Journal of Chromatography A, 1218 (2011) 6348–6355

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Defining intrinsic hydrophobicity of amino acids' side chains in random coil conformation. Reversed-phase liquid chromatography of designed synthetic peptides vs. random peptide data sets

Dmitry Shamshurin [a], Vic Spicer [b], Oleg V. Krokhin [a,c,*]

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

ARTICLE INFO

Article history:
Received 30 March 2011
Received in revised form 21 June 2011
Accepted 27 June 2011
Available online 3 July 2011

Keywords:
Peptide reversed-phase HPLC
Amino acids' hydrophobicity
Peptide retention prediction

ABSTRACT

The two leading RP-HPLC approaches for deriving hydrophobicity values of amino acids utilize either sets of designed synthetic peptides or extended random datasets often extracted from proteomics experiments. We find that the best examples of these two methods provide virtually identical results – with exception of Lys, Arg, and His. The intrinsic hydrophobicity values of the remaining residues as determined by Kovacs et al. (Biopolymers 84 (2006) 283) correlates with an R^2-value of 0.995+ against amino acid retention coefficients from our Sequence Specific Retention Calculator model (Anal. Chem. 78 (2006) 7785). This novel finding lays the foundation for establishing consensus amino acids hydrophobicity scales as determined by RP-HPLC. Simultaneously, we find the assignment of hydrophobicity values for charged residues (Lys, Arg and His at pH 2) is ambiguous; their retention contribution is strongly affected by the overall peptide hydrophobicity. The unique behavior of the basic residues is related to the dualistic character of the RP peptide retention mechanism, where both hydrophobic and ion-pairing interactions are involved. We envision the introduction of "sliding" hydrophobicity scales for charged residues as a new element in peptide retention prediction models. We also show that when using a simple additive retention prediction model, the "correct" coefficient value optimization (0.98+ correlation against values determined by synthetic peptide approach) requires a training set of at least 100 randomly selected peptides.

1. Introduction

Reversed phase HPLC of peptides has long been recognized as a potent method to determine the hydrophobicity of amino acids [1,2], a crucial parameter in the studies of protein structure and interactions. The hydrophobic interactions in peptide RP separation closely mimic the interaction of peptides and proteins with hydrophobic substrates in biological systems. Some of these chromatographic studies were also directed at improving the prediction of peptide retention in RP-HPLC systems, and have yielded a number of hydrophobicity scales.

Recently Mant et al. [3] reviewed a large collection of RP-HPLC methods to define the hydrophobicity of amino acid side chains. They concluded the hydrophobicities (retention coefficients) derived using designed synthetic peptides correlated poorly against values from models built using random peptide collections. A set of twenty synthetic peptides of the sequence motif *Ac-XGAKGAGVGL-Amide* (where X is the target amino acid residue) was designed to eliminate any secondary structure and nearest-neighbor effects on the intrinsic hydrophobicity of the substituted residue in a random coil conformation [4]. These peptides were separated under reversed-phase conditions and their observed retention values were used to derive the intrinsic hydrophobicity values of the residues. In a RP-HPLC approach based on the separation of random peptide collections, retention coefficients values are determined from the optimization of models that correlate observed and predicted retention times. The phenomenological difference between these two methods is obvious: while for the designed peptides the observed retention of a single compound is used to determine hydrophobicity of each residue, random dataset studies rely on multiple peptides containing the same amino acid to extract its value. The designed peptide approach relies on the elimination of all non-hydrophobic interactions through the careful design of the sequence motif, while random peptide methods eliminate secondary structure effects through the "smoothing" effect of larger datasets. Gilar et al. demonstrated one of the best examples

* Corresponding author at: Manitoba Centre for Proteomics and Systems Biology, 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 © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2011.06.092

of the random peptide collection method recently, using a set of 165 tryptic peptides [5]. They showed optimizations of an additive retention prediction model using retention data sets collected for various column temperatures, concentrations of ion-pairing agent, and types of sorbents, proving that this can reveal the differences in retention coefficients responsible for altering separation selectivity. Over the years we have developed a family of Sequence Specific Retention Calculator (SSRCalc) models based on extensive proteomics-derived datasets collected under various chromatographic conditions [6–8]. Creating accurate retention prediction models was the major goal of these studies, but the values of the retention coefficients we obtained were never the subject of in-depth comparison against other hydrophobicity scales.

The examples of the random peptide collection approach chosen for comparison by Mant et al. [3] were published between 1981 and 1995. Since then, advancements in mass spectrometry and proteomics applications have renewed the interest in developing RP-HPLC peptide retention prediction algorithms by presenting researchers with systems to collect large, confident peptide datasets. A number of peptide retention prediction models were developed in the past decade, providing significant advancements in the field [5–13]. These yielded additional hydrophobicity scales that also merit comparative analysis. In this report we seek to bridge a gap between the two RP-HPLC methods for determining the intrinsic hydrophobicities of amino acid residues: the use of designed synthetic peptides and the use of random peptide dataset derived from proteomics experiments.

2. Materials and methods

2.1. Retention coefficients (intrinsic hydrophobicity) values

All hydrophobicity scales were normalized using the approach as described by Mant et al. [3] with the average of the retention coefficients values scaled to 0.0 and the standard deviation to 1.0. Appendix A shows the entire collection of retention coefficients extracted from the literature data in both raw format and as normalized scales.

2.2. Materials

Deionized (18 MΩ) water and HPLC-grade acetonitrile were used for preparation of the eluents. Trifluoroacetic acid was sourced from Sigma–Aldrich (St. Louis, MO). Four peptides of designed sequences of motif *Ac-XGAKGAGVGL-Amide* (with His, Gly substitution, identical to ones from [4]) and *Ac-XGAKGAGLLL-Amide* (with His, Gly) were synthesized by JPT Peptide technologies GmbH (Berlin, Germany).

