log P} \qquad (4)$$

$n = 30$; $R = 0.956$; $F = 92$; $s = 3.13$; $p < 5 \times 10^{-14}$

That description of t_R by eq. (4) is good is apparent from the criteria of statistical quality. All regression coefficients are highly statistically significant ($p < 0.03$) as is the whole equation ($p < 5 \times 10^{-14}$). The values of the multiple correlation coefficient, R, the standard error of the estimate, s, and the value of the F-test of significance, F, are also satisfactory.

The descriptive power of the equation was verified by calculating retention times for peptides from the set used to derive the QSRR equation. Comparison of experimental and calculated retention times clearly reflected the reliability of the equation – the correlation coefficient was R = 0.956 (*RMSE* = 2.93) (Fig. 2). Equation 4 provides the predictive model based on the experimentally obtained descriptor (log Sum_{AA}) improved by implementation of two molecular-modelling-based descriptors (log VDW_{Vol} and $c\log P$). The experimentally obtained descriptor (log Sum_{AA}) seemed

to make a very significant contribution to the retention of the peptides. The retention times measured for the amino acids used to determine the peptide structure descriptor, Sum_{AA}, are listed in Table II. Some amino acids passed the column easily, some were more or less strongly retained; this was because of the different structures of the individual amino acids. It is evident that some amino acids are excluded from stationary phase. The t_R data depend on the similar (and dissimilar) behaviour of the amino acids in HPLC and, consequently, the sum of these data ($\log Sum_{AA}$) for the amino acids contained in the peptides is of great importance for predicting peptide retention under the same conditions as were used for the individual amino acids. It is sufficient to note the high significance ($p = 8 \times 10^{-6}$) of $\log Sum_{AA}$ in eq. (4). Obviously, $\log Sum_{AA}$ has little in common with octanol–water partition coefficient, either for individual amino acids or for the peptide. The analytes were highly ionizable and only a minute fraction of molecules exists in the non-ionized form in solution. Only for that fraction does $\log P$ ($c\log P$) properly reflect the ability to partition between aqueous and hydrophobic phases. Sum_{AA} does not mimic $c\log P$; it actually reflects differences between the polarities of the peptides. Instead, $c\log P$ is an auxiliary peptide structure descriptor – a correction for $\log Sum_{AA}$. In QSRR eq. (2) the significance of the $c\log P$ term ($p = 2 \times 10^{-4}$) is two orders of magnitude less than that of $\log Sum_{AA}$ ($p = 8 \times 10^{-6}$). Peptides considered in the study were selected to cover a wide range of structural diversity, including some post-translational modifications. The issue of the effect of the modifications on peptide retention was, therefore, taken into account in the work. Acetylated and post-translationally modified peptides were included in our test series. Certainly these modifications affect retention. They also appropriately affect calculation of the structural descriptors $\log VDW_{Vol}$ and $c\log P$, however. The advantage of our QSRR model is that it takes into account structural changes within peptides resulting both from different sequences and from slight modifications of the component amino acids. These changes are quantitatively reflected by descriptors readily calculable from the peptide molecular formula. The retention time calculated by us from QSRR is, therefore, not just a sum of the retentions of the component amino acids. What is more, in contrast with the models of Palmblad et al. [7] and Petritis et al. [8], 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 that set.

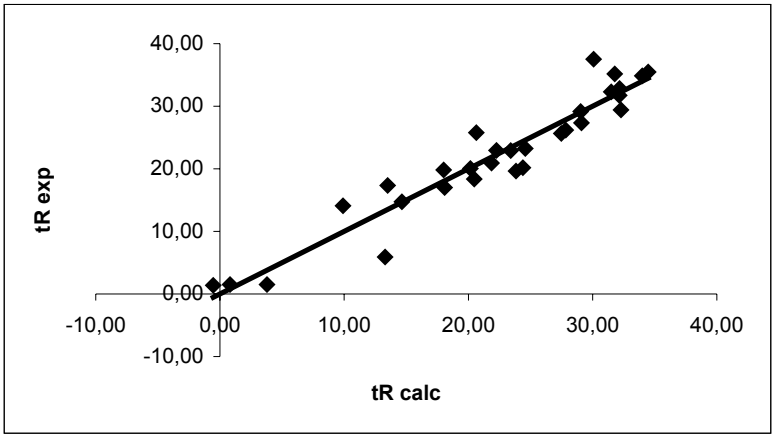

Fig. 2

Correlation between retention times calculated by use of QSRR (eq. 4) and experimental retention times for the model set of peptides used to derive the QSRR equation

The derived QSRR model was then tested by use of an independent set of peptides, not previously used to derive eq. (4). For the peptides listed in Table III the three proposed calculated structural descriptors were used to predict retention times under the same HPLC conditions as for the model set of peptides. The experimental gradient retention times, t_{Rexp}, and those calculated by use of eq. (4), t_{Rpred}, are given in Table IV. The power of prediction of the QSRR model described by eq. (4) is illustrated in Fig. 3 (correlation coefficient 0.944; *RMSEP* = 3.04). The correlation obtained for a set of structurally diverse peptides not used to derive the QSRR equation, proves its usefulness.

