Journal of Chromatography A, 1217 (2010) 489–497

Contents lists available at ScienceDirect

Journal of Chromatography A

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

A model for predicting slopes *S* in the basic equation for the linear-solvent-strength theory of peptide separation by reversed-phase high-performance liquid chromatography

Hoangkim Vu [a], Vic Spicer [a], Alexander Gotfrid [a], Oleg V. Krokhin [b,c,*]

[a] Department of Physics and Astronomy, University of Manitoba, Winnipeg, R3T 2N2, Canada
[b] Manitoba Centre for Proteomics and Systems Biology, University of Manitoba, 799 JBRC, 715 McDermot Avenue, Winnipeg, R3E 3P4, Canada
[c] Department of Internal Medicine, University of Manitoba, 799 JBRC, 715 McDermot Avenue, Winnipeg, R3E 3P4, Canada

ARTICLE INFO

Article history:
Received 25 August 2009
Received in revised form 21 October 2009
Accepted 23 November 2009
Available online 27 November 2009

Keywords:
Linear-solvent-strength theory
Peptide reversed-phase HPLC
Peptide retention prediction

ABSTRACT

A model for predicting the slope (*S*) in the fundamental equation of linear-solvent-strength theory for peptidic compounds was developed. Our approach is based on the novel assumption that three well-defined molecular descriptors: peptide length (*N*), charge (*Z*) and hydrophobicity index (*HI*) are the major contributors to the value of *S*. Following the definition of the model's variables, the retention of a number of Arg-terminated synthetic peptides was investigated under isocratic elution conditions (100 Å pore size C18 phase, 0.1% trifluoroacetic acid as ion-pairing modifier). The peptide sequences were systematically designed to span the properties of the typical tryptic peptides that are analyzed in proteomic experiments. Experimental data show that slopes *S* increase with the independent increase in both peptide charge and peptide length when the two other parameters are held constant. The influence of peptide hydrophobicity is more complex: depending on peptide length and charge, stronger RP-HPLC retention can either decrease or increase the values of *S*. We postulate a general function to explain this behavior: $S = C1 \times Z^{C2} + C3 \times N^{C4} + C5 \times HI^{C6} + C7/Z + C8/N + C9/HI + C10 \times ZN + C11 \times ZHI + C12 \times NHI + B$. A simple optimization using a "random walk" through parameter-space was used to determine the optimal coefficients compared to the measured *S*-values of 37 peptides. The model gives a ~0.97 R^2 correlation between the measured and predicted *S*-values: it was verified against previously published data on a human growth hormone protein tryptic digest and some synthetic analogues from that mixture.

1. Introduction

It is hard to overestimate the role of reversed-phase (RP) HPLC in studying biologically active compounds, particularly proteins and peptides. It provides insight into structure and function, as well as aiding the development of methods for purification and large-scale production of biologically active compounds. This role has further increased with the accelerated development of proteomic studies in recent years. Thousands of research laboratories and industrial facilities use the peptide/protein RP-HPLC technique on daily basis.

Fundamental understanding of the separation mechanism of peptidic molecules on alkyl-bonded silicas was developed in the 1980–1990s [1–3], following the extension of the linear-solvent-strength (LSS) theory of RP-HPLC for the separation of peptides and proteins [4]. Typically peptide separation is achieved via water:organic solvent gradients, while mobile phases are supplemented with ionogenic (ion-pairing) modifiers, which improve peak shape and maintain the desired pH of the eluent solution. In general, retention of peptides and proteins under RP conditions is described by the relationship: $\log k = \log k_w - S\phi$; where k is the retention factor at an organic solvent volume fraction ϕ (ϕ = ACN%/100) and k_w is the retention factor at $\phi = 0$ [4], similar to RP-HPLC of low molecular weight compounds.

One of the distinct features of the LSS theory of peptide RP-HPLC is the extremely high *S*-value for peptidic molecules [5], which tends to increase with molecular weight. In other words, peptidic compounds elute from alkyl-bonded media in a very narrow range of acetonitrile concentrations. This makes the use of gradient elution mandatory when separating complex peptide mixtures. Either knowing, or being able to predict *S*-values for peptides is important from the point of view of optimizing the separation selectivity in RP-HPLC, in order to provide complete resolution of chromatographic bands under gradient or isocratic conditions. It has been shown on multiple occasions that altering column size and gra-

* Corresponding author at: Department of Internal Medicine, University of Manitoba, 799 JBRC, 715 McDermot Avenue, Winnipeg, R3E 3P4, Canada.
Tel.: +1 204 789 3283; fax: +1 204 480 1362.
E-mail address: krokhino@cc.umanitoba.ca (O.V. Krokhin).

0021-9673/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2009.11.065

dient/elution parameters could result in a significant change in separation selectivity [5,6]. For proteomic applications, the complete separation of the components is not as critical, since modern mass spectrometric detectors can handle detection (identification) of co-eluting compounds. However, the influence of slope S on separation selectivity becomes important when peptide retention prediction algorithms are used to strengthen peptide identification, or when LC–MS data sets obtained using different column sizes, flow rates, or gradient slopes are compared for systematic collection [7]. Thus, the use of identical column sizes and gradient conditions is preferable for comparison of the selectivity of RP-HPLC columns for peptide separation. Otherwise, a variation in selectivity might be observed simply as a consequence of various S-slopes, rather than from a difference in separation selectivity itself.

It was initially suggested that the magnitude of S-values is linked directly to the molecular weight of the analytes. Snyder and co-workers [4] described this dependence as $S = 0.44(MW)^{0.21}$ for a set of polypeptides ranging 600–14,000 Da in molecular weight. Following this pioneering work, it was assumed that a similar relationship should be observed for other sets of peptides. However, subsequent measurements performed by Hearn and co-workers [8–11] on smaller analytes with a narrower coverage of MW range showed that this formula was incorrect; they concluded that the parameters of $S = a(MW)^b$ should be varied to obtain a better correlation for each particular set of peptides and chromatographic conditions. They also found that the slope S does not follow a simple dependency on the hydrophobicity of these peptides [8]. Altogether, these studies led to the conclusion that the slope S is determined not by molecular weight alone, but rather by a combination of the "magnitude of hydrophobic contact area and the number of interaction sites" [9,10]. It is conceivable that an increase in molecular weight of a peptide could itself lead to an increase in hydrophobicity and the number of contact sites. However, there have also been multiple observations where, on the one hand, well-retained peptide solutes exhibit low S-values, and on the other, very hydrophilic small peptides exhibit high S-values. In addition, some peptides, upon interaction with hydrophobic alkyl-silica surfaces, can adopt a preferred folding conformation; this could be a reason for such deviations making the overall picture even more complex. Altogether, complexity and mixed character retention mechanisms (hydrophobic, ion-pairing, interactions with free silanol groups) have precluded development of a quantitative model for predicting S-values for peptidic compounds.

