Journal of Chromatography, 599 (1992) 211–225
Elsevier Science Publishers B.V., Amsterdam

CHROMSYMP. 2553

Prediction of peptide retention time in reversed-phase high-performance liquid chromatography

Claire Chabanet

Station de Biométrie, INRA, 78352 Jouy-en-Josas (France)

Mireille Yvon*

Station de Recherches Laitières, INRA, 78352 Jouy-en-Josas (France)

ABSTRACT

Peptide retention in reversed-phase chromatography depends mainly on the amino acid composition of peptides and can therefore be predicted by summing the relative hydrophobic contributions of each constitutive amino acid residue. The prediction is correct for small peptides but overestimates the retention times of peptides larger than 10–15 residues. A new prediction model is proposed in which the contribution to peptide retention of each amino acid residue is not a constant but a decreasing function of peptide length. From the retention times of 104 peptides, the parameters of decreasing functions were estimated by a non-linear multiple regression analysis. The contribution to peptide retention of charged, polar and non-polar residues appears to be differently affected by peptide length. The secondary structure of most peptides during reversed-phase high-performance liquid chromatography could be responsible for this. The high correlation between the predicted and observed retention times of peptides which were not used to establish the model indicates a good predictive accuracy of the new model.

INTRODUCTION

Several workers [1–3] have reported the advantages of predicting the retention of peptides of known composition in reversed-phase high-performance liquid chromatography (RP-HPLC). Briefly, a knowledge of the retention times of given peptides would simplify their chromatographic separation and purification from complex mixtures. In addition, it would allow one to predict the solubility of peptides in precipitating agents such as trichloroacetic acid or sulphosalicylic acid solutions, as this solubility has been found to be highly correlated with the peptide retention time in RP-HPLC [4]. For preparative applications, such precipitation techniques would allow a first mixture separation before the use of RP-HPLC as the final step of the isolation procedure.

It is now recognized that the retention times of small peptides can generally be predicted by summing the relative hydrophobic contributions of each constitutive amino acid residue. Therefore, retention coefficients for amino acid residues have been determined for different chromatographic systems (mobile phase, stationary phase, pH) using either the retention times of a wide range of peptides of varied composition and length [1,5–10] or the retention time of a synthetic octapeptide model in which two residues were successively substituted by each of the 20 amino acids found in proteins [2].

The retention times of peptides larger than 10–15 residues are less than that predicted by summing the retention coefficients of each constitutive amino acid residue. Several researchers [3,11] have reported an exponential relationship between the observed retention times and the peptide length for a series of peptide polymers. Moreover, Mant *et al.* [3] demonstrated that this exponential relationship varied from one series of peptide polymers to another, depending on their hydrophobicity. They therefore

0021-9673/92/$05.00 © 1992 Elsevier Science Publishers B.V. All rights reserved

introduced a correction factor for peptide retention time prediction, taking into account the hydrophobicity and the length of the peptide [3,12]. Recently, the same group [13] added to their prediction method a further correction factor, to predict the retention behaviour of amphipathic α-helices during reversed-phase chromatography.

Assuming that this chromatographic behaviour for large peptides could be due to a decrease in the accessibility of certain residues or to a removal of certain residues from contact with the stationary phase, we attempted to establish and to test a new model, in which the contribution of each amino acid residue to peptide retention times would not be a constant value but a decreasing function of peptide length, depending on the non-polar, polar or charged nature of the residues.

EXPERIMENTAL

Materials

HPLC-grade acetonitrile was obtained from Baker (Deventer, Netherlands) and trifluoroacetic acid (TFA) from Pierce (Rockford, IL, USA). Water was purified by passage through a Milli-Q water purification system (Millipore, Bedford, MA, USA).

Most of the peptides were obtained by enzymatic (tryptic, chymotryptic, pepsic, plasmic) degradation of milk proteins (α_{s1}-casein, α_{s2}-casein, β-casein, κ-casein, β-lactoglobulin, lactoperoxidase) as described previously [4,14–16]. Other peptides were purchased from Sigma (St. Louis, MO, USA). The origin of the peptides is noted in the tables.

Apparatus

The HPLC instrument consisted of two M510 pumps, a Wisp 710B injector and a Lambda Max M481 spectrophotometer (Waters Assoc., Milford, MA, USA). The system was coupled to a computer equipped with Baseline 810 software (Waters Assoc.).

Chromatographic measurements

Chromatographic measurements were made at room temperature using a Waters μBondapack C_{18} (10 μm) column (250 × 4.6 mm I.D.).

Linear elution was carried out with 0.11% TFA in water and 0.1% TFA in acetonitrile–water (60:40) over a gradient slope of 1.2% acetonitrile/min at a flow-rate of 1 ml/min. The elution was monitored by measuring the absorption at 220 nm.

The retention time was expressed as the acetonitrile concentration in the solvent at the elution time. This was calculated by subtracting the gradient elapsed time from the peak elution time and then multiplying by the percentage of acetonitrile per minute in the linear gradient. The gradient elapsed time (t_g) was previously defined by Guo *et al.* [2] as the time for the gradient to reach the detector from the proportioning valve via the pump, injection system and column. This value was measured as described previously [2].

Establishment of the model

Retention times of the small peptides were recorded and fitted to the linear relationship

$$T_{\mathrm{ret}_i} = \sum_{j=1}^{19} n_{ij} a_j + b_0 + \varepsilon_i \qquad (1)$$

where

n_{ij} = number of amino acid residues j in peptide i;
a_j = retention coefficient for residue j;
b_0 = retention coefficient for α-NH_2 and α-COOH terminal functions;
ε_i = independent errors which are assumed to be normally distributed, with the same variance.

In order to simulate a decrease in the contribution of residues in larger peptides, we considered a new model in which the contribution of each residue to peptide retention time (A_j) is a decreasing function of peptide length (l_i). The decreasing function was chosen to have a slope equal to zero when $l_i = 0$, an inflection point and a lower asymptote:

$$T_{\mathrm{ret}_i} = \sum_{j=1}^{19} n_{ij} A_j(l_i) + b_0 + \varepsilon_i \qquad (2)$$

where

$A_j(l) = (a_j - I_j)\mathrm{e}^{-b_j l^2} + I_j$
$I_j = a_j/k_j$

In small peptides the contribution of each residue (A_j) is close to a_j (the retention coefficient for residue j). In very long peptides this contribution is I_j (lower asymptote) which, is proportional to a_j ($I_j = a_j/k_j$).

The b_j parameter of the function sets the curve slope.

From this general model, two sub-models were considered. In the first, assuming that the peptide chain length effects may be the same on all residues, we imposed k_j and b_j to be similar for all residues. In the second, we assumed the decrease in the contribution of residues to be dependent on the amino acid residue considered. This model thus consisted of nineteen functions and each of them contained three parameters, a_j, b_j and k_j. Therefore, 58 ($19 \times 3 + b_0$) parameters had to be estimated. It was impossible to estimate as many parameters from the 104 observed data. In order to decrease the number of parameters, three groups of amino acid residues were formed: Gly, Ala, Val, Met, Ile, Leu, Phe, Trp were considered as non-polar residues, Asp, Asn, Thr, Ser, Glu, Gln, Pro, Tyr, His as polar residues and Lys and Arg as charged residues. Based on the accessibility of residues in proteins [17,18], the same k_j was imposed for all residues in the same group (which means that the same proportion of the residues surface area was assumed to be accessible in proteins) and for each group the decreasing function $A_j(l_i)$ was supposed to have the same shape (same b_j).