2.3. Instrumentation and chromatographic conditions

RP-HPLC analysis used a micro-Agilent 1100 Series system (Agilent Technologies, Wilmington, DE) with a UV detector operated at 214 nm and a manual injector (loop size 10 μl). All chromatographic experiments were conducted at room temperature (22–25 °C). Gradient elution conditions were applied using an in-house packed column (Luna C18(2) 100 Å, 5 μm pore size (Phenomenex, Torrance, CA), 100 mm × 1 mm size) at a 150 μl/min flow rate. Binary solvent settings were used with both eluents A (water) and B (acetonitrile), containing 0.1% trifluoroacetic acid (TFA) and gradient slope of 1% acetonitrile per minute starting from 0% acetonitrile.

Stock solutions of peptides (~1 mg/ml) were prepared by dissolving each peptide in 1 ml of 0.1% TFA in water. Ten microliters of the sample was injected following sample dilution with buffer A to provide ~0.5–1.0 μg of injection of each component.

2.4. Calculations and programming

The additive retention prediction model optimization was performed using a ~5000 peptide retention dataset originally used in the development of our SSRCalc 100 Å TFA algorithm [7]. Subsets of 20–1000 species were randomly selected from this data set. These were used to optimize an additive retention prediction model with a correction for peptide length: $t_R = (1 - a \times \mathrm{Ln}(N)) \times \left(\sum nR_{ci}\right)$; where N, peptide length; R_{ci}, retention coefficients for individual amino acids; n, number of a particular residues in a peptide sequence [5]. An automated multi-pass single parameter optimization of the 20 R_{ci} values and the coefficient a (following our earlier approaches [6]) provided the best correlation t_R experimental vs. t_R predicted, measured as Pearson's correlation coefficient.

This optimization was performed five times for each size of the data set, with fresh peptides being randomly selected for each execution cycle, and the resulting correlation values were averaged across the five repetitions. Programs were written in Perl 5.8.8 and executed on an AMD 955-X4 (Advanced Microdevices, Sunnyvale, CA) based workstation running Yellow Dog Enterprise Linux (Fixstars, Tokyo, Japan).

3. Results and discussion

3.1. The choice of hydrophobicity scales for comparison

Table 1 and Appendix A show the hydrophobicity scales chosen for comparison in this study. It includes seven different scales for peptides in random coil conformation reported by Hodges and co-workers using their synthetic peptide approach [4,14,15]. Kovacs et al. [4] designed synthetic peptides of common Ac-XGAKGAGVGL-Amide composition where position X is substituted with 20 naturally occurring amino acids (Ac-XG- as in Table 1). It was suggested that RP-HPLC measurements for these peptides provides the most realistic picture of intrinsic hydrophobicity of amino-acid side chains due to the elimination of possible secondary structure and nearest-neighbor effects. In addition to this data, we choose to compare the retention coefficients for C-terminal substitutions with free carboxy (-GX-OH) and amide groups (-GX-Amide), N-terminal substitution with free amino group (NH_3-XG-) and the internal substitution within the 11 mer peptides (-GXG-) [14]. We also considered the data presented by the same group in 1986 (Guo et al. [15]), which featured the same TFA-based eluent system and wide pore (300 Å) sorbent.

Comparison of hydrophobicity coefficients obtained by Meek [2], Meek and Rossetti [16], Browne et al. [17] and Wilce et al. [18] with intrinsic hydrophobicities by Kovacs et al. [4] was reported recently showing generally poor correlations [3]. In our comparative study we tried to include all hydrophobicity scales derived from RP-HPLC retention prediction studies in the past 8 years (Table 1). Most of them used acidic pH eluents (formic acid and TFA). Baczek et al. [11] used the measured RP retention values of individual amino acids as a component of their QSSR model. It was shown, however, that the properties of amino acids differ significantly when they are linked through peptide bonds [3]. In case of the "kernel function with support vector machine" based optimization the retention coefficients were not reported [12]. Both of these models used less than 100 peptides for optimization, making their results difficult to evaluate, as discussed in the following sections. Klammer et al. [13] reported support vector regression optimization using 12 different datasets of 150–2384 peptides. Scales by Petritis et al. [9] and Shinoda et al. [10] were obtained using an artificial neural network optimization with ~7000 and 834 peptides, respectively.

Table 1
Cross correlations between hydrophobicity (retention coefficients) values for different models.

Model	Correlation values for 15 residue/20 residue sets			
	Ac-XG-[a] [4]	SSRCalc 100 Å TFA [7]	SSRCalc 300 Å TFA [7]	SSRCalc 100 Å FA [8]
Designed synthetic peptides models				
-GX-OH (100 Å TFA) [14]	0.990/0.985	0.983/0.936	0.992/0.965	0.989/0.891
-GX-Amide (100 Å TFA) [14]	0.999/0.988	0.993/0.942	0.994/0.967	0.994/0.886
Ac-XG- (100 Å TFA)[a] [4]	1/1	0.996/0.939	0.996/0.963	0.995/0.884
NH_2-XG- (100 Å TFA) [14]	0.976/0.983	0.976/0.909	0.961/0.928	0.963/0.832
-GXG- (100 Å TFA) [14]	0.993/0.969	0.994/0.971	0.995/0.978	0.993/0.942
Guo et al. (300 Å TFA) [15]	0.963/0.920	0.957/0.952	0.974/0.958	0.971/0.952
Random peptide dataset models (regression analysis)				
SSRCalc 300 Å TFA [7]	0.996/0.963	0.997/0.992	1/1	0.997/0.960
SSRCalc 100 Å TFA [7]	0.996/0.939	1/1	0.997/0.992	0.996/0.983
SSRCalc 2004 300 Å TFA [6]	0.983/0.941	0.979/0.961	0.991/0.983	0.985/0.920
SSRCalc 100 Å FA [8]	0.995/0.884	0.996/0.983	0.997/0.960	1/1
Gilar (100 Å TFA) [5]	0.988/0.974	0.985/0.986	0.986/0.982	0.985/0.927
Gilar (100 Å FA) [5]	0.989/0.883	0.986/0.971	0.984/0.945	0.986/0.988
Random peptide dataset models (support vector regression analysis)				
Klammer et al. (100 Å FA) [13]	0.935/0.752	0.940/0.898	0.955/0.864	0.944/0.943
Random peptide dataset models (analytical neural network)				
Petritis et al. (100 Å TFA–FA) [9]	0.646/0.677	0.653/0.700	0.683/0.729	0.677/0.695
Shinoda et al. (100 Å FA) [10]	0.806/0.686	0.795/0.764	0.827/0.747	0.823/0.827