To our knowledge, the charge of a peptide/protein was never considered as a parameter that might impact S-values. Nevertheless, increasing the molecular weight of naturally occurring peptides typically leads to a larger number of basic residues (R, K, H). These residues carry positive charge at the pHs of the mobile phases typically used for peptide RP-HPLC, and are involved in ion-pairing interactions. In light of the previous discussion on the number of interacting sites, it seems reasonable that this would directly influence the slopes S. Recently we showed that ion-pairing must be taken into account for accurate peptide retention prediction [12]. Further, the hydrophobicity of N-terminal amino acids and the residues adjacent to positively charged ones are significantly affected by the formation of ion pairs [13]. This also causes selectivity variations upon switching between various additives to the mobile phase (for example trifluoroacetic vs. formic acid). Given its significant impact on peptide interactions in RP-HPLC, it seems logical to consider peptide charge as one of the parameters affecting the slopes S.

The first goal of this study was to define a set of molecular descriptors of peptide molecules, which can be easily calculated (or estimated) and used to derive S-values in the basic LSS equation. While the general definition of both hydrophobic contact area and the number of interaction sites leaves some ambiguity in the numerical expression of these parameters, peptide length, charge and hydrophobicity can be calculated or measured with sufficient accuracy. Following the definition of the parameters and the space for variation of these variables (typical for tryptic peptides in proteomic studies) we have attempted to develop a model for predicting the slopes S based on the experimental measurements for a set of model peptides.

2. Materials and methods

2.1. Materials

Deionized (18 MΩ) water and HPLC-grade acetonitrile were used for the preparation of eluents. All chemicals were sourced from Sigma–Aldrich (St. Louis, MO). The 37 model peptides and 5 peptides corresponding to the human growth hormone sequence were custom synthesized by BioSynthesis Inc. (Lewisville, TX) and the peptide ALILTLVS was purchased from Bachem Americas (Torrance, CA). Table 1 shows the list of peptides, together with their core properties: molecular weight, charge, length, and hydrophobicity.

2.2. Instrumentation

A micro-Agilent 1100 Series system (Agilent Technologies, Wilmington, DE), was used with a manual injector (loop size 10 μl) and a UV detector operated at 214 nm. All chromatographic experiments were conducted at a controlled temperature of 25 °C.

Peptide identity was confirmed by high-accuracy (10 ppm) mass measurements (both MS and MS/MS) using the Manitoba/Sciex prototype MALDI quadrupole/TOF (time-of-flight, QqTOF) mass spectrometer [14]. Peptide samples were mixed 1:1 with 2,5-dihydroxybenzoic acid MALDI matrix solution (150 mg/ml in 1:1 water:acetonitrile), deposited on a stainless steel target, and air dried prior to MALDI acquisition.

2.3. Chromatographic conditions

Both gradient and isocratic experiments were performed using a Luna C18(2) 100 Å, 5 μm (Phenomenex, Torrance, CA), 100 mm × 1 mm column, at 150 μl/min flow rate and binary solvent setting with both eluent A (water) and B (acetonitrile) containing 0.1% trifluoroacetic acid (TFA). Gradient conditions of 1% acetonitrile per minute starting from 0% were applied to determine the hydrophobicity of the synthetic peptides in hydrophobicity index (HI) units [15].

Isocratic measurements were performed using programmable mixing of solvents A and B. The accuracy of these measurements depends strongly on the proportioning accuracy of pumps A and B in a binary eluent system. We verified the accuracy by preparing an exact v/v water/acetonitrile mixture (with 0.1% TFA) in one eluent bottle, and using it as an 100% eluent in isocratic mode. Three different eluent compositions were tested in this mode (10, 20 and 30% acetonitrile for P2, P4 and P6, respectively). All of them showed virtually identical retention values to those obtained when isocratic conditions were created using programmable proportional mixing. Obtaining identical retention for these three concentrations supported the assumption that proportioning works correctly for the whole range of acetonitrile content studied.

2.4. Sample preparation

Stock solutions of peptides (~1 mg/ml) were prepared by dissolving each peptide in 1 ml of 0.1% TFA in water or a 20% acetonitrile solution. Ten microliters of sample was injected in

Table 1
Synthetic peptides used for the model development.

Peptide series	Charge, *Z*	Length, *N*	Internal index number	Sequence	Molecular weight (Da)	Calculated hydrophobicity index, *HI*[a]	Retention time at 1% ACN/min (min)	Measured hydrophobicity index, *HI*[b]	Slope
Standard peptides	+2	11	P1	LGGGGGGDGSR[c]	888.41	5.05	11.99	4.01	23.6
		10	P2	LGGGGGGDFR	891.42	13.47	18.67	10.76	18.8
		9	P3	LLGGGGDFR	890.46	15.70	23.30	15.44	18.1
		8	P4	LLLGGDFR	889.50	20.49	28.45	20.70	16.8
		7	P5	LLLLDFR	888.54	26.30	33.81	26.11	14.7
		8	P6	LLLLLDFR	1001.63	28.09	37.29	29.59	14.6
	+1	8		ALILTLVS	828.53	27.14	33.02	25.00	13.5
#1	+2	8	1-1	LASAADFR	849.42	13.13	21.19	13.34	17.5
		8	1-2	LLSAADFR	891.47	16.48	26.17	18.34	17.4
		8	1-3	LLSLADFR	933.52	22.24	31.41	23.61	15.3
		8	1-4	LLSLLDFR	975.57	26.21	34.57	26.78	14.9
#2	+3	8	2-1	LASAAHFR	871.47	12.24	19.72	11.81	20.0
		8	2-2	LLSAAHFR	913.52	15.48	23.73	15.89	19.9
		8	2-3	LLSLAHFR	955.56	20.47	26.45	18.66	18.3
		8	2-4	LLSLLHFR	997.61	24.14	32.11	24.44	16.8
#3	+4	8	3-1	LAHAAHFR	921.56	11.15	19.04	11.11	21.1
		8	3-2	LLHAAHFR	963.54	14.38	22.42	14.56	21.2
		8	3-3	LLHLAHFR	1005.58	18.76	27.00	19.23	20.1
		8	3-4	LLHLLHFR	1047.62	22.14	30.45	22.74	19.8
#4	+1	8	4-1	LASAADFG	750.36	14.06	21.63	13.75	14.8
		8	4-2	LLSAADFG	792.40	17.78	27.38	19.61	14.7
		8	4-3	LLSLADFG	834.45	24.86	32.87	25.21	13.8
		8	4-4	LLSLLDFG	876.49	32.23	36.68	29.10	12.2
#5		8	5-1	LAVAAHFR	883.50	14.57	23.00	15.15	19.6
	+3	8	5-2	LLVAAHFR	925.53	17.98	25.53	17.73	18.5
		8	5-3	LLVLAHFR	967.59	22.02	29.53	21.80	17.5
		8	5-4	LLVLLHFR	1009.64	25.30	32.88	25.22	16.5
#6	+2	11	6-1	LAGSASADAFR	1064.52	14.35	20.69	13.27	19.5
		11	6-2	LLGSLSLDAFR	1190.66	24.88	32.48	24.55	17.4
#7	+2	14	7-1	LAGGSASSADAAFR	1279.62	15.69	20.90	13.47	21.2
		14	7-2	LLGGSLSSLDAAFR	1405.74	25.18	32.76	24.82	19.4
#8	+2	17	8-1	LAGGGSASSSADAAAFR	1494.71	14.22	19.41	12.04	22.4
		17	8-2	LLGGGSLSSSLDAAAFR	1620.84	23.34	30.20	22.37	20.9
#9	+3	11	9-1	LAGSASAHAFR	1086.56	13.11	19.41	12.04	21.3
		11	9-2	LLGSLSLHAFR	1212.68	23.30	30.27	22.44	20.9
#10	+3	14	10-1	LAGGSASSAHAAFR	1301.64	13.64	19.61	12.23	22.2
		14	10-2	LLGGSLSSLHAAFR	1427.77	22.25	30.18	22.36	23.3
#11	+3	17	11-1	LAGGGSASSSAHAAAFR	1516.74	12.68	19.84	12.46	23.5
		17	11-2	LLGGGSLSSSLHAAAFR	1642.87	21.70	30.63	22.78	25.2