Therefore, the retention times of the total peptides were recorded and fitted to eqn. 2 where:

in model 1:

$A_j(l) = (a_j - I_j)e^{-bl^2} + I_j$ for all residues
$I_j = a_j/k$

and in model 2:

$A_j(l) = (a_j - I_j)e^{-b_1 l^2} + I_j$ for non-polar residues
$I_j = a_j/k_1$

$A_j(l) = (a_j - I_j)e^{-b_2 l^2} + I_j$ for polar residues
$I_j = a_j/k_2$

$A_j(l) = (a_j - I_j)e^{-b_3 l^2} + I_j$ for charged residues
$I_j = a_j/k_3$

A non-linear multiple regression analysis was done to estimate $a_j(19)$, b_0, b, k of the first model and $a_j(19)$, b_0, b_1, b_2, b_3, k_1, k_2, k_3 of the second model using the maximum-likelihood method and the statistical package "NL" [19].

RESULTS AND DISCUSSION

Retention behaviour of peptides

The retention times of 104 peptides tested are listed in Table I. We used peptides of different length and composition so that all residues often enough appear to permit an accurate determination of their contribution to peptide retention time according to the peptide length.

TABLE I

OBSERVED AND PREDICTED RETENTION TIMES OF PEPTIDES USED TO ESTABLISH THE MODEL

The retention times were predicted using eqn. 2 with the parameter values in Table II.

Peptide	Sequence	No. of residues	t_R observed [CH_3CN (%)]	t_R predicted [CH_3CN(%)]	
				Model 1	Model 2
1[a]	Y	1	0.2	0.6	0.9
2[a]	F	1	2.1	4.7	3.4
3[a]	W	1	7.3	6.4	5.4
4[a]	FA	2	2.5	4.1	3.1
5[a]	YA	2	0.2	0	0.4
6[a]	FG	2	2.9	4.8	3.9
7[a]	FY	2	8.4	8.7	9.3
8[a]	LV	2	3.0	4.5	3.8
9[a]	SY	2	1.3	1.3	1.8
10[a]	YV	2	2.8	2.7	3.2
11[a]	WG	2	6.2	6.4	5.8
12[a]	LL	2	8.5	8.4	7.7
13[a]	LY	2	5.8	6.5	7.2
14[a]	HF	2	3.8	3.3	2.2

(*Continued on p. 214*)

TABLE I (*continued*)

Peptide	Sequence	No. of residues	t_R observed [CH_3CN (%)]	t_R predicted [CH_3CN(%)]	
				Model 1	Model 2
15[a]	MP	2	6.5	3.9	3.7
16[a]	LP	2	4.9	4.7	4.7
17[a]	LF	2	12.1	10.6	9.8
18[a]	TL	2	1.7	2.8	1.9
19[a]	WA	2	5.6	5.7	5.1
20[a]	LM	2	4.8	7.6	6.7
21[a]	LW	2	13.8	12.2	11.8
22[a]	WE	2	5.2	6.9	6.3
23[a]	WGG	3	5.0	6.5	6.3
24[a]	YVG	3	6.4	2.7	3.5
25[a]	YYL	3	14.2	10.5	12.6
26[a]	IPI	3	10.8	9.7	9.7
27[a]	FGG	3	2.3	4.8	4.3
28[a]	MLF	3	18.8	15.5	15.2
29[a]	MLG	3	4.7	7.6	7.2
30[a]	LLL	3	17.0	14.1	14.0
31[a]	EPM	3	3.4	4.4	4.5
32	LRF	3	13.9	11.3	10.8
33	YQL	3	9.4	6.8	7.9
34	RFF	3	14.3	13.5	12.9
35[a]	FGFG	4	14.7	15.5	12.7
36	LQSW	4	12.0	12.9	13.5
37	FRQF	4	12.4	13.6	13.7
38	RQFY	4	9.6	9.7	10.6
39	HIQK	4	0.9	0.4	0.2
40	HPIK	4	1.8	2.3	2.4
41[a]	YGGFL	5	17.8	15.7	15.4
42	LHSMK	5	4.7	6.1	6.2
43	RLKKY	5	5.3	6.0	6.7
44[a]	WIIWLQL	6	27.4	25.6	27.4
45	VNELSK	6	4.4	5.8	5.6
46	EAMAPK	6	2.2	2.5	2.8
47	EMPFPK	6	12.5	13.6	14.2
48	TTMPLW	6	22.1	19.2	20.0
49[a]	MEHFRWG	7	18.4	18.4	19.4
50[a]	YPFPGPI	7	21.6	19.7	20.3
51	VLPVPQK	7	9.8	10.2	11.0
52	DMPIQAF	7	17.8	14.4	16.7
53	GPFPIIV	7	24.3	20.8	21.6
54	AVPYPQR	7	7.3	7.1	7.7
55	EDVPSER	7	3.9	2.7	3.0
56	EKVNELSK	8	5.9	5.4	5.3
57	DAYPSGAW	8	13.0	10.7	11.4
58	KKYKVPQL	8	14.0	8.5	9.2
59	YVPLGTQY	8	13.6	14.2	14.4
60	VAPFPQVF	8	21.3	19.2	20.2
61	LGYLEQLL	8	22.3	23.1	24.2
62	EGIHAQQK	8	3.2	0.5	1.5
63	YYVPLGTQY	9	15.7	17.5	16.6
64[a]	PHPFHFFVYK	10	22.3	24.6	25.1
65	QLDAYPSGAW	10	15.3	15.6	16.0

TABLE I (*continued*)