Correlations are shown for 15 (no Cys, Pro, Arg, Lys, His) and all 20 residues.
-GX-OH – from Tripet et al. [14]; NH_2-GAGAGVGLGX-OH, C18 100 Å, TFA based eluent.
-GX-Amide – from Tripet et al. [14]; NH_2-GAGAGVGLGX-Amide, C18 100 Å, TFA.
[a] Ac-XG- – from Kovacs et al. [4]; Ac-XGAKGAGVGL-Amide, C18 100 Å, TFA; intrinsic hydrophobicity scale.
NH_2-XG- – from Tripet et al. [14]; NH_2-XGAKGAGVGL-Amide, C18 100 Å, TFA.
-GXG- – from Tripet et al. [14]; LGLGXGLGLGK, C18 100 Å, TFA.
Guo et al. [15]; Ac-GXXLLLKK-Amide; C18 300 Å, TFA.
SSRCalc 300 Å TFA – from Krokhin [7]; ~4000 random tryptic peptides, C18 300 Å, TFA.
SSRCalc 100 Å TFA – from Krokhin [7]; ~5000 random tryptic peptides, C18 100 Å, TFA.
SSRCalc 2004 300 Å TFA – from Krokhin et al. [6]; 346 random tryptic peptides, C18 300 Å, TFA.
SSRCalc 100 Å FA – from Dwivedi et al. [8]; ~4000 random tryptic peptides, C18 100 Å, formic acid.
Gilar (100 Å TFA) – from Gilar et al. [5]; 165 random tryptic peptides, C18 100 Å, TFA.
Gilar (100 Å FA) – from Gilar et al. [5]; 165 random tryptic peptides, C18 100 Å, formic acid.
Klammer et al. (100 Å FA) – from Klammer et al. [13]; 2080 random tryptic peptides, C18 100 Å, formic acid.
Petritis et al. (100 Å TFA–FA) – from Petritis et al. [9]; ~7000 random tryptic peptides, C18 100 Å, formic acid–TFA mixture.
Shinoda et al. (100 Å FA) – from Shinoda et al. [10]; 834 random Lys-C peptides, C18 100 Å, formic acid.

Retention coefficients reported by Gilar et al. [5] and Krokhin [7] were obtained using more traditional optimization approaches with linear regression analysis on 165 and ~4000–5000 peptide datasets, respectively. The unique feature of the latter was the introduction of a number of sequence specific corrections to take into account nearest neighbor and helicity effects.

Hydrophobicity/hydropathy scales determined by classical methods such as the accessible surface area approach show very poor correlation with RP-HPLC derived scales (below 0.5 R^2-value, Appendix A). The whole residue scales determined by Wimley et al. [19] for water–octanol interface showed the closest match (a correlation of 0.72) with the Kovacs et al. [4] values.

3.2. Consensus rules for RP-HPLC hydrophobicity scales

In their review, Mant et al. [3] outlined general requirements for hydrophobicity scales obtained by a RP-HPLC experiment on C18 sorbents for peptides in random coil conformation. Most of these rules are based on a general knowledge of the structures of amino acids, and can be used for rapid assessment of a particular scale. For example, the hydrophobicity/retention coefficients for C18 sorbents should increase as follows: (i) Gly < Ala < Val < Ile < Leu and (ii) Asp < Glu; Asn < Gln; Ser < Thr. If any of these criteria are not met then the hydrophobicity scale, its underlying data or analysis methods required some additional evaluation. The hydrophobicity scales in Appendix A indicate that the artificial neural approaches are "multiple violators" of these rules. However, these models are still included in the detailed comparison provided in Table 1.

3.3. Detailed comparison of hydrophobicity scales: unique role of charged residues and proline. "Goodness" of the fit

Plotting retention coefficients of the models against each other, as in Fig. 1, can provide a detailed comparison of hydrophobicity scales. We selected our SSRCalc 100 Å TFA model as the most accurate retention predictor for the comparison with hydrophobicity scales generated by the synthetic peptides approach (Fig. 1A and C), as most of them employ the same separation conditions – C18 100 Å columns with TFA based eluents [4,14]. Fig. 1A shows correlation of normalized intrinsic hydrophobicities and SSRCalc retention coefficients. An R^2-value correlation of 0.939 was found for all 20 residues. This plot allows us to easily spot correlation outliers: Cys, Pro, Arg, His, and Lys.

3.3.1. Cysteine

The Cys residue in our studies was always alkylated with iodoacetamide (a common proteomics practice), in contrast to the free Cys group in the designed peptide approach used by Hodges and co-workers. Not surprisingly, the hydrophobicity values obtained were significantly different. Note that comparison of SSRCalc scales with Gilar et al.'s [5] values show very similar results for Cys as a result of their identical alkylation chemistry (Fig. 1G and H).

3.3.2. Proline

The intrinsic hydrophobicity for Pro as determined by Kovacs et al. [4] was found to be larger than its retention coefficient in SSRCalc (Fig. 1A and B). We believe the Pro residue's unique structure of an α-amino group as a part of side chain is responsible for