[a] $HI = H \times 0.5835 + 3.1551$, where H is the peptide hydrophobicity calculated using 100A-TFA version of SSRCalc (http://hs2.proteome.ca/SSRCalc/SSRCalc33B.html).
[b] *HI* values for P1–P6 peptides were measured under isocratic conditions as shown in Fig. 1a and for the rest of peptides using retention time measurements with 1% per minute gradient separations.
[c] P1 peptide was not considered for the model development because of low *HI* value.

both isocratic and gradient elution modes. Individual peptides were diluted to provide ~0.5–1 μg injection of each component.

2.5. Measurements, calculations and model development

The dead volume of the column and connecting tubings was measured using the injection of a non-retained compound (water) and measuring the elution time of the negative peak. Retention factors for isocratic elution were calculated using the formula: $k = (t_R - t_0)/t_{0c}$; where t_R is the retention time, t_0 is the system (column and tubings) dead time, and t_{0c} is the column dead time.

S-Values for both the model and test peptides were determined as the average of three slopes measured from $\log k$ vs. ϕ plots in three independent measurements. Each experimental slope was generated using 4–6 data points (ϕ values). Retention time measurements for each data point were also done in triplicate; overall, ~2000 isocratic chromatographic separations were carried out in the course of this data collection.

A predictive model was developed using a simple "random walk" through parameter-space to find a suitable (but not necessarily the only, or even the best) fit against the "training" dataset of 37 custom synthesized peptides. The code was written in less than 300 lines of Perl on a Mac Pro computer running the OS-X variant of UNIX. We suggest a very general function of the form: $S = C1 \times Z^{C2} + C3 \times N^{C4} + C5 \times HI^{C6} + C7/Z + C8/N + C9/HI + C10 \times ZN + C11 \times ZHI + C12 \times NHI + B$; where for each peptide the values of *HI* (measured hydrophobicity index), *Z* (peptide charge), *N* (peptide length) and an observed *S* (slope) were known.

3. Results and discussion

3.1. The choice of the parameters of the model and its variation range

As was shown in the introduction, determining the properties of peptides responsible for variation the slopes in the basic LSS equation is a daunting task. In our own research we arrived at basically the same conclusion. Recently we studied the set of structurally related peptides of virtually identical 888–891 Da molecular weights, which cover wide range of hydrophobicities [15]. The goal of that study was to develop a calibrating mixture of peptides that can be used in different RP-HPLC modes to plot retention time vs. hydrophobicity dependencies for more accurate retention prediction and data alignment. In these measurements, the molecular weight was kept almost constant by substituting -Gly-Gly- into -Leu-, or -Gly-Ser- into -Phe-, resulting in an increase in hydrophobicity accompanied by a decrease in peptide

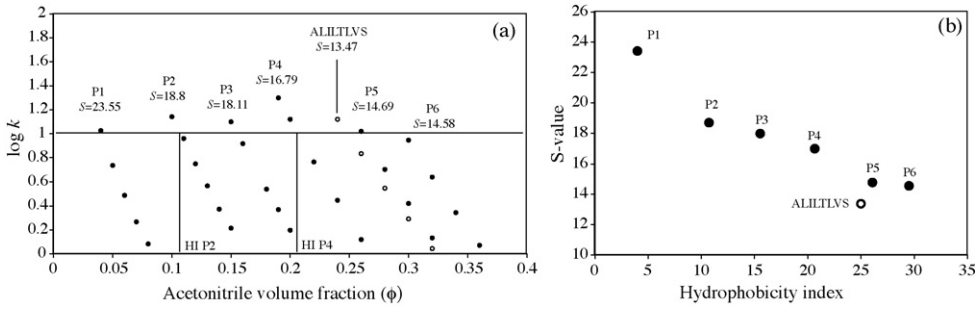

Fig. 1. Determination of *S*-slope and hydrophobicity index values for the set of test P1–P6 peptides. (a) log *k* vs. ϕ plots for ALILTLVS and P1–P6 peptides. Slope values and procedure for assigning *HI* for P2 and P4 are shown. (b) *S* vs. *HI* plots for these species. Peptide sequences and characteristics are given in Table 1.

length (LGGGGGGDGSR, LGGGGGGDFR, LLGGGGDFR, LLLGGDFR, LLLLDFR). A systematic decrease in *S* values was observed for these peptides (peptides P1–P5 in Table 1, Fig. 1). This finding contradicts the original assumption that the increase in *S* is related to *MW*. It also does not agree with the conclusion of Hearn and Aguilar [10], who suggested that a larger hydrophobic contact area and a number of contact sites would leads to an increase in *S*. Instead, the higher *S*-values for hydrophilic peptides could be explained by an increase in the number of residues. These observations led us to use the number of amino acids as a parameter for the model, instead of using the molecular weight. Nevertheless, there is a strong correlation between the number of amino acids in a peptide chain and the molecular weight. Looked at from this perspective, a correlation between the *MW* and *S* for a random set of peptides should be observed, supporting the findings of Snyder and co-workers [4]. We note that alternative approaches to link *S* and *MW* have been proposed by Sakamoto et al. [16], and more recently by Gilar and Neue [17]: $S = a \times \ln(MW) - B$ and $\log(S) = a \times \log(MW) + B$, respectively.