Peptide	Sequence	No. of residues	t_R observed [CH_3CN (%)]	t_R predicted [CH_3CN(%)]	
				Model 1	Model 2
66	YLGYLEQLLR	10	27.7	26.3	26.4
67	QLDAYPSGAWY	10	16.6	18.6	17.9
68	SLSQSKVLPVPE	12	14.7	12.8	15.1
69	FFVAPFPQVFGK	12	29.8	30.4	31.3
70	QLDAYPSGAWYY	12	18.3	21.4	19.5
71	QFYQLDAYPSGAW	13	23.8	24.2	23.3
72	RPKHPIKHQGLPQE	14	9.0	9.6	9.9
73	VPQLEIVPNSAEER	14	15.4	15.4	14.4
74	AVPYPQRDMPIQAF	14	18.8	20.2	19.3
75	HQGLPQEVLNENLLR	15	20.8	20.5	20.8
76	YYVPLGTQYTDAPSF	15	20.3	22.2	19.0
77	FQSEEQQQTEDELQDK	16	8.1	8.9	9.4
78	YKVPQLEIVPNSAEER	16	17.7	16.8	15.1
79	YQEPVLGPVRGPFPIIV	17	28.9	29.3	28.4
80	LYQEPVLGPVRGPFPIIV	18	29.7	32.2	31.7
81	LLYQEPVLGPVRGPFPIIV	19	31.2	34.8	34.6
82	LTLTDVENLHLPLPLLQSW	19	32.0	35.7	36.3
83	EPMIGVNQELAYFYPELFR	19	30.7	34.1	32.6
84	RPKHPIKHQGLPQEVLNENL	20	18.8	17.6	17.8
85	SDIPNPIGSENSEKTTMPLW	20	22.6	24.3	22.7
86	RPKHPIKHQGLPQEVLNENLLRF	23	25.9	23.8	24.6
87	RPKHPIKHQGLPQEVLNENLLRFF	24	27.6	27.2	27.9
88	TDAPSFSDIPNPIGSENSEKTTMPLW	26	23.9	23.9	23.0
89	SLSQSKVLPVPQKAVPYPQRDMPIQA	26	19.2	19.6	20.7
90	SLSQSKVLPVPQKAVPYPQRDMPIQAF	27	22.8	22.8	23.8
91	SLSQSKVLPVPQKAVPYPQRDMPIQAFL	28	26.1	24.7	25.7
92	SLPQNIPPLTQTPVVVPPFLQPEVMGVSK	29	30.0	26.9	27.0
93	YYVPLGTQYTDAPSFSDIPNPIGSENSEK	29	23.1	20.9	21.4
94	YPVEPFTESQSLTLTDVENLHLPLPLPLLQSW	30	34.9	32.3	31.7
95	YVPLGTQYTDAPSFSDIPNPIGSENSEKTTMPLW	34	25.7	25.4	25.8
96	YYVPLGTQYTDAPSFSDIPNPIGSENSEKTTMPLW	35	28.8	26.3	27.1
97	LQPEVMGVSKVKEAMAPKHKEMPFPKYPVQPFTESQS	37	21.4	20.1	20.8
98	KVPQLEIVPNSAEERLHSMKEGIHAQQKEPMIGVNQEL	38	19.9	18.7	18.4
99	KVPQLEIVPNSAEERLHSMKEGIHAQQKEPMIG VNQELAYF	41	21.6	21.8	21.2
100	KVPQLEIVPNSAEERLHSMKEGIHAQQKEPMIGVN QELAYFYPELF	46	28.4	27.6	26.6
101	QFYQLDAYPSGAWYYVPLGTQYTDAPSFSDIPNPI GSENSEK	42	28.5	26.7	29.5
102	IHPFAQTQSLVYPFPGPIPNSLPQNIPPLTQTPVVVPPFLQ PEVMGVSK	49	32.7	35.2	34.8
103	YPVEPFTESQSLTLTDVENLHLPLPLLQSWMHQPHQPL PPTVMFPPQSVLSLSQSK	56	37.6	40.3	37.9
104	MAIPPKKNQDKTEIPTINTIASGEPTSTPTIEAVESTVAT LEASPEVIESPPEINTVQVTSTAV	64	25.4	25.4	26.6

[a] Peptides purchased from Sigma. The other peptides were obtained by enzymatic degradation of α_{s1}-, β- and κ-caseins [4].

Calculation of retention parameters

The retention coefficients for amino acid residues (a_j) were first determined from retention times of small peptides containing either up to ten residues (67 peptides) or up to seven residues (55 peptides) using the linear eqn. 1. Retention coefficients could not be calculated from retention times of peptides smaller than six residues because some residues did not appear. The correlation between observed and predicted retention times was much better with the retention times calculated by summing the coefficients determined with peptides containing up to seven residues. Moreover, all but two coefficients were estimated with a better precision although the sample was smaller. These results indicate that the linear model is no longer satisfactory for peptides containing more than seven residues. The same observation was previously made by Mant *et al.* [3], who reported observed retention times less than those expected from the sum of the retention coefficients of each constitutive residue for some ten-residue peptides.

From the retention times of all peptides and using the new model 1 of retention time prediction (eqn. 2), the 22 parameters ($19a_j + b_0 + b + k$) of the function were determined simultaneously. Using model 2 it was impossible to estimate all the parameters of the function together and therefore some parameters had to be set. Because only two residues are charged and their retention coefficients are small, we chose to set a parameter for this group. Taking into account that about 50% of their surface area remains accessible in proteins [17], I_j for charged residues was set to $a_j/2$ and the other parameters were calculated. The k_1 and k_2 values were calculated to be 4.9 and 2.9, respectively. I_j was thus approximated and set to $a_j/5$ for non-polar, $a_j/3$ for polar and $a_j/2$ for charged residues and the other parameters, $a_j(19)$, b_1, b_2 and b_3, were re-calculated. It was found that b_3 was not significantly different from zero and the sub-model with $b_3 = 0$ was accepted using the likelihood ratio test. This indicated that the contribution of a charged residue to the peptide retention time was not significantly dependent on the peptide length. Hence the contribution to the peptide retention time of charged residues [$A_j(I)$] was set equal to a_j and $k_3 = 2$ had no meaning.

The calculated parameters of both models and the retention coefficients determined with the linear model and the small peptides are listed in Table II.

Comparison and validity of the two new models

As neither model 1 or 2 was a sub-model of the other, they could not be compared by the likelihood ratio test and therefore four criteria were used to compare and to evaluate them.

(1) Comparison of the calculated retention coefficients for amino acid residues (a_j) with those determined with small peptides. The retention coefficients (a_j) calculated with both new models were well related with those determined with the linear model and the small peptides ($\leqslant 7$ residues) (Fig. 1). However, the sum of squares of differences between the calculated coefficients was the least with the second model (13.46 against 14.6). The a_j values determined with the second model are slightly higher than those obtained with the first model, especially for polar residues, whereas the b_0 value was much smaller.

(2) Relationship between predicted and observed retention times. The retention times calculated with both new models were plotted against the observed retention times for the 104 peptides used to establish the model (Fig. 2). A high correlation was observed in both instances; the coefficients were 0.98 and 0.99 for models 1 and 2, respectively.

(3) The likelihood. The parameters of both models were calculated by using the maximum likelihood method. For a similar number of calculated parameters (degrees of freedom = 89) the likelihood was higher for model 2 [-2 log (likelihood) = 433 and 460.7 for models 2 and 1, respectively].

(4) Structure of reduced residuals. The reduced residuals (difference between observed and calculated retention times of peptide i divided by the estimated standard error) must be randomly distributed around zero.

With both models the plot of reduced residuals *versus* calculated retention time did not have any structure but a structure was observed in the plot of reduced residuals *versus* peptide length with the first model (Fig. 3).

With this first model, depending on the peptide chain length, the predicted retention time was either underestimated (for peptides containing between four and ten residues and those containing over twenty residues) or overestimated (for very short

TABLE II

RETENTION COEFFICIENTS OF AMINO ACID RESIDUES (a_j) AND PARAMETERS OF THE FUNCTION $A_j(l)$

The retention coefficients [CH_3CN (%)] were predicted either (1) from retention times of small peptides (≤ 7 residues) and from eqn. 1, or (2) from retention times of all peptides and from eqn. 2 with model 1 or (3) from retention times of all peptides and from eqn. 2 with model 2. The numbers in parentheses represent the number of amino acids used for each calculation and S.E. is the estimated standard error.