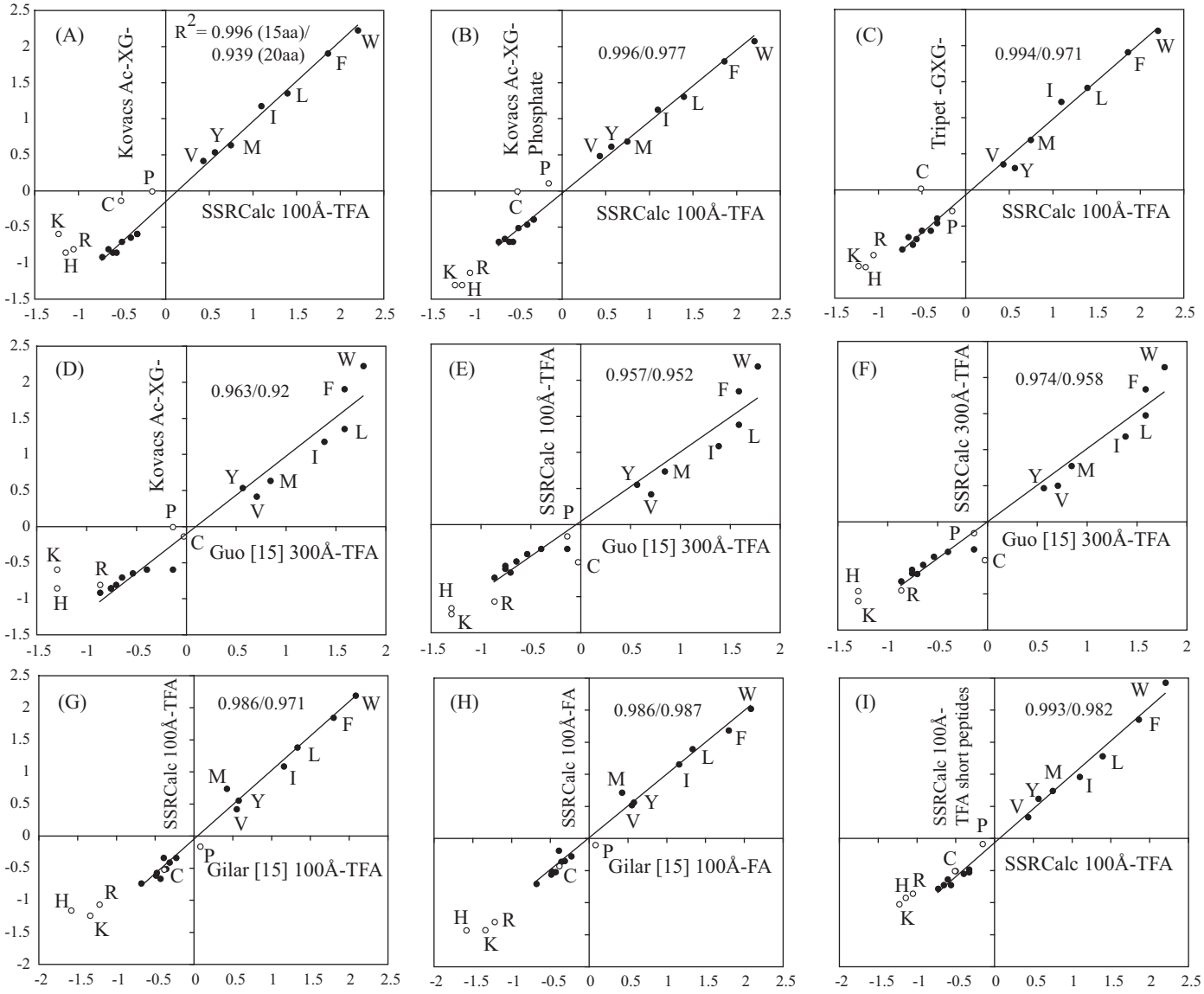

Fig. 1. Correlation between normalized hydrophobicity scales determined by different approaches. R^2-value correlations are shown for both 15 (excluding Cys, Pro, Arg, Lys, and His) and all 20 residues (e.g. 0.996/0.977). For the assignment of the scales, please refer to Table 1 and Appendix A.

this, making Pro a very rigid residue. We postulate this makes it difficult to access the hydrophobic/hydrophilic properties of Pro. We do agree that the intrinsic hydrophobicity/hydrophilicity values of Pro as determined for Ac-XGAKGAGVGL-Amide substituted series of peptides correctly represents the properties of the residue within these specific peptides, but overall the impact of Pro on the hydrophobicity of a peptide/protein is different when it is positioned inside the sequence. When the designed peptide approach data set was constrained to only the LGLGXGLGLGK species, the hydrophobic contribution of Pro was found to be nearly identical to that from the SSRCalc model (Fig. 1C).

3.3.3. Arg-His-Lys

The intrinsic hydrophobicity values of the charged residues Arg, His and Lys vary greatly across the Kovacs et al. [4] and SSRCalc 100 Å TFA models [7] (Fig. 1A): these residues are more hydrophobic in the designed peptide approach. Kovacs's coefficients obtained for the more hydrophilic phosphate counter-ion show a much better correlation with SSRCalc TFA values (Fig. 1B). But it is common knowledge that the hydrophobicity of charged residues is affected by hydrophobicity of its counter-ions: Arg, Lys, and His are more hydrophilic in a phosphate buffer system at the same pH. Tripet et al. [14] also noticed that Arg, His and Lys become more hydrophobic at the terminal positions as compared to the internal locations by comparing values of the LGLGXGLGLGK and Ac-XGAKGAGVGL-Amide series. This results in slightly better correlation when SSRCalc 100 Å TFA is compared to -GXG- set of peptides (Fig. 1C).

A detailed explanation of the observed discrepancies and the unique role of charged residues is given in the following sections. However, for the purpose of comparing hydrophobicity scales derived by RP-HPLC we propose to simply exclude Pro, Cys, Arg, His and Lys from consideration. Positional constraints for (Pro) and the alkylation status (Cys) can be applied as a filter for inclusion of these residues into consideration, while the pH 2 determination of hydrophobicity (Arg, His, and Lys) is ambiguous.

Comparison of retention coefficients (intrinsic hydrophobicity) values between the best designed synthetic peptides [4] and random peptide data sets [7] methods shows that there is virtually no difference between the scales obtained using these different approaches. All SSRCalc models (independent of pore size of the sorbent or ion-pairing modifier (TFA or FA)) demonstrate a 0.995+

R^2-value correlation against the intrinsic hydrophobicities for 15 residues obtained using the Ac-XGAKGAGVGL-Amide series of peptides (Fig. 1A and Table 1). This high degree of similarity permits the creation of consensus hydrophobicity scales determined using a RP-HPLC approach.