A similar measurement was performed on the peptide ALILTLVS ($N = 8$); this made the overall picture even more complicated (Fig. 1). It has almost identical affinity to the C18 phase as does LLLLDFR ($N = 7$), which has a larger number of the residues but lower molecular weight (by ~60 Da). The measured slope of 13.47 is lower than the value of 14.69 for the P5-peptide. This decrease could be explained as: (i) lower *MW* – this agrees with the finding of Snyder and co-workers, but contradicts our measurements on the P1–P5 peptides; (ii) increasing *N*, which is opposite to the behavior of the P1–P5 peptides. The only evident molecular descriptor that could account for this discrepancy is the charge of the peptide. This led us to include charge as a model variable in addition to the number of amino acids and peptide hydrophobicity.

Due to the extreme diversity of peptidic compounds in terms of their physical properties, it is hard to approach the development of a comprehensive predictive model. The studies performed earlier [8–11] concentrated on sets of related peptides, often representing the sequence of some biologically active compounds. This, is our opinion, prevented the development of a more accurate understanding of variations in *S*-values. Thus, our study was designed to cover a typical group of analytes for a tryptic digestion – the most popular enzymatic protocol in proteomics. Consequently we have targeted particular ranges of parameter variations in designing the sequences of test peptides, rather than targeting particular sets of related peptides. To further illustrate this, we have used the data set that we collected for the optimization of our Sequence Specific Retention Calculator (SSRCalc) peptide retention prediction model, as a typical example of the sequences identified in proteomic analyses [13]. Fig. 2 shows the frequency distribution within this set of ~5000 peptides depending on peptide charge, length and hydrophobicity. Based on Fig. 2a, one can conclude that the majority of tryptic peptides in this set are carrying a +2 charge at the acidic eluent conditions used (0.1% TFA). The two charged groups include the N-terminal amino group and the side chains of the C-terminal Lys or Agr residues. Triply and quadruply charged peptides–the second and third most prevalent case–feature one and two internal basic amino acid (His, Lys, Arg), respectively.

A single positive charge is characteristic of peptides representing the C-terminal sequence of proteins with no internal basic residues. Therefore, the test set of peptides was designed to cover the +1 to +4 charge range, with an emphasis on doubly and triply charged species.

The distribution of peptide length strongly depends on the type of mass spectrometer and the identification procedures used. Automated MS/MS analyses rarely produce confident identification for peptides shorter than 7–8 residues. Very short peptides (2–5 residues long) are rarely observed also, due to the higher level of background noise for both ESI and MALDI techniques. As peptide length increases, the probability of occurrence goes down, as shown in Fig. 2b. We selected $N = 8$ as the shortest peptide included in the study, as it corresponds to the maximum in the distribution. This group was the most abundant, giving 20 species plus peptides with 3-mer increment increase: 11, 14 and 17 residues (4 of each, as shown in Table 1). The five peptides of various sizes from P1–P6 set were also added to the training set, giving a total of 37 sequences investigated.

Fig. 2c shows the distribution of calculated peptide hydrophobicities for our 5000 peptide optimization set; we used our own SSRCalc algorithm to perform these calculations. SSRCalc is one of the most accurate models developed to date and typically provides ~0.98 R^2-value correlation for retention time vs. hydrophobicity plots for the tryptic peptides [13]. Recently we proposed to use a peptide hydrophobicity index (*HI*) in an acetonitrile percentage scale to express calculated hydrophobicity [15]. As in early work by Valko and Slegel [18] we define the *HI* value as the acetonitrile percentage at which a particular peptide has a retention factor equal 10 when eluted under isocratic conditions. These values were carefully measured for P1–P6 peptides, as depicted in Fig. 1a, and mapped onto our 5000 peptide optimization data set. The measurements expressed in *HI* provide a very straightforward representation of the gradient RP-HPLC process; Fig. 2c shows that all peptides in our data set are eluting from a C18 100 Å column between 0 and 35 acetonitrile percentage, when TFA conditions are used. In this distribution the maximum at 15–17 SSRCalc *HI* units is due to the lower abundance of very small (hydropholic) and very big (hydrophobic) peptides in the peptide data set. We chose to work with an *HI* range of ~10–25% acetonitrile, corresponding to a hydrophobicity of P2–P5 in our original 6-peptide set: it covers ~75% of tryptic peptides typically observed in RP-HPLC/MS experiments. Hydrophilic analytes ($HI < 10$) were excluded, as anomalously high *S*-values are often observed under these conditions [19]. We also found the same significant deviation for the peptide P1 (Fig. 1). Table 1 shows both SSRCalc predicted and gradient condition measured *HI* values for the 37 test pep-

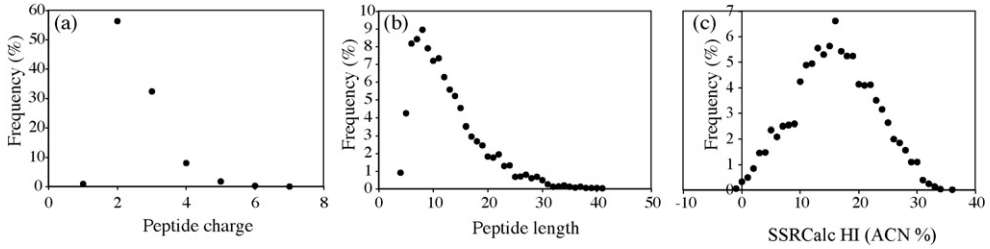

Fig. 2. Frequency distribution of basic peptides properties for a typical set of tryptic peptides obtained in proteomic experiments [13]: (a) peptide charge, (b) length, (c) hydrophobicity index.

tides; the experimentally measured *HI* values were used for model development. However, the calculated values could have been successfully utilized as well: they show a ~0.973 R^2-correlation against the measured values.

3.2. Designing peptide sequences in the test set and measurement of S-values

Synthetic peptides from the sets 1–5 (Table 1) were designed to have the same length, different charge states +1 to +4, and hydrophobicities following those of P2–P5: ~10, ~15, ~20, ~25% acetonitrile. This was achieved through Ala-Leu substitutions, which resulted in consecutive 42 Da mass increases. Peptides in set 6-11 were designed to probe the influence of peptide length for the species having +2 and +3 charges and border hydrophobicity values of ~10 and ~25*HI*. Constructing the set of test peptides in this way allowed us to study the effect of any descriptor under investigation while holding the two other variables constant. For example, the influence of peptide charge can be followed for the 8-mer most hydrophilic members of the sets (*HI* ~ 10) when the slopes for 1-1, 2-1, 3-1, 4-1, 5-1 are compared. Similarly, the effect of peptide length can be examined for the doubly charged most hydrophobic analytes (*HI* ~ 25) by comparing the $\log k$ vs. ϕ dependencies of 1-4, 6-2, 7-2 and 8-2. This approach allows us to assess the influence of one particular physical property of the peptide on the slope *S*, and to compare these findings against prior literature, as well as to relate it to the modeling of the peptide RP-HPLC mechanism.