Amino acid residue and parameter		Retention coefficient					
		1		2		3	
		a_j	S.E.	a_j	S.E.	a_j	S.E.
	Trp	10.24	0.56 (11)	9.81	0.65 (25)	10.64	0.59 (25)
	Phe	8.81	0.43 (22)	8.15	0.39 (64)	8.65	0.35 (64)
	Leu	6.91	0.33 (27)	5.93	0.30 (134)	6.51	0.29 (134)
	Tyr	5.45	0.45 (16)	4.10	0.38 (71)	6.16	0.49 (71)
	Ile	6.16	0.52 (11)	5.50	0.53 (63)	5.90	0.46 (63)
	Met	5.15	0.58 (11)	5.13	0.67 (35)	5.54	0.60 (35)
	Pro	2.39	0.36 (21)	2.26	0.38 (178)	3.58	0.41 (178)
	Val	2.55	0.57 (9)	2.03	0.51 (91)	2.56	0.43 (91)
	Ser	0.58	0.77 (5)	0.64	0.52 (85)	1.14	0.65 (85)
	Gln	−0.41	0.64 (8)	0.30	0.43 (102)	1.11	0.50 (102)
	Arg	0.74	0.60 (9)	0.84	0.63 (30)	1.06	0.44 (30)
	Glu	0.24	0.56 (12)	0.56	0.40 (94)	1.03	0.48 (94)
	Asn	−1.10	1.14 (2)	1.02	0.79 (43)	1.03	0.97 (43)
	Thr	0.73	0.69 (3)	0.31	0.47 (55)	0.65	0.50 (55)
	Gly	−0.05	0.37 (15)	0.12	0.41 (63)	0.50	0.36 (63)
	Asp	1.30	1.25 (2)	−0.40	0.97 (32)	0.31	1.21 (32)
	Lys	−1.35	0.41 (18)	−0.55	0.39 (70)	−0.18	0.26 (70)
	Ala	−0.39	0.55 (7)	−0.61	0.46 (57)	−0.27	0.42 (57)
	His	−0.96	0.69 (6)	−1.35	0.60 (31)	−1.24	0.74 (31)
α-Amino + α-COOH	b_0	−4.66	0.49 (55)	−3.45	0.55 (104)	−5.25	0.48 (104)
	b_1	–	–	0.0019	0.0001	0.0017	0.0001
	b_2	–	–	0.0019	0.0001	0.0136	0.0028
	b_3	–	–	0.0019	0.0001	0	
	k_1	–	–	3.11	0.13	5	
	k_2	–	–	3.11	0.13	3	
	k_3	–	–	3.11	0.13	2	

peptides with less than three residues and those containing between ten and twenty residues). Such a structure did not exist with the second model.

All the criteria and especially the structure of reduced residuals *versus* peptide length indicate that the second model was slightly better than the first. Consequently, the peptide retention time prediction was improved by considering three groups of amino acid residues with their contributions following different functions of peptide length. However, from these data, the first model cannot be ruled out categorically.

On the other hand, in the second model three groups of amino acids may not be enough and the amino acids distribution in the different groups may not be accurately chosen, but we would need much more data to be able to estimate the b_j *and* k_j parameters for each residue.

Contribution of residues to peptide retention times as a function of peptide length $A_j(l)$

Model 1. With this model, the contribution of all residues to peptide retention times followed the same relationship:

$$A_j(l) = (a_j - I_j)e^{-0.0019l^2} + I_j$$
$$I_j = a_j/3.11$$

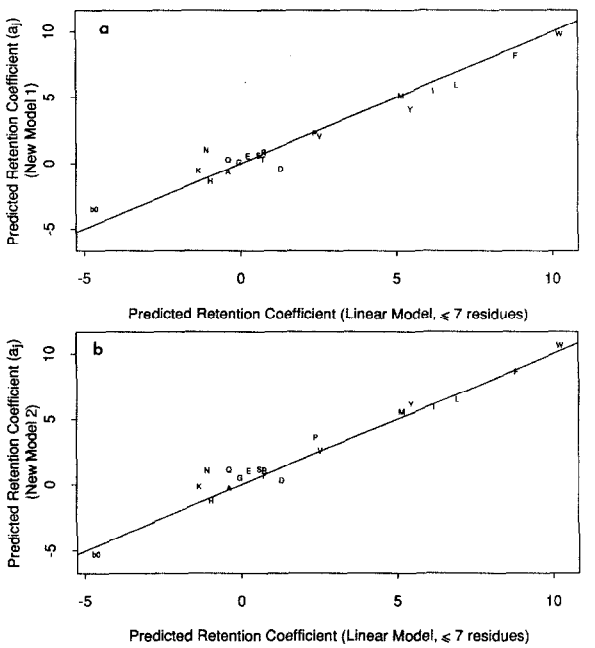

Fig. 1. Comparison of the sets of predicted retention coefficients for amino acid residues (Table I). (a) Relationship between the set determined with all peptides and model 1, and that determined with eqn. 1 and small peptides ($\leqslant 7$ residues). (b) Relationship between the set determined with all peptides and model 2, and that determined with eqn. 1 and small peptides ($\leqslant 7$ residues).

(Fig. 4a). It mainly decreased when the peptide length increased from 7 to about 30 residues.

Model 2. The contribution of non-polar residues to peptide retention times followed the relationship

$$A_j(l) = (a_j - I_j)e^{-0.0017l^2} + I_j$$
$$I_j = a_j/5$$

(Fig. 4b), that of polar residues the relationship

$$A_j(l) = (a_j - I_j)e^{-0.014l^2} + I_j$$
$$I_j = a_j/3$$

(Fig. 4c) and that of charged residues

$$A_j(l) = a_j$$

(Fig. 4d).

The contribution of the residues to peptide retention time decreased rapidly when the peptide length increased, especially for polar residues. In a pentapeptide the contribution of polar residues would be lower than in a dipeptide. These results could explain why the retention coefficients for polar

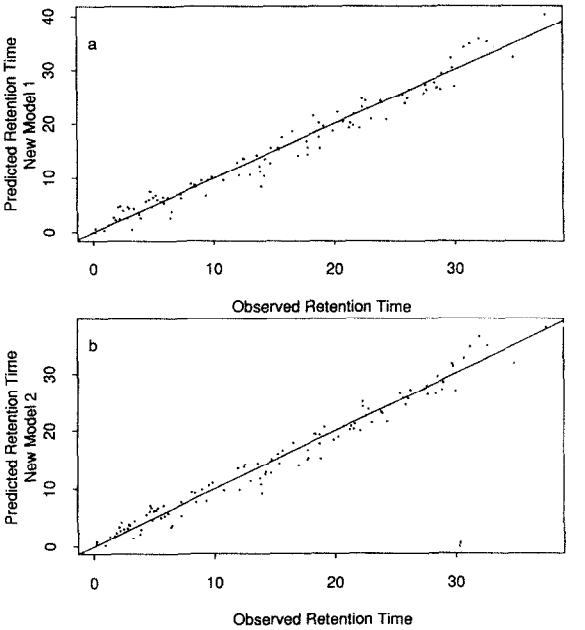

Fig. 2. Relationship between observed retention times of peptides used to establish the model and their predicted retention times (Table I) using (a) model 1 and (b) model 2. The correlation coefficients are 0.98 and 0.99, respectively.