Finally, the power of the QSRR approach for prediction of peptide retention, with the objective of identifying individual peptides in proteomics, was verified with a protein digest. Myoglobin digest, obtained online by use of a bioreactor based on trypsin immobilized on a monolithic support and analyzed by LC–MS–MS, was identified first with the Sequest software without any restrictions. Separation and identification was performed for the 19 peptides listed in Table V. Those data, with the Sequest database search statistics for all the peptides presented in Table VI, were used for analysis of the performance of the QSRR approach, treated as a potential additional identification constraint. Correlation analysis performed on the experimental and calculated retention times for the peptides from myoglobin digest yielded good results (Fig. 4; $R = 0.928$ and *RMSEP* = 4.74).

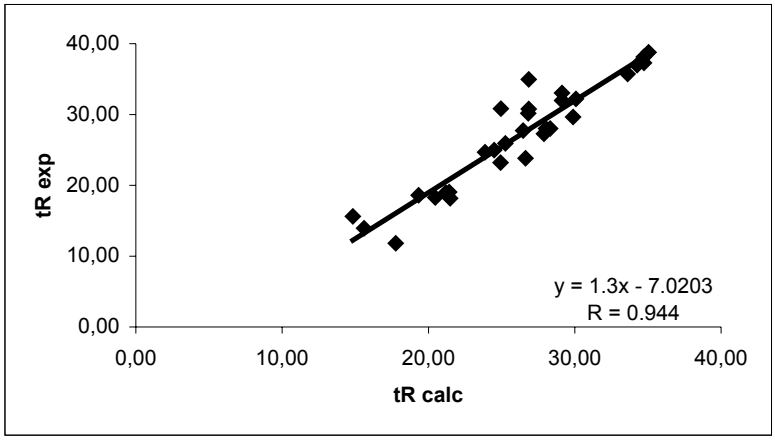

Fig. 3

Correlation between retention times calculated by use of QSRR (eq. 4) and experimental retention times for the testing set of peptides not used to derive the QSRR equation

Table VI

Collection of Sequest database searching statistics used for testing the QSRR model as the peptide identification constraint. Symbols are explained in the text

No.	Sequence	*m/z*	CH	X_{corr}	ΔC_n	S_p	RS_p	I
1	K.ELGFQG.-	650.32	1.00	1.15	1.00	201.5	1.00	7/10
2	K.HKIPIK.Y	735.49	2.00	1.85	1.00	651.9	1.00	9/10
3	K.YKELGFQG.-	941.47	2.00	2.91	1.00	725.7	1.00	12/14
4	K.HLKTEAEMK.A	1086.56	2.00	3.08	1.00	873.7	1.00	14/16
5	K.ALELFRNDIAAK.Y	1360.76	2.00	3.11	1.00	1016.2	1.00	17/22
6	K.HGTVVLTALGGILK.K	1378.84	2.00	4.97	1.00	1545.0	1.00	23/26
7	K.HPGDFGADAQGAMTK.A	1502.67	2.00	1.06	1.00	57.5	1.00	7/28
8	K.HGTVVLTALGGILKK.K	1506.94	2.00	4.58	0.92	2605.8	1.00	24/28
9	R.NDIAAKYKELGFQG.-	1553.80	2.00	2.16	1.00	644.5	1.00	14/26
10	K.VEADIAGHGQEVLIR.L	1606.86	2.00	4.09	1.00	1221.5	1.00	20/28
11	K.ALELFRNDIAAKYK.E	1651.92	2.00	2.50	1.00	518.3	1.00	13/26
12	K.HGTVVLTALGGILKK.K	1506.94	3.00	4.17	1.00	1808.6	1.00	32/56
13	K.HGTVVLTALGGILKK.K	1506.94	3.00	3.97	1.00	2444.7	1.00	30/56
14	K.GHHEAELKPLAQSHATK.H	1853.96	3.00	0.40	1.00	206.6	1.00	15/64
15	K.HLKTEAEMKASEDLKK.H	1857.97	3.00	3.69	1.00	1392.7	1.00	27/60
16	K.YLEFISDAIIHVLHSK.H	1885.02	3.00	2.54	1.00	535.2	1.00	24/60
17	K.KGHHEAELKPLAQSHATK.H	1982.06	3.00	4.50	1.00	1483.9	1.00	33/68
18	K.ALELFRNDIAAKYKELGFQG.-	2283.21	3.00	4.45	1.00	1835.9	1.00	30/76
19	R.LFTGHPETLEKFDKFKHLKTEAEMK.A	3004.56	3.00	1.86	1.00	187.1	1.00	18/96