The slopes *S* in the basic LSS equation can be measured in both gradient and isocratic elution modes. The calculations allowing us to convert gradient retention data obtained using different gradient slopes/flow rates into isocratic $\log k$ vs. ϕ dependencies have been described elsewhere [4,8], and have been successfully used in previous studies. The reported R^2-value correlations for these linear dependencies were found to be typically in the range of 0.95–0.98 [8]; in our opinion, such variations leave some uncertainty in assigning *S*-values. Our isocratic measurements typically resulted in correlations of 0.996 and higher. Despite being more labor-intensive compared to gradient measurements, we opted to use isocratic measurements. However, the choice of isocratic measurements instead of gradient ones precluded any testing of the model using real tryptic digests, as the presence of multiple components in the mixture requires gradient separation. Therefore, to test the model we used data in the literature on the chromatographic behavior of tryptic peptides from human growth hormone (see Section 3.5), and some synthetic species from that mixture.

3.3. The influence of peptide hydrophobicity, charge and length

Preliminary results for ALILTLVS and the six standard peptides P1–P6 described above suggested that *S*-values depend on peptide length and charge. To look at the influence of peptide length in greater detail, plots of *S* vs. *N* were constructed for the most hydrophilic and hydrophobic members of the relevant peptide sets, as shown in Fig. 3a and b. For example, to monitor the influence of length, we use the first members of the doubly charged sets of peptides in Fig. 3a: analytes 1-1, 6-1, 7-1 and 8-1 (Table 1) were considered. Similarly for triply charged we use 2-1, 9-1, 10-1, 11-1, etc. Generally, linear dependencies with R^2-value correlations ~0.99 were observed, while *S*-values grow faster for more hydrophobic species in Fig. 3b. Peptide elongation was achieved by uniform insertion of the residues that typically do not contribute significantly to peptide RP retention – Gly, Ser and Ala. Their role in the variation of *S* was to provide spacing between the hydrophobic residues Leu or Phe, resulting in a larger hydrophobic contact area. This influence becomes more prominent when peptide hydrophobicity increases. Overall, the rate of increase in S with peptide elongation partially depends on the hydrophobicity of the peptide as well as its charge.

An increase in peptide charge leads invariably to higher *S*-values for peptides of the same length and hydrophobicity. Fig. 3c and d shows this effect for the most hydrophilic and hydrophobic members of the sets. The sets of 8-mer peptides were designed to carry from 1 to 4 charges (Table 1), while the longer peptides have only charges +2 and +3. The respective dependences for the first members of 8-mer sets include peptides 4-1, 1-1, 2-1, 5-1 (both +3) and 3-1, while 11-, 14- and 17-mers all have only two data points: 6-1, 9-1 (11-mer); 7-1, 10-1 (14-mer); 8-1, 11-1 (17-mer). Linear correlations with R^2-value of 0.99 and better were observed for all four subsets of 8-mer species: addition of each charge increases *S* approximately by 2 (similar dependencies for the peptides of intermediate *HI* and $N = 8$ are not shown). This demonstrates the importance of the contribution of charged residues to peptide retention in RP-HPLC. Despite being hydrophilic in nature, they provide an additional means of peptide–sorbent interaction, and increase the number of contact sites. This effect is more pronounced when it occurs in long hydrophobic peptides. Thus, the highest *S*-value of all 37 species was observed for the triply charged 17-mer peptide 11-2 with $HI = 21.6$ (Fig. 3b and d).

It has been suggested that peptide hydrophobicity is one of the major parameters determining *S*-slopes. However, no simple correlation has been found [8–10]. Analyzing the measurements for equally sized/equally charged subsets of the model peptides in this study helped us to draw more detailed conclusions. We have confirmed the absence of a direct connection between *S* and *HI*: slopes can decrease and increase with hydrophobicity depending on peptide charge and size. Fig. 4a shows a decrease in *S* values for a number of 8-mer species of different charges when the hydrophobicity increases in the ~10 to ~25*HI* range studied. All four peptides from each subset in Table 1 were used to generate these plots. Thus, the *S*-values for peptides 4-1, 4-2, 4-3 and 4-4 were utilized to build the 8-mer singly charged species graph. While all *S* vs. *HI* dependencies plateau at 10–15*HI*, the slopes decrease faster at higher peptide hydrophobicity. The same trend was observed when doubly charged peptides of different sizes were considered (Fig. 3b). The sets of longer peptides (11, 14 and 17 residues) contained only two species, however it is possible to find a trend showing

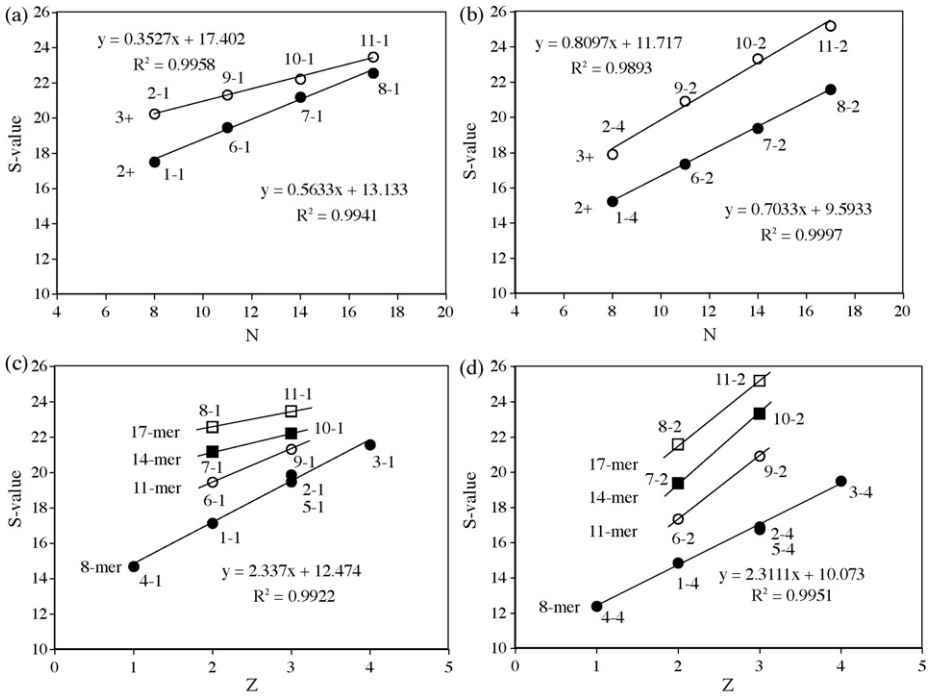

Fig. 3. Effect of peptide size and charge on *S*-value. (a and b) *S* vs. *N* for the first and last members of respective sets, most hydrophilic and hydrophobic, respectively. Peptides are grouped based on charge; (c and d) *S* vs. *Z* for the first and last members of the sets, most hydrophilic and hydrophobic, respectively. Peptides are grouped based on length.

a decrease in *S* similar to the 8-residue ones in Fig. 4a. The picture changes significantly for triply charged peptides of different length: *S* grows for 14- and 17-mer species and remains almost constant for 11-mer ones when hydrophobicity increases. Although this conclusion was drawn using only two data points, we believe it reflects the general trend of having the largest *S*-values for long, highly charged analytes.