Based on these findings we suggest new criteria for determining "goodness of fit" of different hydrophobicity scales derived from C18 100 Å sorbents:

(1) As discussed, the expected order for retention should be observed: Gly < Ala < Val < Ile < Leu; Asp < Glu, Asn < Gln and Ser < Thr. We suggest that additionally Trp be the most hydrophobic residue, followed by Phe, as in Mant et al. [3]. However, some discrepancies were observed; in the results reported by Guo et al. [15] the coefficients for Leu and Phe were equal. Hydrophobicity values for the Val-Tyr-Met triad were found very close in most of the scales (especially for Val-Tyr). However agreement between the Kovacs et al. [4] and SSRClac scales [7] indicate that the order should be Val ≤ Tyr < Met, yielding a hydrophobicity order of: Gly < Ala < (Val ≤ Tyr) < Met < Ile < Leu < Phe < Trp.
(2) We feel that a more stringent criteria than the R^2-value of 0.9–0.95 in Mant et al. [3] should be used. If correlation between hydrophobicity values is below 0.98 the datasets or underlying mathematics behind coefficient derivation should be examined.

3.4. Comparison of random peptide dataset RP-HPLC approach with intrinsic hydrophobicity scales determined by Kovacs et al. [4]

As we noted before, poor correlations were found between the Kovacs et al. [4] scales and all examples of RP-HPLC on a random datasets selected by Mant et al. for comparison [3]. In our opinion, most of the discrepancies observed were due to the limited number of peptides (25–100) used in these earlier studies [2,16,17]. We illustrate in subsequent sections that these are insufficient numbers for the proper evaluation of retention coefficients using regression analysis. However, the data set reported by Wilce et al. [18] contained 1738 entries but gave the poorest correlation (0.847 R^2-value) [3] with intrinsic hydrophobicity values. It was noted that Trp and Met were the most "notorious" outliers in the Wilce et al.'s studies, but no explicit explanation was given [3]. We postulate that the low observed hydrophobicity values for Trp and Met were likely caused by their oxidation, again illustrating the power of mass spectrometry when used intelligently by chromatographers: had these peptides being filtered through MS detection these incorrect peptide retention values would have been excluded from the Wilce et al. [18] datasets.

In considering hydrophobicity scales derived in the proteomics era, one feature becomes obvious: the application of advanced optimization techniques such as artificial neural networks [9,10] provides the poorest correlations (Table 1). It is difficult to isolate the mechanism behind such deviations as the quality of retention datasets was probably good based on the level of experience in the respective groups. Support vector regression analysis produced an intermediate quality of correlation of ~0.94–0.95, but with a number of violations of the consensus hydrophobicity rules (Appendix A). Note that in Table 1 from the Klammer et al. data [13], we used only the best results obtained with standard digestion/separation protocols (trypsin/formic acid conditions with 60 cm column) for a training set of 2080 peptides. Overall, the authors used 12 different datasets to derive hydrophobicity contributions of individual amino acids, including results obtained using the MudPit approach and utilizing different enzymes. The introduction of an additional separation phase with a high salt eluent component in MudPit lowers the correlations. The same was observed for elastase and chymotrypsin datasets, but in these cases a dataset with a low number of peptides (150) was used [13].

Gilar et al. [5] applied a linear regression analysis using a classical additive retention prediction approach with a correction for the peptide length, against a dataset of 165 peptides. Retention coefficients were optimized for formic acid and TFA as ion-pairing modifiers and across different column temperatures. We find the reported retention coefficients correlate to 0.984+ against both the intrinsic hydrophobicities of Kovacs et al. [4] and SSRCalc [7] (Table 1 and Fig. 1G and H). Another distinctive difference observed in Gilar's scale is variation in Val-Tyr-Met triad: Met < Val < Tyr was typical.

The best correlations (>0.995) were seen between the intrinsic hydrophobicity values reported by Kovacs et al. [4] and those from the SSRCalc models. The latter were optimized using 4000–5000 peptide data sets, and attempted to take into account sequence dependent features of peptide retention. Nearest neighbor effects and the stabilization of amphipathic helical structures were suggested as a major reasons for the inability of the random peptide approach to reproduce hydrophobicity scales obtained by the designed peptide approach [3]; SSRCalc attempts to account for these effects. We suggested that ion-pairing formation, which involves charged residues, affects the apparent hydrophobicity of neighboring residues [6,7]. We also correct peptide retention based on the presence of amphipathic helical stretches of motifs such as XXOOXX and XXOXX, where X corresponds to hydrophobic residues and O to any other residues. These corrections were introduced based on empirical rules and improved the SSRCalc prediction accuracy [7]. As we show here, these corrections facilitated a more accurate assignment of retention coefficients by partially limiting these effects. The difference between 0.984+ and 0.995+ correlations for Gilar et al.'s [5] and SSRCalc [7] in Table 1 reinforces the value of introducing of sequence specific features, and the use of larger peptide datasets.

3.5. Sensitivity of retention coefficients method towards variation of pore size and ion-pairing modifier

Peptide's retention is very sensitive to the variation of sorbent's chemistry, the type of ion pairing agent, the temperature of the column, and the pore size of the sorbent. The influence of polar end-capping on separation selectivity was demonstrated by random peptide sets [20] and by designed peptide's approaches [21]. Gilar et al. [5] evaluated the effect of column temperature on retention coefficients using their collection of 165 peptides. In our study we considered it worthwhile to investigate if variations in pore size and the type of ion pairing modifier reflect in amino acid hydrophobicity values across the different models. For example, the Guo et al. [15] model was created for the series of designed peptides separated on C18 300 Å sorbent with TFA based eluent (Fig. 1D–F). Not surprisingly its 15 retention coefficients correlates best with SSRCalc 300 Å TFA (R^2 = 0.974, Fig. 1F), as well as the original SSRCalc reported in 2004 [6] also for 300 Å TFA (R^2 = 0.986, results not shown). As shown in Fig. 1G and H, the Gilar et al. [5] 100 Å FA model correlates much better with SSRCalc 100 Å FA (R^2 = 0.988, for all 20 residues), compared to SSRCalc 100 Å TFA (R^2 = 0.971). Comparison of retention coefficients with and without the basic residues included is also a good indication of the identity of the ion-pairing modifiers between two models. Thus for Gilar (100 Å FA)–SSRCalc (100 Å FA) pair correlation was 0.986 and 0.988 for 15 and all 20 residues, respectively (Fig. 1H). In case of the Gilar (100 Å FA)–SSRCalc (100 Å TFA) pair shown in Fig. 1G, correlations were 0.986 and 0.971: the inclusion of charged residues into consideration results in a lowered correlation for non-identical ion-pairing chemistries. Similar comparison allows to assign Klammer et al.'s