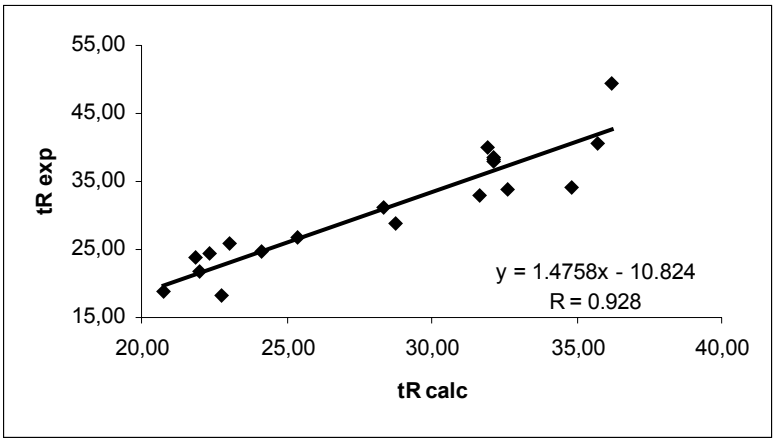

Fig. 4

Correlation between retention times calculated by use of QSRR (eq. 4) and experimental retention times for a testing set of peptides originating from a myoglobin digest not used to derive the QSRR equation

The data were then further analyzed. Taking into account the confidence level of 0.9 for the correlation analysis and the cross-correlation score values for singly, doubly, and triply charged tryptic peptides, considered as correctly identified with the Sequest software ($X_{corr} > 1.9$ (+1), 2.2 (+2) and 3.75 (+3), respectively), several conclusions could be drawn. For seven peptides (K.YKELGFQG.-, K.HLKTEAEMK.A, K.HGTVVLTALGGILKK.K, R.NDIAAKYKELGFQG.-, K.HGTVVLTALGGILKK.K, K.KGHHEAELKPLAQSHATK.H, and K.ALELFRNDIAAKYKELGFQG.-) the correctness of Sequest identification was confirmed by the QSRR approach (Fig. 5). Also, bearing in mind the QSRR-based predictions of peptide retention, confirmation of the incorrect identification as suggested by X_{corr} values was obtained for six other peptides (K.ELGFQG.-, K.HKIPIK.Y, K.HPGDFGADAQGAMTK.A, K.GHHEAELKPLAQSHATK.H, K.YLEFISDAIIHVLHSK.H, and R.LFTGHPETLEKFDKFKHLKTEAEMK.A). Finally, for five identified peptides, according to the filtering criteria used, the QSRR model constraint indicated potential false positives (K.ALELFRNDIAAK.Y, K.HGTVVLTALGGILK.K, K.VEADIAGHGQEVLIR.L, K.ALELFRNDIAAKYK.E, and K.HGTVVLTALGGILKK.K). We also identified a potential false negative, K.HLKTEAEMKASEDLKK.H.