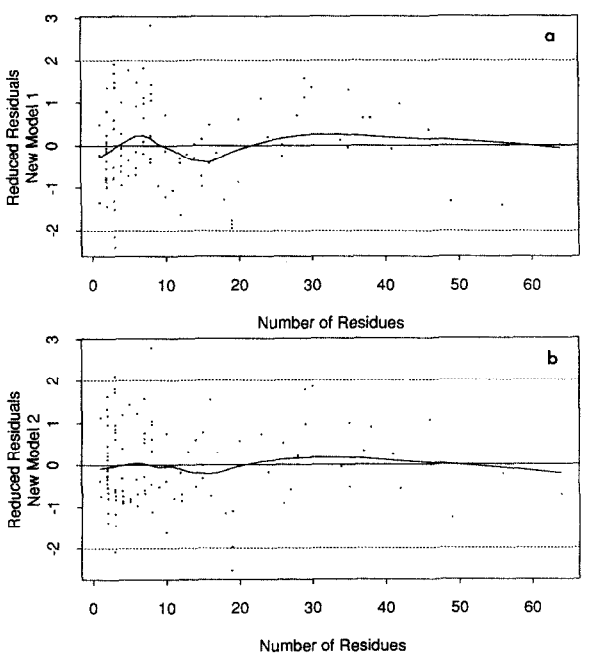

Fig. 3. Plots of the reduced residuals (difference between the observed and predicted retention times divided by the estimated standard error) *versus* peptide length, (a) with model 1 and (b) with model 2.

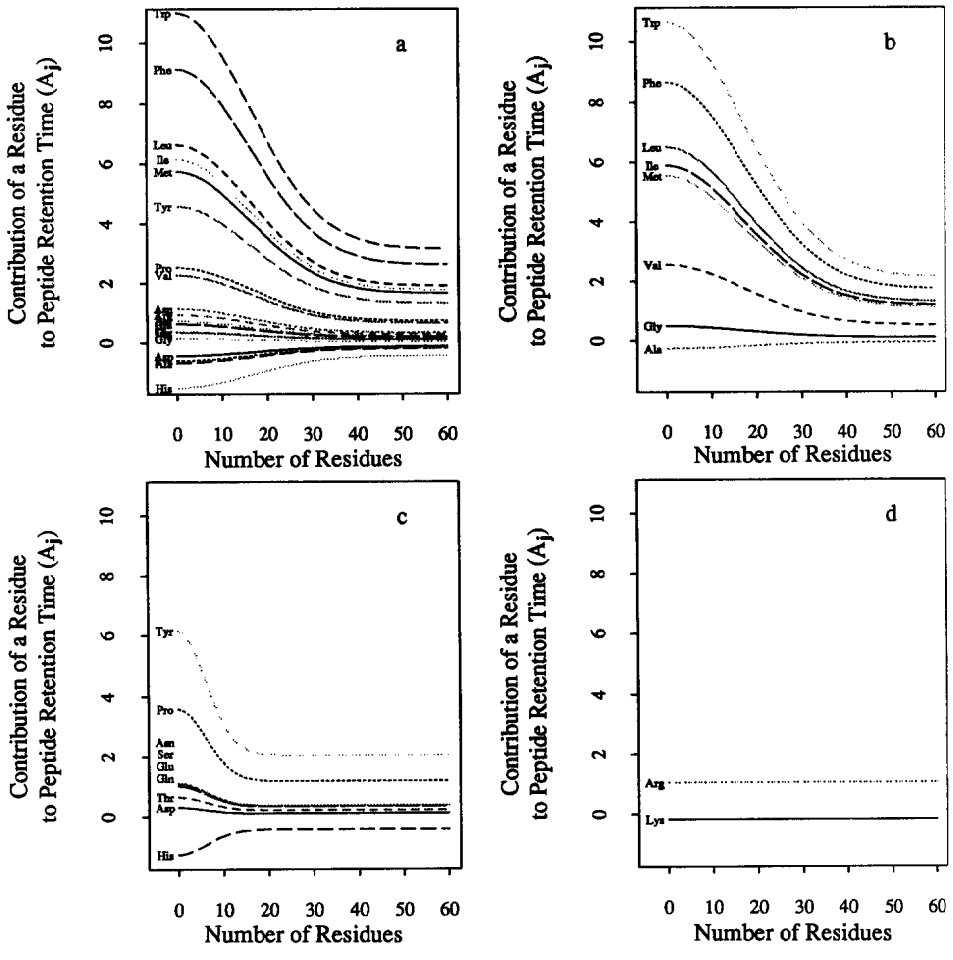

Fig. 4. Plots of predicted residue contribution (A_j) *versus* peptide length (l). (a) In model 1, for all residues the relationship is

$$A_j(l) = \left(a_j - \frac{a_j}{3.11}\right)e^{-0.0019l^2} + \frac{a_j}{3.11}$$

(b) In model 2, for non-polar residues the relationship is

$$A_j(l) = \left(a_j - \frac{a_j}{5}\right)e^{-0.0017l^2} + \frac{a_j}{5}$$

(c) In model 2, for polar residues the relationship is

$$A_j(l) = \left(a_j - \frac{a_j}{3}\right)e^{-0.014l^2} + \frac{a_j}{3}$$

(d) In model 2, for charged residues the relationship is

$$A_j(l) = a_j$$

residues predicted with the linear model from small peptides (less than seven residues) were slightly smaller than the a_j of the new model 2 (Fig. 1).

Our a_j values were compared with retention constants determined by other workers using chromatographic conditions close to ours (C_{18} column and aqueous TFA as the mobile phase and acetonitrile as the mobile phase modifier in a linear gradient elution system). With the weighted fit retention constants determined by Sasagawa *et al.* [1] (Fig. 5), the correlations were good (0.93 and 0.94 for models 1 and 2, respectively) and were better than that found with the non-weighted constants. The retention constants obtained using weighted analysis probably reflect more realistic constants than those obtained using non-weighted analysis because information on the conformation of large peptides was incorporated in the analysis.

The correlation between the a_j values of model 2 and the retention coefficients determined by Guo *et al.* [2] was poor. But as they used an octapeptide

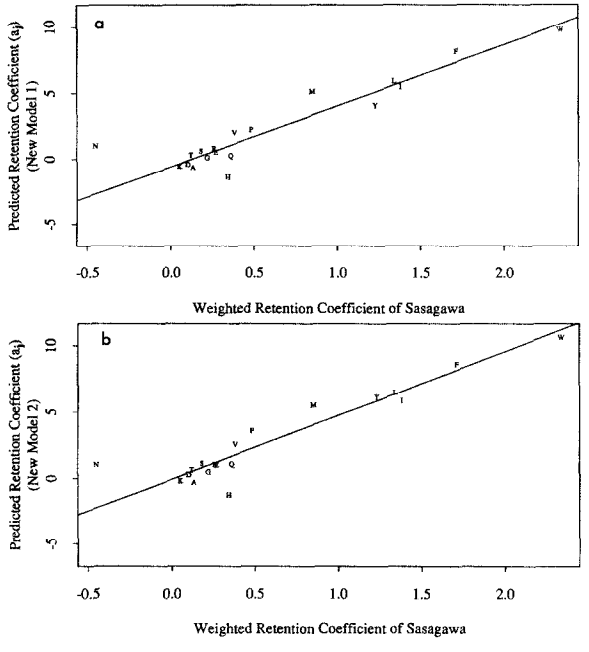

Fig. 5. Relationship between Sasagawa *et al.*'s retention constants for amino acid residues calculated by weighted curve fitting [1] and our retention coefficients (a_j) predicted by (a) model 1 (correlation coefficient 0.93) and (b) model 2 (correlation coefficient 0.94).