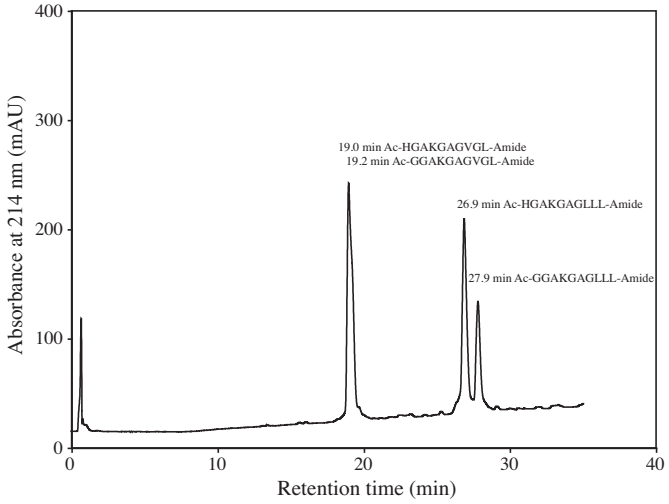

Fig. 2. Chromatographic separation of Gly and His substituted peptides of common *Ac-XGAKGAGVGL-Amide* and *Ac-XGAKGAGLLL-Amide* compositions. Luna C18(2) 100 Å column; linear 1% acetonitrile per minute starting from 0% acetonitrile, 0.1% TFA.

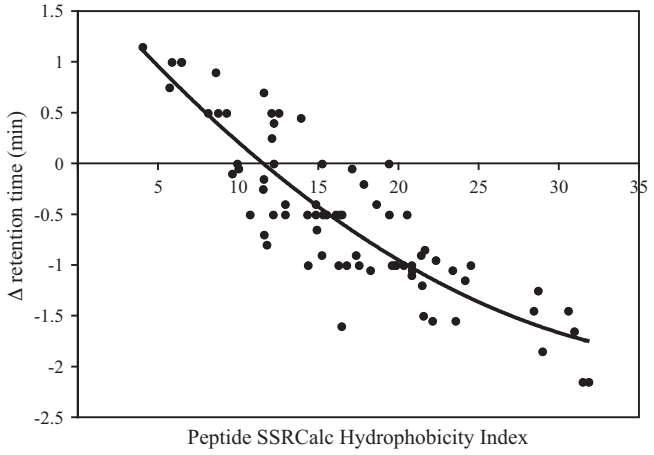

Fig. 3. The effect of the N-terminal addition of Arg on peptide retention on a C18 100 Å column (from dataset of ~5000 tryptic peptides, TFA-based eluents, 0.75% per minute acetonitrile gradient [7]). 75 pairs of peptides differing by one extra Arg residue were used to generate this plot (e.g. TPIAVR–RTPIAVR, ANVMDYR–RANVMDYR).

[13] and Shinoda et al.'s [10] models being conducted using formic acid based eluents even without knowing their experimental conditions. As shown in Table 1 when comparing to SSRCalc (100 Å FA) their correlations are identical in both 15- and 20-residue plots.

3.6. Intrinsic hydrophobicity of charged residues

All charged residues in a peptide sequence are involved in ion pairing interactions under RP-HPLC conditions. It has been shown on multiple occasions that variation in the ion-pairing chemistry under pH 2 conditions results in the alteration of hydrophilicity values for Arg, His, and Lys [22,23]. Leaving aside these effects, we intended to pursue the more complicated question: why are the hydrophobicity values for charged residues inconsistent even when using identical sorbent and ion pairing chemistry?

As an example, the hydrophobicity values for Lys, His and Arg reported in the Kovacs et al. TFA [4] scale were found to be higher than those in Guo et al. [15] (Fig. 1D). Tripet et al. [14] noted the same – their series *LGLGXGLGLGK* exhibited much higher hydrophilicity of Lys, His, and Arg compared to the *Ac-XGAKGAGVGL-Amide* in Kovacs et al. [4] as well as all other models with N- or C-terminal as the position of substitution [3]. It was suggested that charged residues are more hydrophobic in the N-terminal position than when they are internal. Analyzing 5 sets of designed synthetic peptides ([3,14], shown in the captions for Table 1) one can conclude that the sequence for the *-GXG-* series is the most hydrophobic among them (contains 4 Leu). The same is true for Guo et al. [15] peptides also exhibiting more hydrophilic character for Lys, His, Arg: 3 Leu in a framework sequence. These initial observations suggested that the hydrophobicity of basic amino acids could be determined not by position, but rather overall hydrophobicity of the peptide.

To test this hypothesis we synthesized a series of peptides, two of which were identical to ones reported by Kovacs et al. *Ac-XGAKGAGVGL-Amide* (X = Gly, His) and two were the same length and composition, but with the last 3 residues being Leu: *Ac-XGAKGAGLLL-Amide* with Gly and His in N-terminal position. Fig. 2 shows separation of these two pairs under 100 Å C18 conditions with 1% acetonitrile per minute gradient (TFA based eluent system). Confirming the previous findings [4], the substitution of Gly with basic His does not change the retention of a peptide in the *Ac-XGAKGAGVGL-Amide* framework. *Ac-GGAKGAGLLL-Amide* and *Ac-HGAKGAGLLL-Amide* have the latter series length and position of substituted residue, but exhibit a substantial decrease in retention: $\Delta t_R = -1$ min following the Gly-His substitution. These findings suggest that the hydrophobicity/hydrophilicity of charged residues is determined by the overall peptide hydrophobicity rather than by its position in the peptide.