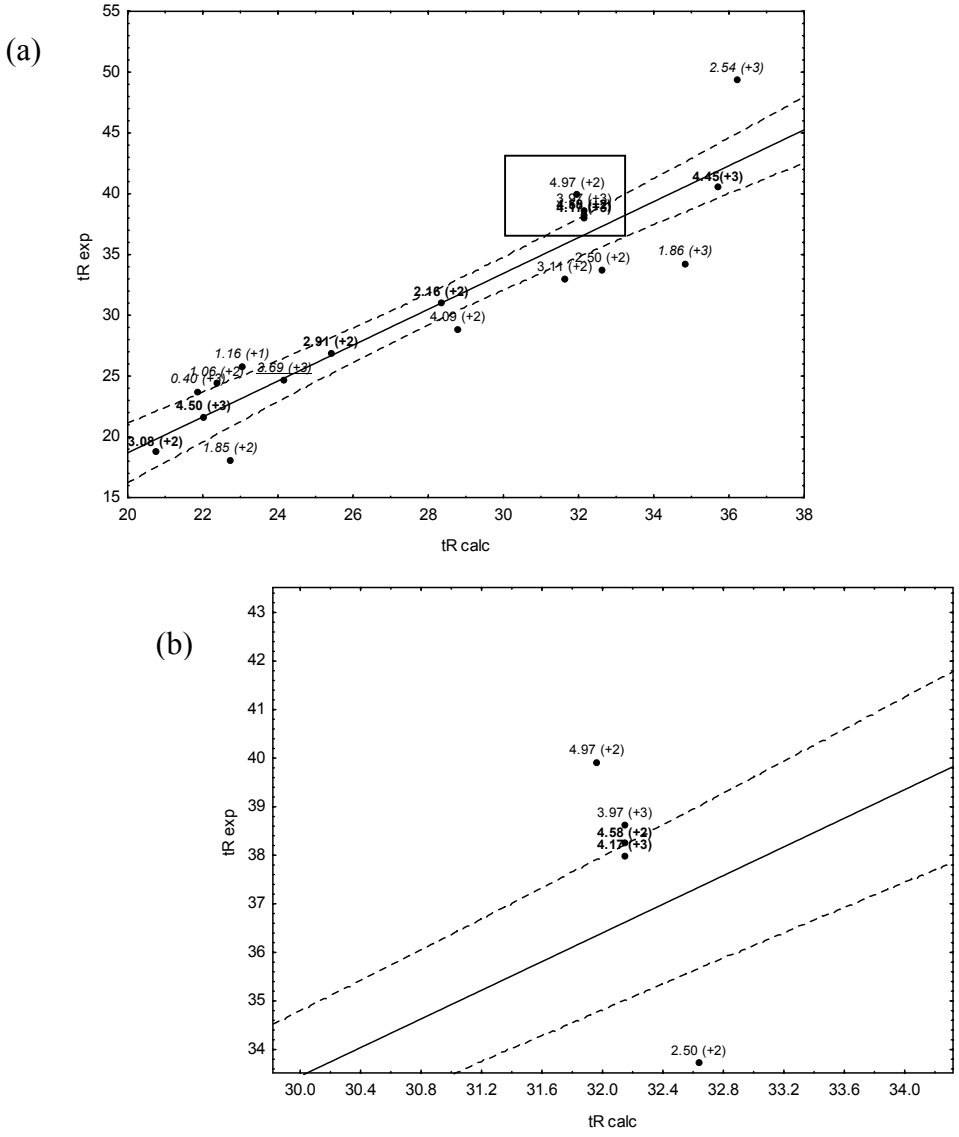

Fig. 5

(a) Correlation between experimental gradient retention times and gradient retention times calculated by use of eq. (4) for all the peptides identified with the system based on on-line digestion and LC–ESI-MS–MS analysis. The peptides indicated in bold are those for which correct identification was confirmed. The peptides indicated in italics are those for which incorrect identification was confirmed. Normal print indicates the peptides recognized as possible false positives. The peptide indicated by underlining and italics was recognized as a possible false negative. The area within the confidence level = 0.9 is indicated with a dotted line. (b) Enlargement of the region indicated with a square in (a)

CONCLUSIONS

The QSRR equations obtained in our work enabled prediction of the retention times of tryptic peptides after on-line digestion of myoglobin. To characterize the HPLC system used for peptide separation, measurements of retention were performed individual for naturally occurring amino acids and for a representative series of peptides. Thirty different peptides were used to derive a statistically significant model QSRR equation. Using the same structural descriptors for other peptides, chromatographed under the same LC conditions, one can calculate their retention times. Thus QSRR information from liquid chromatography can be used for protein identification. The peptide retention predictions based on QSRR can be regarded as an additional constraint verifying the correctness of peptide MS–MS ion search. The simple QSRR model contains only three descriptors of peptides calculable from the structural formula – log Sum_{AA}, log VDW_{Vol}, and *c*log *P*.

ACKNOWLEDGEMENTS

The work was supported in part by the Polish State Committee for Scientific Research Projects 2 P05F 012 27 and 2 P05F 041 30 and by the Italian MURST Research Project 2004038884_004.