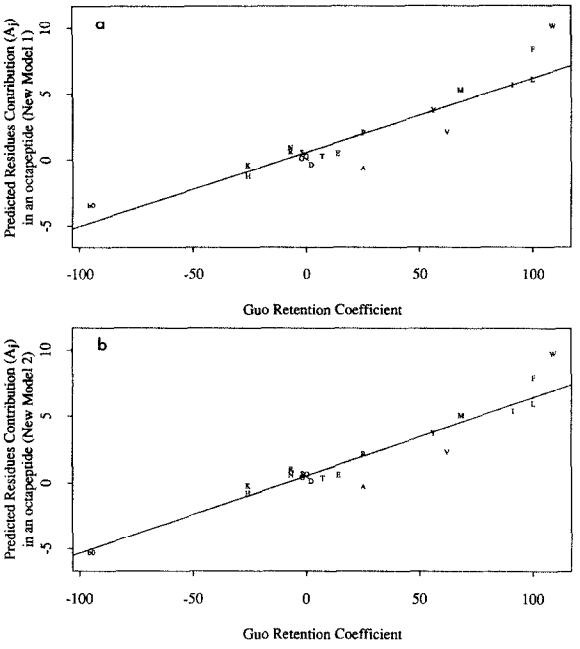

Fig. 6. Relationship between Guo *et al.*'s retention constants for amino acid residues obtained from synthetic octapeptides [2] and our values of residue contributions (A_j) in an eight-residue peptide. (These contributions were calculated with the equations in Fig. 4.)

model for the determination, their retention coefficients were compared with our predicted residues contributions in an eight-residue peptide (Fig. 6). These residue contributions were calculated using the functions $A_j(l)$ established for the three groups of amino acid residues with $l = 8$. The correlation was good (0.94) even though slight differences were found for four non-polar residues (Ala, Val, Phe and Trp). Similar discrepancies were observed with model 1 with the correlation being slightly lower (0.93). These differences might be due to the different column packing material used (we used μBondapack C_{18} instead of Synchropack C_{18}), especially for Ala and Val, as the constants determined by Sasagawa *et al.* [1], who used a μBondapack column, were similar to ours. Another possible explanation is that we imposed a similar function $A_j(l)$ for all non-polar residues and this function could be slightly different for each of them.

The calculated retention coefficient for His agreed with that calculated from the linear model and with

Guo *et al.*'s and Sasagawa *et al.*'s constants, although it was considered as a polar residue, whereas it could have been considered as a charged residue under our chromatographic conditions.

In the first model we assumed that whatever the peptide chain length, every residue contributed equally to the overall peptide hydrophobicity. Hence this model looks like the prediction method developed by Mant and co-workers [3,12] and suggested that the main peptide length effect is due to a decrease in the proportion of the peptide which binds to the stationary phase at any time and that there is no specific portion of the polypeptide chain interacting with the stationary phase. However, the structure of reduced residuals *versus* peptide length suggested that this model had a "bias", which means that, depending on the peptide length, the predicted retention time was either underestimated or overestimated. Although the interpretation of this feature is not easy, we can suggest an explanation. With this model, most of the a_j values were slightly lower than those obtained with the small peptides and with the second model. Therefore, generally, peptide retention was underestimated except for peptides containing between 7 and 25 residues. For these peptides the overestimation was probably due to an overestimation of the proportion of polar residues contributing to peptide retention (Fig. 4a and c). This high proportion was related to the low b value (model 1), which was due to the constraint $b_1 = b_2 = b_3$, and hence it seems that the peptide length affected the contributions of polar and non-polar residues differently.

In the second model, we assumed the effect of the peptide length on the contribution of residues to the peptide retention times to be dependent on the nature of the amino acid residues. This could be due to conformational constraints which removed certain residues from contact with the stationary phase or which led to a decrease in the accessibility of residues.

The b_j parameter of the function $A_j(l)$ sets the curve slope. The calculated b value for polar residues (b_2) was significantly higher than that for non-polar residues (b_1) as the sub-model with $b_1 = b_2$ (and with $k_1 = 5, k_2 = 3, k_3 = 2$) was rejected using the likelihood ratio test. The curve patterns (Fig. 4b and c) indicate that the decrease in the contribution of polar residues occurred mainly when the peptide length increased from 2 to 20 residues and thus could be due to the secondary structure, whereas the contribution of non-polar residues mainly decreased when the peptide length increased from 10 to 60 residues.

An explanation for the decrease in the contribution of polar residues to the peptide retention time during the formation of the secondary structure was proposed by Zhou *et al.* [13]. They showed that the formation of an amphipathic α-helix results in a preferred binding domain which is non-polar. Hence the polar residues which are not in the preferred binding domain would contribute less to peptide retention. For a peptide series having an α-helical structure with hydrophobic residues almost equally distributed on both sides of the helix, only a slight discrepancy between the observed and predicted [13] peptide retention time was observed. We cannot say that all peptides used in our study can form an amphipathic α-helix, but as the hydrophobicity of the column can induce secondary structures, most of them could adopt a structure (helical or other) on binding, leading to a preferred binding domain which would be the most non-polar but would not include all the non-polar residues. This could explain why the decrease in the non-polar residue contribution was lower than that of polar residues during the formation of the secondary structure.

With this model, the calculated residue contribution in a very large peptide (I_j) was in good agreement with the accessibility of residues in proteins determined by Chothia [17]: for non-polar residues, the average proportion of the residue surface area that remains accessible in proteins is 15.6%, for polar residues it is 33% and for charged residues 50%. This seems to indicate that the accessibility of residues in large peptides during RP-HPLC was close to that observed in proteins. However, imposing the k_j value to be similar for all residues ($k_1 = k_2 = k_3$), but $b_1 \neq b_2 \neq b_3$, led to a model almost as good as the previous one, without a structure of reduced residuals. In this model the k_j value was calculated between k_1 and k_2 values (4.05) and b_j and a_j values very close to those determined with the previous model ($b_1 = 0.0016, b_2 = 0.0130$ and b_3 close to 0). Moreover, as it has been demonstrated that RP-HPLC is a strong denaturant of tertiary and quaternary structures [11] of proteins and polypeptides, it is extremely difficult to argue

TABLE III

COMPARISON OF PREDICTED AND OBSERVED RETENTION TIMES OF 47 PEPTIDES NOT USED TO ESTABLISH THE MODEL

Retention times were predicted using eqn. 2 and model 2 with the parameter values in Table II (3).