3.4. Modeling behavior of S-values in 3-parameter-space

Understanding the rules of variation of *S*-values and developing an approach for its quantitative estimation is extremely important for applied RP-HPLC of peptides. They will help to estimate expected selectivity variations when chromatographic conditions are optimized for better resolution of components of the mixture, or retention prediction models developed for one gradient conditions (or column size) applied to another one. There are at least two distinct ways to achieve this once molecular descriptors for the model have been determined. The first approach assumes the synthesis and measurement of *S*-values for an extended set of synthetic peptides that cover a wide range of molecular descriptor variation. Following such measurements, lookup tables can be constructed and used when the *S*-value for a particular sequence needs to be estimated. While peptide charge and length have a finite set of integral values, peptide hydrophobicity changes continuously. Depending on the chosen frequency of the data points covering the hydrophobicity scale (4 or 2 points in our case) it may require the synthesis and experimental measurements for a set of hundreds of peptides. A second approach assumes the measurement for a number of selected peptide species and creates a model to describe the experimental data. We chose the latter scenario due to the extreme time and resources that would be consumed by the first approach.

We assume a general equation with twelve coefficients to describe the variation of *S*: $S = C1 \times Z^{C2} + C3 \times N^{C4} + C5 \times HI^{C6} + C7/Z + C8/N + C9/HI + C10 \times ZN + C11 \times ZHI + C12 \times NHI + B$.

Our optimization program executes a simple "random walk" of all coefficient values through a parameter-space, with the volume of this space decreasing from 10 to 0.0125 units through successive

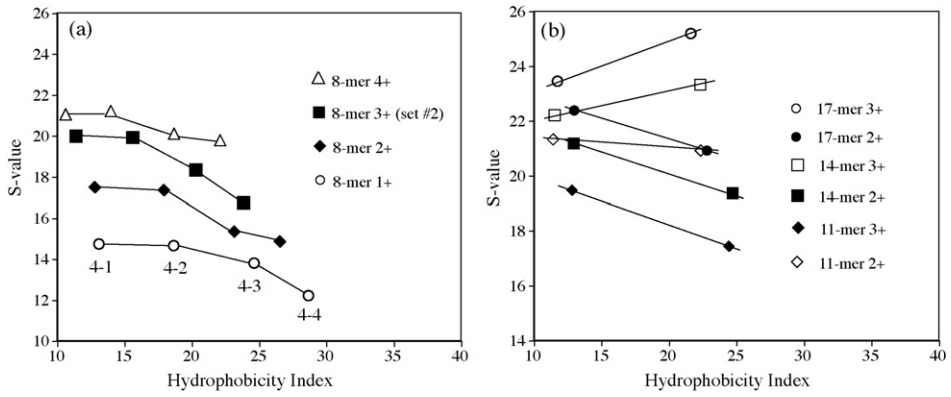

Fig. 4. Effect of peptide's hydrophobicity on its *S*-value. (a) 8-mer peptides of various charges; (b) 11-, 14- and 17-mer doubly- and triply charged peptides.

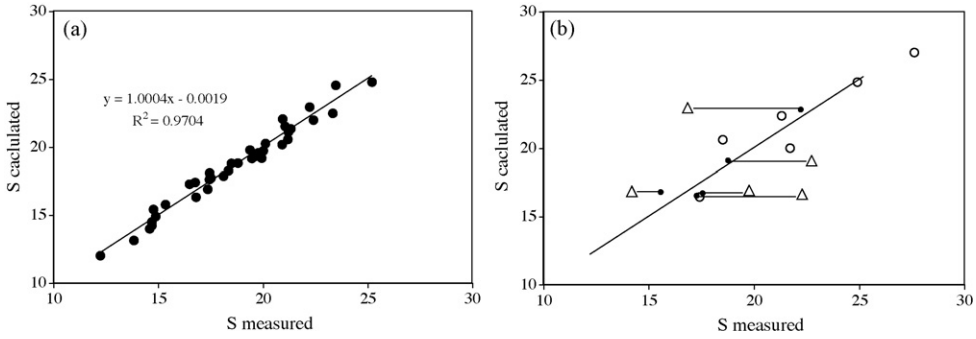

Fig. 5. Accuracy of predictive model for (a) test set of 37 peptides and (b) tryptic peptides from human growth hormone molecule: (△ and ○) measured values from [19] and (●) measured values for synthetic analogues.

cycles in several "batch runs" of 50,000–200,000 cycles each, in which the previous batch's best values are used as a starting point for the current batch. The measurement of goodness-of-fit is a simple R^2 linear regression between the equation and the observed data; this step also yields the intercept constant B. The first batch run holds parameters $C7$–$C12$ to zero and typically gives an R^2 value of ~0.92 after 50,000 cycles, indicating the major independent influence of each variable. The second batch permutes through all 12 parameters and converges to $R^2 \sim 0.95$. The program ultimately permutes through ~800,000 cycles in the course of converging to a final solution, with a correlation of R^2-value ~0.97 for experimental vs. calculated S-values: $S = 1.6824Z^{0.7482} + 2.4747N^{0.0707} - 0.0481HI^{1.5505} - 0.7352/Z - 0.2636/N - 0.2975/HI + 0.0887Z \times N + 0.0022Z \times HI + 0.0215N \times HI + 011.2904$ (Fig. 5a).

Since dependence of S from peptide hydrophobicity is very complex it was of interest to model the behavior of the slopes depending on two parameters only: charge and peptide length. A similar model optimization procedure resulted in only 0.89 R^2-value correlation: $S = 163.0468Z^{-1.1195} + 73.2917N^{0.0748} - 188.0112/Z - 21.1523/N + 0.0016ZN - 51.7578$. This shows that all three parameters should be taken into account when building an accurate model for predicting slopes in the basic LSS equation for peptides.