REFERENCES

[1] R. Kaliszan, T. Bączek, A. Cimochowska, P. Juszczyk, K. Wiśniewska, and Z. Grzonka, Proteomics, **5**, 409 (2005)
[2] T. Bączek, P. Wiczling, M.P. Marszałł, Y. Vander Heyden, and R. Kaliszan, J. Proteome Res., **4**, 555 (2005)
[3] J.L. Meek, Proc. Natl. Acad. Sci. USA, **77**, 1632 (1980)
[4] C.A. Browne, H.P.J. Bennett, and S. Solomon, Anal. Biochem., **124**, 201 (1982)
[5] V. Casal, P.J. Martin-Alvarez, and T. Herraiz, Anal. Chim. Acta, **326**, 77 (1996)
[6] D. Guo, C.T. Mant, A.K. Taneja, J.M.R. Parker, and R.S. Hodges, J. Chromatogr., **359**, 499 (1986)
[7] M. Palmblad, M. Ramström, K.E. Markides, P. Håkansson, and J. Bergquist, Anal. Chem., **74**, 5826 (2002)

[8] K. Petritis, L.J. Kangas, P.L. Ferguson, G.A. Anderson, L. Pasa-Tolic, M.S. Lipton, K.J. Auberry, E.F. Strittmatter, Y. Shen, R. Zhao, and R.D. Smith, Anal. Chem., **75**, 1039 (2003)
[9] E.F. Strittmatter, L.J. Kangas, K. Petritis, H.M. Mottaz, A.A. Anderson, Y. Shen, J.M. Jacobs, D.G. Camp II, and R.D. Smith, J. Proteome Res., **3**, 760 (2004)
[10] T. Kawakami, K. Tateishi, K. Yamano, T. Ishikawa, K. Kuroki, and T. Nishimura, Proteomics, **5**, 856 (2005)
[11] C.T. Mant, N.E. Zhou, and R.S. Hodges, J. Chromatogr., **476**, 363 (1989)
[12] R. Kaliszan, Quantitative Structure–Chromatographic Retention Relationships, Wiley, New York, 1987
[13] R. Kaliszan, J. Chromatogr., **656**, 417 (1993)
[14] E. Forgacs and T. Cserhati, Molecular Bases of Chromatographic Separations, Boca Raton, CRC Press, 1997
[15] R. Kaliszan, M.A. van Straten, M. Markuszewski, C.A. Cramers, and H.A. Claessens, J. Chromatogr. A, **855**, 455 (1999)
[16] T. Bączek and R. Kaliszan, J. Chromatogr. A, **962**, 41 (2002)
[17] T. Bączek and R. Kaliszan, J. Chromatogr. A, **987**, 29 (2003)
[18] R. Kaliszan, T. Bączek, A. Buciński, B. Buszewski, and M. Sztupecka, J. Sep. Sci., **26**, 271 (2003)
[19] D.N. Perkins, D.J.C. Pappin, D.M. Creasy, and J.S. Cottrell, Electrophoresis, **20**, 3551 (1999)
[20] J.K. Eng, A.L. McCormack, and J.R. Yates III, J. Am. Soc. Mass Spectrom., **5**, 976 (1994)
[21] J.R. Yates III, J.K. Eng, A.L. McCormack, and D. Schieltz, Anal. Chem., **67**, 1426 (1995)
[22] D.C. Anderson, W. Li, D.G. Payan, and W.S. Noble, J. Proteome Res., **2**, 137 (2003)
[23] D.L. Tabb, W.H. McDonald, and J.R. Yates III, J. Proteome Res., **1**, 21 (2002)
[24] M.P. Washburn, D. Wolters, and J.R. Yates III, Nat. Biotechnol., **19**, 242 (2001)
[25] J. Peng, J.E. Elias, C.C. Thoreen, L.J. Licklider, and S.P. Gygi, J. Proteome Res., **2**, 43 (2003)
[26] L. Florens, M.P. Washburn, J.D. Raine, R.M. Anthony, M. Grainger, J.D. Hayness, J.K. Moch, N. Muster, J.B. Sacci, D.L. Tabb, A.A. Witney, D. Wolters, Y. Wu, M.J. Gardner, A.A. Holder, R.E. Sinden, J. Yates, and D.J. Carucci, Nature, **419**, 520 (2002)

[27] J.N. Adkins, S.M. Varnum, K.J. Auberry, R.J. Moore, N.H. Angell, R.D. Smith, D.L. Springer, and J.G. Pounds, Mol. Cell. Proteomics, **1**, 947 (2002)
[28] W.-J. Qian, T. Liu, M.E. Monroe, E.F. Strittmatter, J.M. Jacobs, L.J. Kangas, K. Petritis, D.G. Camp II, and R.D. Smith, J. Proteome Res., **4**, 53 (2005)
[29] E. Calleri, C. Temporini, E. Perani, C. Stella, S. Rudaz, D. Lubda, G. Mellerio, J.-L. Veuthey, G. Caccialanza, and G. Massolini, J. Chromatogr. A, **1045**, 99 (2004)
[30] E. Antonini and M. Brunori, Hemoglobin and Myoglobin in their Reactions with Ligands, North Holland Publishing, Amsterdam, 1971