Peptide	Sequence	No. of residues	t_R predicted [CH_3CN(%)]	t_R observed [CH_3CN(%)]	Error [CH_3CN(%)]
1[a]	VP	2	0.8	1.1	−0.3
2[a]	VW	2	7.9	8.9	−1.0
3[a]	GF	2	3.9	3.3	0.6
4[a]	FR	2	4.4	1.6	2.8
5[a]	GW	2	5.8	6.8	−1.0
6[a]	YL	2	7.2	6.1	1.1
7[a]	VA	2	−3.0	0	–
8[a]	KY	2	0.5	0.1	0.4
9[a]	VY	2	3.2	2.1	1.1
10[a]	GR	2	−3.7	0	–
11[a]	VS	2	−1.6	0	–
12[a]	GY	2	1.2	0.5	0.7
13[a]	VD	2	−2.4	0	–
14[a]	GWG	3	6.3	5.8	0.5
15[a]	VGG	3	−1.7	0	–
16[a]	GFG	3	4.3	3.0	1.3
17[a]	EVF	3	6.8	8.6	−1.8
18[a]	GHG	3	−5.4	0	–
19[a]	ETY	3	2.0	2.7	−0.7
20[a]	GPGG	4	−0.7	0	–
21[a]	PFGK	4	6.6	4.1	2.5
22[a]	MRFA	4	9.5	9.8	−0.3
23	SLLFM	5	22.0	20.0	2.0
24	VAGTWY	6	12.6	14.3	−1.7
25	ALPMHIR	7	14.0	13.0	1.0
26	TKIPAVFK	8	12.5	14.8	−2.3
27[a]	RGFFYTPKA	9	17.2	16.4	0.8
28[a]	RPPGFSPFR	9	19.5	16.3	3.2
29[a]	DRVYIHPFHL	10	20.4	20.9	−0.5
30	VLVLDTDYKK	10	14.1	13.9	0.2
31	TKVIPYVRYL	10	19.4	20.9	−1.5
32[a]	RPKPQQFFGLM	11	25.5	22.4	3.1
33[a]	ELYENKPRRPYIL	13	20.9	19.3	1.6
34	TPEVDDEALEKFDK	14	11.1	16.3	−5.2
35	IVGYLDEEGVLDQNR	15	18.6	16.1	2.5
36	WLPAEYEDGLALPFGWTQR	19	33.6	28.1	5.5
37	AMKPWIQPKTKVIPYVRYL	19	28.8	23.6	5.2
38	SLAMAASDISLLDAQSAPLR	20	21.2	23.7	−2.5
39	VYVEELKPTPEGDLEILLQK	20	23.9	25.4	−1.5
40	DLYKTPDNIDIWIGGNAEPM	20	25.5	23.4	2.1
41	KNTMEHVSSSEESIISQETYKQEK	24	11.1	14.4	−3.3
42	EQINAVTSFLDASLVYGSEPSLASR	25	21.5	29.3	−7.8
43	TVYQHQKAMKPWIQPKTKVIPYVRYL	26	25.3	22.6	2.7
44[a]	HSQGTFTSDYSKYLDSRRAQDFVQWLMNT	29	23.9	22.0	1.9
45	QKWIPPYQGYRNSVDPRISNVFTFAFRFGHM	31	32.3	30.1	2.2
46	NAVPITPTLNREQLSTSEENSKKTVDMESTEVFTKK	36	15.4	17.4	−2.0
47	KTKLTEEEKNRLNFLKKISQRYQKFALPQYLKTVYQH QKAMKPWIQPKTKVIPYVRYL	58	35.5	29.6	5.9

[a] Peptides purchased from Sigma. The other peptides were obtained by enzymatic or chemical (CNBr) degradation of α_{S2}-casein [14], β-lactoglobulin [15] and lactoperoxidase [16].

that the tertiary structure of peptides explains our results concerning both the I_j values and the decrease in the contribution of non-polar amino acid residues when the peptide length increases from 10 to 60 residues.

Accuracy of peptide retention prediction

The value of a predictive method must be assessed by its accuracy in predicting the retention times of peptides that were not used to determine the model. Therefore, model 2 was applied to predict the retention times of 47 peptides (Table III). These peptides were of various origins and their length varied from 2 to 58 residues. They were chromatographed under the conditions used to establish the model. As expected, several peptides predicted to have negative retention times had an observed retention time of zero. The predicted retention times were plotted against the observed retention times (Fig. 7). The relationship was linear with a correlation coefficient of 0.97.

We also used our model for predicting the retention times of peptides tested under chromatographic conditions close to ours (same stationary and mobile phases) by Sasagawa *et al.* [1] (Table IV). The correlation between their observed retention times and our predicted retention times for 71 peptides was 0.93 (Fig. 8). We only tested 71 peptides out of the 100 used by Sasagawa *et al.* because the others contained aminoethylcysteine, carboxymethylcysteine and trimethyllysine residues, for which we had not calculated the retention coefficients.

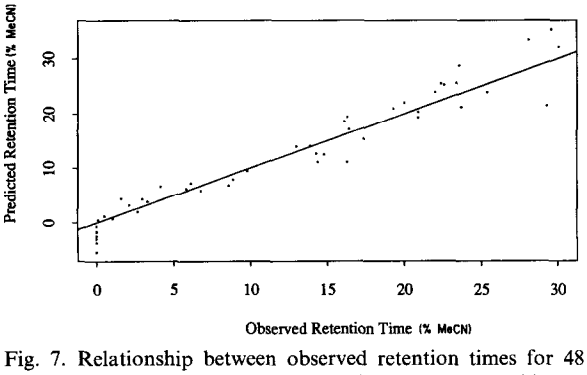

Fig. 7. Relationship between observed retention times for 48 peptides which were not used to establish the model (Table III) and their predicted retention times using the new model. The correlation coefficient is 0.97. MeCN = Acetonitrile.

The observed and predicted retention times were in good agreement except for 8 or 9 out of the 118 peptides tested (47 + 71). As these peptides (Nos. 34, 36, 37, 42 and 47 in Table III and Nos. 40, 46, 63 and 70 in Table IV) have different lengths (7, 8, 14, 19, 20, 25, 58 and 93 residues), the discrepancies observed might be due to sequence specific conformations or to nearest neighbour effects as defined by Mant *et al.* [3]. Nevertheless the new model permits a good prediction of peptide retention times.

In this empirical study, from the retention times of a large number of peptides of various lengths and compositions and mainly coming from milk proteins, we calculated the average contribution to the peptide retention time of each amino acid residue according to the peptide length. From the results, it seems that an almost similar proportion of polar and non-polar residues contributes to the peptide retention of small peptides (<5 residues). In the same way, a similar proportion of polar and non-polar residues contributes to the peptide retention of large peptides (>25 residues). However, for peptides containing between 7 and 25 residues, the proportion of non-polar residues contributing to the peptide retention is higher than that of the polar residues (1.5–2-fold higher). Such a feature may be due to the secondary structure of most peptides during RP-HPLC which leads to a most hydrophobic preferred binding domain, as shown previously by Zhou *et al.* [13] for peptides with an amphipathic α-helix structure.

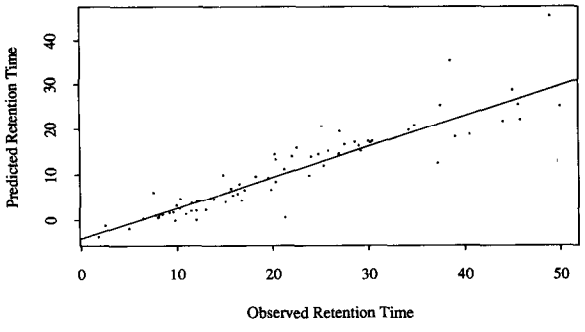

Fig. 8. Comparison of retention times observed by Sasagawa *et al.* [1] for 71 peptides (2–93 residues) with predicted retention times using our new model (eqn. 2). The chromatographic conditions used by Sasagawa *et al.* were close to ours: Waters μBondapack C_{18} column (300 × 4.0 mm I.D.), gradient from 0.1% aqueous TFA to 0.07% TFA in acetonitrile at 1% acetonitrile/min and flow-rate 2 ml/min.