3.5. Testing the model using literature data for tryptic digest of human growth hormone

We have chosen to test the model using literature data for a tryptic digest of human growth hormone. Grego et al. [20] described the peptide identification for this mixture and tested various peptide retention prediction models developed using these data. Later Chloupek et al. [21] and Hancock et al. [19] applied computer simulation to achieve the complete resolution of peptides in this mixture and reported S-values measured for the components. Table 2 shows the peptide sequence, calculated HI, measured retention times and slopes, as well as the S-values calculated according to our proposed model. Correctness of the peak assignment in Table 2 can be confirmed by plotting retention times [21] vs. SSRCalc hydrophobicity of non-modified tryptic peptides. This gave a correlation R^2-value of ~0.97 (not shown here). In total there are 11 non-modified tryptic peptides with $HI > 10$ which we used for the model testing. These species range from 6 to 21 residues in length, +2 to +4 in charge and ~12 to 36 units on the HI scale; they represent a typical set of tryptic peptides. Simple visual analysis of Fig. 5b and the S-values in Table 2 indicates significant deviations from the predicted S-values for five species out of 11 (labeled as triangles). The literature data [19] was obtained using gradient elution conditions at a column temperature of 20 °C, while our results were obtained at 25 °C. Since this difference could be responsible for some deviation in determining the slopes [22], we decided to examine synthetic analogs of the five peptides in question: TGQIFK, SNLQLLR, KDMDKVETFLR, FPTIPLSR, DLEEGIQTLMGR. Table 2 and Fig. 5b (labeled as solid circles) show that the measured values for these species vary significantly from the literature data, but agree well with our predictive model. It is difficult to explain such significant differences by temperature variation or some inconsistency when S-values are measured under gradient conditions. However, the case of SNLQLLR–KDMDKVETFLR is more obvious. The reported S-value of 22.3 for T8 SNLQLLR is very close to the one we predicted (22.9) and measured (22.2) for T17-18-19 KDMDKVETFLR. The reverse also applies: the reported value of 16.8 for T17-18-19 KDMDKVETFLR corresponds to our predicted (16.7) and measured (17.6) slopes for T8-SNLQLLR. This pair of peptides has very close elution times, and change retention order when 30 or 120 min gradients were applied [21]. It seems very likely that the peptide identity was assigned incorrectly when the table representing measured S-values was composed.

Overall, our developed model provides very accurate prediction of S-values for the random set of tryptic peptides tested in this study. It was of interest, however, to monitor species with molecular descriptors outside of our initial parameter-space: those very hydrophobic, or very long. Thus, the ISLLLIQSWLEPVQFLR peptide with an HI value of ~36 showed the largest deviation: 18.5 in literature data vs. 20.6 predicted. The two longest peptides SVFANSLVYGASDSNVYDLLK ($N = 21$) and LHQLAFDTYQEFEEAYIPK ($N = 19$) exhibit very accurate prediction: 24.9 and 27.6 compared to predicted values of 24.9 and 27.0, respectively. Hydrophilic peptides ($HI < 10$) were not included in the test set of 37 model species and in Fig. 5b due to the expected anomalously high S-values. Indeed, calculated values for doubly charged hydrophilic species QTYSK and LEDGSPR are much lower compared to the literature data (Table 2). Another interesting example is provided by the disulfide linked peptides T20-21 ($N = 13$, $Z = 3$, $HI \sim 13$) and T6-16 ($N = 32$, $Z = 4$, $HI \sim 28$). While for the first one the predicted value of 22.4 is close to the literature value of 22.8, the second shows a significant deviation of 41.4 vs. 30.7 in [19]. Most likely this is the result of the T6-16 peptide length being significantly outside the parameter-space used for our model development.

3.6. Overall summary and future development

Experimental data obtained in this study allowed us to critically estimate the correctness of the general conclusions made in the prior literature about the factors affecting the S-values for peptides. We have suggested using peptide charge as one of the numerical parameters that should be taken into account when considering the variation of S. Indeed for all analytes studied, slopes S in the basic LSS equation increase with charge when the peptide length and hydrophobicity remain constant. The involvement of charged functional groups in ion-pairing interactions is consistent with the general statement: the ion-pairing interaction is as important in establishing contacts with the stationary phase as are the

Table 2
Human growth hormone tryptic peptides used for the model verification.

Charge, *Z*	Length, *N*	Index number [18]	Sequence	Calculated hydrophobicity index. *HI*[a]	Retention time for 120 min gradient [21] (min)	Measured slopes [19]	Calculated slopes	Measured slopes using synthetic peptides
+2	5	T14	QTYSK[b]	3.53	12.8	24.3	21.5	
+2	7	T12	LEDGSPR[b]	5.47	18.7	23.7	20.0	
+2	6	T13	TGQIFK	12.40	31.6	14.2	17.0	15.6
+3	13	T15	FDTNSHNDDALLK	13.75	38.2	21.3	22.4	
+2	7	T8	SNLQLLR	16.77	44.6	22.3	16.7	17.3
+4	11	T17-18-19	KDMDKVETFLR	17.67	43.4	16.8	22.9	22.2
+2	8	T2	LFDNAMLR	20.25	47.3	17.4	16.5	
+3	10	T17-18	DMDKVETFLR	18.83	47.8	21.7	20.0	17.6
+2	8	T1	FPTIPLSR	18.92	50.1	19.8	16.7	18.8
+2	12	T11	DLEEGIQTLMGR	19.41	58.1	22.7	18.4	
+3	19	T4	LHQLAFDTYQEFEEAYIPK	27.56	61.9	27.6	27.0	
+2	21	T10	SVFANSLVYGASDSNVYDLLK	27.17	65.1	24.9	24.9	
+2	17	T9	ISLLLIQSWLEPVQFLR	36.02	86.9	18.5	20.6	

[a] Same as in Table 1.
[b] These peptides were not considered for the model testing because of low *HI* value.

hydrophobic contact sites. This finding is consistent with a recent report by Wang and Carr [23], showing the significant influence of ion-pairing on slopes *S* for basic drugs and peptides.

Increasing the peptide length also led invariably to an increase in *S*. We achieved this length variation through the uniform insertion of "hydrophobically neutral" residues into the peptide sequence. We supposed that this increases the hydrophobic contact area while keeping the number of hydrophobic contact sites constant, giving an increase in *S*. Simultaneously we found that peptide length (rather than molecular weight) is the parameter that should be taken into account when building a predictive model for *S*-values. There is always a close-to-linear dependence between S and the number of amino acids for peptides of same charge and hydrophobicity; it is conceivable that *N* and *MW* have a very strong correlation. However, in some cases, the use of molecular weight might lead to discrepancies in assigning *S*-values for peptides, as was pointed out in the prior literature [8–10]. A specific feature of tryptic species is the presence of positively charged groups on both termini; increasing the peptide length leads to a further physical separation of the charged groups. Since a strong influence of peptide charge on *S*-values was found, it will be interesting to understand the behavior of the slopes when adjacent residues are carrying positive charge.