It is known that the contribution of a particular amino acid in peptide retention depends on the overall size of a peptide: the larger peptide, the smaller the contribution. Thus Mant et al. [24] introduced a correction factor related to a peptide length. It is also known that peptide length correlates with its hydrophobicity for the random peptide data sets: the longer the sequence, more hydrophobic the peptide is expected to be. These two relations make it difficult to determine which property of a peptide – length, or hydrophobicity, is the dominant factor behind deviations under additive models. These deviations are responsible for the concave character of t_R vs. hydrophobicity plots, which require correction for peptide length [5,24] or both length and hydrophobicity [6,7]. However, independent of peptide size or hydrophobicity, the substitution of Gly for more hydrophobic residues (Ala < (Val ≤ Tyr) < Met < Ile < Leu < Phe < Trp) will always result in higher retention of peptide in random coil conformation. This increase will be smaller for longer, more hydrophobic peptides.

The same scaling rules will not be applicable to charged residues. The contribution of the charged residues compared to Gly in TFA-based eluents can be positive (Lys, Arg [4]), near zero (His [4], Fig. 2) or negative, as shown for *Ac-GGAKGAGLLL-Amide* and *Ac-HGAKGAGLLL-Amide* in Fig. 2 as well as for *LGLGXGLGLGK* [14] and *Ac-GXXLLLKK-Amide* [15] series.

SSRCalc uses different sets of retention coefficients for short ($N < 9$) and long peptides [7]. The rationale behind these corrections was developed empirically; the resulting hydrophobicity values were never compared to the other scales. Fig. 1I shows such a comparison across SSRCalc 100 Å TFA retention coefficients between short and long peptides. Confirming the previous finding from the designed peptide approach, charged residues were found to be more hydrophobic for short (relatively hydrophilic) peptides.

We found additional proof of variability of charged residue's retention contribution upon variation of peptide hydrophobicity when studied changes in chromatographic behavior caused by the N and C-terminal additions of Lys and Arg. A typical tryptic digest often contains peptides with missed cleavages when the protein sequence features two or more adjacent cleavage sites: -RR-, -KK-,

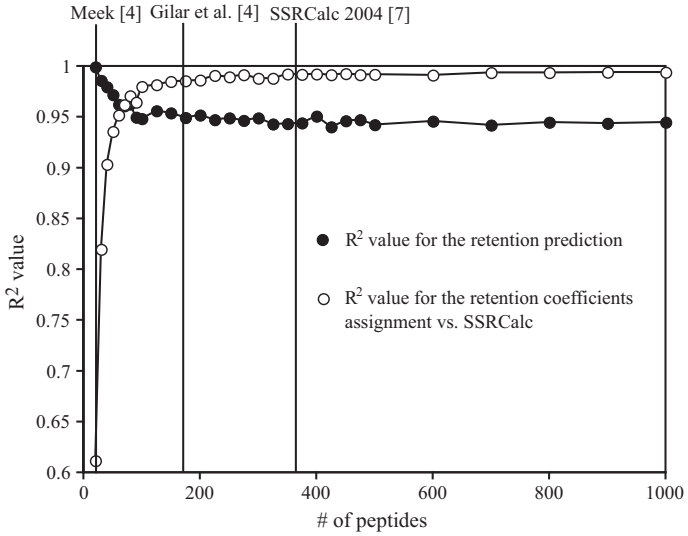

Fig. 4. The accuracy of (1) retention prediction (R^2-value against observed peptide retention times) and (2) assignment of the retention coefficients (R^2-value against the SSRCalc scale), as a function of the size of a random peptide optimization data set. Each reported value is the average across 5 optimization runs, with new randomly selected peptides in each run. ●, retention prediction; ○, retention coefficients.

-RK-, -KR-, etc. In this case, the missed cleavage results in the appearance of two or more related species differing by one extra charged residue (for example: HPDYSVVLLLR and RHPDYSVVLLLR from human albumin). Fig. 3 shows how the retention time of a peptide changes upon N-terminal addition of Arg depending on peptide hydrophobicity. The observed trend is similar to the process discussed in this section: peptide retention can both increase and decrease, depending on the hydrophobicity of parent molecule.

Collectively, this data shows that the correct assignment of hydrophobicity of charged residues is crucial for building accurate retention prediction models. We envision the introduction of "sliding" hydrophobicity scales, where retention contribution of the charged residues will vary depending on peptide hydrophobicity and/or size. The same will apply to Asp and Glu at neutral and basic pHs. Therefore the specific determination of hydrophobicity/hydrophilicity of charged residues requires taking into account the overall hydrophobicity (size) of a peptide. Such a drastic difference in behavior of neutral and charged amino acids arises from dualistic (hydrophobic and ion-pairing) mechanism of peptide RP-HPLC retention. We think that the contribution of the ion-pairing mechanism varies with the concentration of the organic solvent required for a peptide elution, which in-turn correlates with the hydrophobicity (and indirectly with the length) of the peptides.

3.7. Determining retention coefficients using a random peptide collection: dataset size requirements

Throughout the manuscript we have noted the disadvantage of a random peptide dataset approach when low number of the species (25–100) is used. It was of interest to explore how many peptides are required for "correct" (0.98+ correlation against values from the designed series method) assignment of the retention coefficients. We randomly selected various numbers of peptides (20–1000) from a total of ~5000 peptides originally used in optimizing the SSRCalc 100 Å TFA model, then performed linear regression optimizations for these subsets. The selection of random peptides and optimization was repeated five times for each subset size and the results were averaged. Fig. 4 shows the average values of retention coefficients correlating against SSRCalc's retention hydrophobicity values (for 17 residues), as well as the average accuracy of the additive prediction models (R^2-value of t_R vs. hydrophobicity plots). The size of peptide datasets corresponding to Meek [2], Gilar et al. [5] and Krokhin et al. [6] are indicated on this plot: 25, 165 and 364 species, respectively. This figure provides an excellent illustration that a set of 25 peptides can provide near-perfect 0.99+ accuracy for an additive retention prediction model. But this result is due to an almost complete overfitting of the model against the training data, and these retention coefficients show a very poor correlation (~0.7 R^2) against the consensus hydrophobicity values. This also demonstrates that a simple additive retention prediction model involving a correction for the peptide length alone cannot yield retention prediction accuracy above 0.945 for randomly selected peptides.