TABLE IV

COMPARISON OF DATA FROM LITERATURE WITH PREDICTED RETENTION TIMES

Data were taken from Sasagawa *et al.* [1]. Retention times were predicted using eqn. 2 and model 2 with the parameter values in Table II (3).

Peptide	Sequence	No. of residues	Retention time	
			Observed	Predicted
1	GGG	3	1.8	−3.8
2	PG	2	2.5	−1.3
3	TEEQ	4	5.0	−1.9
4	MTAK	4	6.5	0.3
5	MAR	3	7.8	0.5
6	YK	2	8.0	0.6
7	TPGSR	5	8.1	1.2
8	KYE	3	8.2	1.2
9	GY	2	8.5	1.2
10	TEAEMK	6	9.2	1.6
11	EY	2	9.6	1.7
12	HLK	3	9.8	−0.1
13	FK	2	9.9	3.2
14	IRE	3	10.3	2.6
15	PL	2	10.3	4.7
16	IAE	3	10.9	1.3
17	GF	2	11.5	3.9
18	KMKDTDSEEE	10	11.5	2.0
19	AFR	3	12.0	4.1
20	DIAAK	5	12.0	0
21	QIAE	4	12.0	2.1
22	ASEDLK	6	13.0	2.3
23	EAFR	4	13.5	4.9
24	FDR	3	13.8	4.6
25	VFDKDGNGY	9	14.8	9.9
26	FKE	3	15.0	4.1
27	KVFGR	5	15.6	7.0
28	SLGQNPTEAE	10	15.8	5.5
29	GW	2	16.3	5.8
30	MIRE	4	16.5	7.9
31	SHPETLEK	8	16.7	4.3
32	HGLDNYR	7	17.0	6.6
33	LFK	3	18.2	9.5
34	IAEFK	5	19.5	9.2
35	ADIDGDGQVNYEE	13	19.8	6.6
36	VFDKDGNGYI	10	20.2	14.4
37	ISAAELR	7	20.3	8.4
38	FESNFNTQATNR	12	20.3	13.2
39	ELGTVMR	7	21.2	11.1
40	GHHEAELK	8	21.3	0.5
41	LQDMINE	7	22.0	13.9
42	FVQMMTAK	8	22.5	15.9
43	QIAEFK	6	23.8	9.8
44	RSLGQNPTEAELQDM	15	24.0	13.9
45	MIREADIDGDGQVNYEE	17	24.8	14.4
46	FLTMMAR	7	25.1	20.6
47	VDADGNGTIDFPE	13	25.3	12.0
48	LGTVMRSLGQNPTEAE	16	25.8	15.2
49	NTDGSTDYGILQINSR	15	26.9	14.5

TABLE IV (*continued*)

Peptide	Sequence	No. of residues	Retention time	
			Observed	Predicted
50	VEADVAGHGQDILIR	15	26.9	14.3
51	FLTMMARKMKDTDSEEE	17	27.0	19.5
52	VFDKDGNGYISAAELR	16	27.5	16.6
53	AFRVFDKDGNGYISAAE	17	28.6	17.1
54	VFDKDGNGYISAAEL	15	29.0	16.3
55	IREADIDGDGQVNYEEFVQM	20	29.2	15.2
56	EAFSLFDKDGDGTITTK	17	30.0	17.3
57	ALELFR	6	30.2	17.0
58	AFSLFDKDGDGTITTKE	17	30.4	17.3
59	NKALELFRKDIAAKYKELGYQG	22	34.2	19.8
60	PGYPGVYTEVSYHVDWIK	18	34.8	20.8
61	EADIDGDGQVNYEEFVQMMTAK	22	37.2	17.5
62	INEVDADGNGTIDFPEFLTM	20	37.5	25.3
63	IILHENFDYDLLDNDISLLK	20	38.5	35.7
64	ASSTNLKDILADLIPKEQARIKTFRQQHGNTVVGQITVDM	39	39.0	18.2
65	HGVTVLTALGAILK	14	40.5	18.7
66	SLGQNPTEAELQDMINEVDADGNGTIDFPEFLTM	34	44.0	21.8
67	YLEFISEAIIHVLHSR	16	45.0	28.7
68	VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFK HLKTEAEM	55	45.6	25.5
69	SLGQNPTEAELQDMINEVDADGNGTIDFPEFLTMMAR	37	45.8	22.1
70	FKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDAD GNGTIDFPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAE	93	48.9	44.6
71	KASEDLKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIK YLEFISEAIIHVLHSRHPGNFGADAQGAM	85	50.0	25.2

REFERENCES

1 T. Sasagawa, T. Okuyama and D. C. Teller, *J. Chromatogr.*, 240 (1982) 329.
2 D. Guo, C. T. Mant, A. K. Taneja, J. M. R. Parker and R. S. Hodges, *J. Chromatogr.*, 359 (1986) 499.
3 C. T. Mant, T. W. L. Burke, J. A. Black and R. S. Hodges, *J. Chromatogr.*, 458 (1988) 193.
4 M. Yvon, C. Chabanet and J. P. Pélissier, *Int. J. Pept. Protein Res.*, 34 (1989) 166.
5 J. L. Meek, *Proc. Natl. Acad. Sci. U.S.A.*, 77 (1980) 1632.
6 J. L. Meek and Z. L. Rossetti, *J. Chromatogr.*, 211 (1981) 15.
7 S. J. Su, B. Grego, B. Niven and M. T. W. Hearn, *J. Liq. Chromatogr.*, 4 (1981) 1745.
8 K. J. Wilson, A. Honegger, R. P. Stötzel and G. J. Hughes, *Biochem. J.*, 199 (1981) 31.
9 C. A. Browne, H. P. J. Bennett and S. Solomon, *Anal. Biochem.*, 124 (1982) 201.
10 T. Sasagawa, L. H. Ericsson, D. C. Teller, K. Titani and K. A. Walsh, *J. Chromatogr.*, 307 (1984) 29.
11 S. Y. M. Law, A. K. Taneja and R. S. Hodges, *J. Chromatogr.*, 317 (1984) 129.
12 C. T. Mant, N. E. Zhou and R. S. Hodges, *J. Chromatogr.*, 476 (1989) 363.
13 N. E. Zhou, C. T. Mant and R. S. Hodges, *Pept. Res.*, 3 (1990) 8.
14 D. Le Bars and J. C. Gripon, *J. Dairy Res.*, 56 (1989) 817.
15 M. Dalgalarrondo, J. M. Chobert, E. Dufour, C. Bertrandharb, J. P. Dumont and T. Haertle, *Milchwissenschaft*, 45 (1990) 212.
16 M. M. Cals, P. Mailliart, G. Brignon, P. Anglade and B. Ribadeau-Dumas, *Eur. J. Biochem.*, 198 (1991) 733.
17 C. Chotia, *J. Mol. Biol.*, 105 (1976) 1.
18 J. Janin, *Nature (London)*, 277 (1979) 491.
19 S. Huet and A. Messean, in F. De Antoni, N. Lauro and A. Rizzi (Editors), *Proceedings in Computational Statistics*, Physica Verlag, Heidelberg, Vienna, 1986, p. 326.