Finally, peptide hydrophobicity impacts differently on *S*, confirming the previous finding of the absence of a direct link between the sum of hydrophobic coefficients of the constituent amino acids and the slope in the basic LSS equation [8]. Increased hydrophobicity led to lower *S* for short peptides, and higher *S* values for long highly charged (+3) species. It is unclear how a larger number of very hydrophobic residues can result in a decrease in the hydrophobic contact area. Most likely there are significant changes in the separation mechanism that occur when a number of hydrophobic residues located close to each other form a "hydrophobic cluster". Overall, the effect observed for short peptides can be described from the point of view of the co-existence of two separation mechanisms: hydrophobic and ion-pairing interactions. The dominance of the former leads to lower *S* (similar to low molecular weight organic compounds in RP-HPLC); the latter leads to higher *S*. But this approach does not hold up when dealing with larger species: increased hydrophobicity for long peptides leads to higher slopes.

It is interesting to note that an increase in all three molecular descriptors: charge, length and hydrophobicity (for long, highly charged peptides) will cause a rise of *S*-values. Meanwhile, all three parameters generally do increase with molecular weight. When Snyder and co-workers [4] studied a set of peptide sequences within the 600–14,000 Da mass range, some correlation between *S* and *MW* was found. When a narrower range of molecular weights (typical for tryptic digests) was considered this correlation was not always supported by experimental data. Narrowing the *MW* range results in small changes in peptide structure resulting in non-concurrent effects of molecular descriptors on *S*. Even within a group of typical tryptic species there are specific subgroups, which do not follow general rules. Thus, as we pointed out, there is a "grey" area of hydrophilic species that exhibit extremely high *S*-values. This group of analytes deserves a directed study in the future. Another intriguing subset of peptides for further studies are the ones carrying sequences prone to stabilization of amphiphatic helical structures upon the interaction with C18 surface. This is expected to change dramatically the hydrophobic contact area and the respective *S*-values.

4. Conclusions

A continuing study of peptide separation selectivity aimed at the development of peptide retention prediction models in RP-HPLC has demonstrated the particular importance of the charge of the separated compounds. Similar logic has suggested the importance of this parameter in determining the slopes (*S*) in the fundamental equation of linear-solvent-strength theory for peptidic compounds. Our original assumption was that three well-defined molecular descriptors: peptide length, charge and hydrophobicity index are the major contributors to the value of *S*. It was supported through the measurements of the slopes for a set of synthetic peptides designed specifically to study these effects. The optimized model gives a ~0.97 R^2 correlation between measured and predicted *S*-values using the formula: $S = 1.6824Z^{0.7482} + 2.4747N^{0.0707} - 0.0481HI^{1.5505} - 0.7352/Z - 0.2636/N - 0.2975/HI + 0.0887ZN + 0.0022ZHI + 0.0215NHI + 011.2904$. The accuracy of the model was also tested using previously published data on a human growth hormone protein tryptic digest and some synthetic analogues from that mixture. To our knowledge this is the only numerical model that predicts *S*-slopes for typical tryptic peptides monitored in proteomics experiments, i.e. species falling into particular range of molecular weight, charge and hydrophobicity.

Acknowledgements

The authors want to thank Dr. K.G. Standing for his help during the manuscript preparation. This work was supported in part by grants from the Technology Transfer Office at the University of Manitoba and the Natural Sciences and Engineering Research Council of Canada (O.V.K.).

References

[1] M.T.W. Hearn (Ed.), HPLC of Proteins, Peptides and Polynucleotides. Contemporary Topics and Applications, Wiley, New York, 1991.
[2] L.R. Snyder, J.L. Glajch, J.J. Kirkland, Practical HPLC Method Development, Wiley, New York, 1997.
[3] C.T. Mant, R.S. Hodges, HPLC of Biological Macromolecules, Marcel Dekker, New York, 2002.
[4] M.A. Stadalius, H.S. Gold, L.R. Snyder, J. Chromatogr. 296 (1984) 31.
[5] L.R. Snyder, J.W. Dolan, High-Performance Gradient Elution: The Practical Application of the Linear-Solvent-Strength Model, Wiley, New York, 2006.
[6] J.L. Glaich, M.A. Quarry, J.F. Vasta, L.R. Snyder, Anal. Chem. 58 (1986) 280.
[7] V. Spicer, A. Yamchuk, J. Cortens, S. Sousa, W. Ens, K.G. Standing, J.A. Wilkins, O.V. Krokhin, Anal. Chem. 79 (2007) 8762.
[8] M.I. Aguilar, A.N. Hodder, M.T.W. Hearn, J. Chromatogr. 327 (1985) 115.
[9] M.T.W. Hearn, M.I. Aguilar, J. Chromatogr. 359 (1986) 31.
[10] M.T.W. Hearn, M.I. Aguilar, J. Chromatogr. 392 (1987) 33.
[11] M.T.W. Hearn, M.I. Aguilar, C.T. Mant, R.S. Hodges, J. Chromatogr. 438 (1988) 197.
[12] O.V. Krokhin, R. Craig, V. Spicer, W.E. Ens, K.G. Standing, R.C. Beavis, J.A. Wilkins, Mol. Cell Proteomics 3 (2004) 908.
[13] O.V. Krokhin, Anal. Chem. 78 (2006) 7785.
[14] A.V. Loboda, A.N. Krutchinsky, M. Bromirski, W. Ens, K.G. Standing, Rapid Commun. Mass Spectrom. 14 (2000) 1047.
[15] O.V. Krokhin, V. Spicer, Anal. Chem 81 (2009) 9522.
[16] Y. Sakamoto, N. Kawakami, T. Sasagawa, J. Chromatogr. 442 (1988) 69.
[17] M. Gilar, U.D. Neue, J. Chromatogr. A 1169 (2007) 139.
[18] K. Valko, P. Slegel, J. Chromatogr. 631 (1993) 49.
[19] W.S. Hancock, R.C. Chloupek, J.J. Kirkland, L.R. Snyder, J. Chromatogr. A 686 (1994) 31.
[20] B. Grego, F. Lambrou, M.T.W. Hearn, J. Chromatogr. 266 (1983) 89.
[21] R.C. Chloupek, W.S. Hancock, L.R. Snyder, J. Chromatogr. 594 (1992) 65.
[22] A.W. Purcell, G.L. Zhao, M.I. Aguilar, M.T.W. Hearn, J. Chromatogr. A 852 (1999) 43.
[23] X. Wang, P.W. Carr, J. Chromatogr. A 1154 (2007) 165.