4. Conclusions

Determination of the intrinsic hydrophobicity of amino acids in a random coil conformation using the RP-HPLC of designed synthetic peptides or extended random peptide collections both produce virtually identical results (0.995+ R^2 correlations). This degree of agreement between these two alternative methods is a novel finding. The availability of high quality proteomics-derived data has allowed us to independently confirm findings of Hodges and co-workers, thus laying foundation for assigning consensus hydrophobicity scales derived from RP-HPLC experiments. The accurate assignment of hydrophobicity values by the SSRCalc model was possible due to taking into account such sequence specific features as nearest neighbor effects and the influence of amphipathic helixes formation. Overall peptide retention is determined by amino acid composition. While the influence of secondary structure formation and ion-pairing (affects nearest neighbor's interaction) causes deviations from predicted retention, it's also complicates determination of intrinsic hydrophobic/hydrophilic contribution of the residues. A very close match between retention coefficients determined during our model's optimization and designed peptide approach indicates correctness of our approach to describing these phenomena. The hydrophobic contribution of the residues present in the other conformations (such as amphipathic helix) undoubtedly will be different from ones reported here. Precise determination of these contributions will require detailed studies comprising best practices of random peptide dataset and designed peptide approaches.

The repetitive regression optimization of a simple length-corrected additive retention prediction model using various sizes

of dataset illustrates that at least 100 random peptides are required for a 0.98+ R^2 of amino acid intrinsic hydrophobicity values against those found using the designed peptide approach. The excellent agreement in amino acid intrinsic hydrophobicity values derived from the alternative approaches has also allowed us to precisely assign and explain the observed differences. We show that determining hydrophobicity of charged residues is strongly affected not only by the type and concentration of ion-pairing modifier, but also but by peptide's hydrophobicity. We attribute differences in values for the basic residues to the influence of ion pairing formation during RP-HPLC separation. In our opinion, defining hydrophobicity/hydrophilicity of these residues is only possible for the particular hydrophobicity of a framework peptide (for designed set approach), or using the average hydrophobicity of the species from particular random peptide collection.

Acknowledgements

The authors thank Drs A. Klammer and W.S. Noble for providing the hydrophobicity contribution value data from their support vector regression approach. This work was supported by the grant from the Natural Sciences and Engineering Research Council of Canada (O.V.K.).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.chroma.2011.06.092.

References

[1] K.M. Biswas, D.R. DeVido, J.G. Dorsey, J. Chromatogr. A 1000 (2003) 637.
[2] J.L. Meek, Proc. Natl. Acad. Sci. U.S.A. 77 (1980) 1632.
[3] C.T. Mant, J.M. Kovacs, H.M. Kim, D.D. Pollock, R.S. Hodges, Biopolymers 92 (2009) 573.
[4] J.M. Kovacs, C.T. Mant, R.S. Hodges, Biopolymers 84 (2006) 283.
[5] M. Gilar, H. Xie, A. Jaworski, Anal. Chem. 82 (2010) 265.
[6] 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.
[7] O.V. Krokhin, Anal. Chem. 78 (2006) 7785.
[8] R.C. Dwivedi, V. Spicer, M. Harder, M. Antonovici, W. Ens, K.G. Standing, J.A. Wilkins, O.V. Krokhin, Anal. Chem. 80 (2008) 7036.
[9] 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, R.D. Smith, Anal. Chem. 75 (2003) 1039.
[10] K. Shinoda, M. Sugimoto, N. Yachie, N. Sugiyama, T. Masuda, M. Robert, T. Soga, M. Tomita, J. Proteome Res. 5 (2006) 3312.
[11] T. Baczek, P. Wiczling, M. Marszall, Y.V. Heyden, R. Kaliszan, J. Proteome Res. 4 (2005) 555.
[12] N. Pfeifer, A. Leinenbach, C.G. Huber, O. Kohlbacher, BMC Bioinformatics 8 (2007) 468.
[13] A.A. Klammer, X. Yi, M.J. MacCoss, W.S. Noble, Anal. Chem. 79 (2007) 6111.
[14] B. Tripet, D. Cepeniene, J.M. Kovacs, C.T. Mant, O.V. Krokhin, R.S. Hodges, J. Chromatogr. A 1141 (2007) 212.
[15] D. Guo, C.T. Mant, A.K. Taneja, J.M.R. Parker, R.S. Hodges, J. Chromatogr. 359 (1986) 499.
[16] J.L. Meek, Z.L. Rossetti, J. Chromatogr. 211 (1981) 15.
[17] C.A. Browne, H.P.J. Bennett, S. Solomon, Anal. Biochem. 124 (1982) 201.
[18] M.C.J. Wilce, M.I. Aguilar, M.T.W. Hearn, Anal. Chem. 67 (1995) 1210.
[19] W.C. Wimley, T.P. Creamer, S.H. White, Biochemistry 35 (1996) 5109.
[20] V. Spicer, A. Yamchuk, J. Cortens, S. Sousa, W. Ens, K.G. Standing, J.A. Wilkins, O.V. Krokhin, Anal. Chem. 79 (2007) 8762.
[21] C.T. Mant, D. Cepeniene, R.S. Hodges, J. Sep. Sci. 33 (2010) 3005.
[22] D.C. Guo, C.T. Mant, R.S. Hodges, J. Chromatogr. 386 (1987) 205.
[23] M. Shibue, C.T. Mant, T.W. Burke, R.S. Hodges, J. Chromatogr. A 1080 (2005) 68.
[24] C.T. Mant, T.W. Burke, J.A. Black, R.S. Hodges, J. Chromatogr. 458 (1988) 193.