Journal of Chromatography, 632 (1993) 11-18
Elsevier Science Publishers B.V.. Amsterdam

CHROMSYMP. 2601

High-performance liquid chromatography of amino acids, peptides and proteins

CXXII[☆]. Application of experimentally derived retention coefficients to the prediction of peptide retention times: studies with myohemerythrin

M. C. J. Wilce, M. I. Aguilar and M. T. W. Hearn

Department of Biochemistry and Centre for Bioprocess Technology, Monash University, Clayton, Victoria 3168 (Australia)

ABSTRACT

Amino acid retention coefficients were derived from the experimental retention data of 118 overlapping peptide heptamers related to the primary amino acid sequence of myohemerythrin. Individual retention coefficient values for each amino acid were derived by a multiple linear regression matrix approach. Retention data were derived for five different experimental conditions including different organic modifiers (acetonitrile, methanol or 2-propanol), different mobile phase additives (trifluoroacetic acid or potassium phosphate) and different silica-based stationary phase ligands (octadecyl or phenyl groups). A high degree of correlation was observed between these experimentally derived amino acid coefficients (EXP) and the coefficients (LIT) which we recently reported derived from the retention data of over 2000 peptides [M. C. J. Wilce ***et al., J. Chromatogr., 536 (1991) 165*** and 548 (1991) 105]. These results demonstrated that the LIT and EXP coefficients can be used for the prediction of the retention of any peptide set. The effect of peptide length was also further investigated and the correlation results demonstrated the importance of peptide flexibility on the final value of the amino acid coefficient.

INTRODUCTION

Reversed-phase high-performance liquid chromatography (RP-HPLC) continues to provide a very powerful technique for the analysis and purification of peptides and proteins. In addition, significant information on the physicochemical basis of the interaction between peptides and proteins and the stationary phase ligand has been derived using RP-HPLC [1–5]. However, fully developed mechanistic models are not yet available to describe, in molecular terms, the interactive processes that occur between the stationary phase, the mobile phase and the peptide or protein solute. The development of such models requires large and systematic data bases and associated structure-retention relationships to be determined. In addition, the influence of chromatographic parameters on the secondary and tertiary structure of the solutes also needs to be addressed. As part of our on-going studies into the mechanism of peptides and protein retention in RP-HPLC, we have previously derived individual amino acid group retention coefficients (GRCs) using a multiple linear regression analysis approach [6]. Several sets of retention coefficients have been re-

Correspondence to: Professor M. T. W. Hearn, Department of Biochemistry and Centre for Bioprocess Technology, Monash University, Clayton, Victoria 3 168, Australia.

[☆] For part CXXI, see ref. 15

ported which have been derived by a number of different methods (e.g. refs. 779). Our matrix-based approach allows GRCs to be derived from a statistically large peptide data set and also allows a more detailed statistical analysis of the amino acid GRCs to be carried out. This analysis includes (1) assessment of the relationship between the variability of the GRCs for a particular amino acid using different experimental conditions, and (2) the co-correlation of the GRCs of different amino acid residues in different sequence arrangements. Categorisation of the peptide retention behaviour allowed the influence of various chromatographic parameters on the GRCs to be evaluated. In these earlier studies, the various chromatographic systems which we examined included three different chemically bonded II-alkylsilica stationary phases, namely, RP-C_{18}, RP-C_8 and RP-C_4 and two different organic solvents in the mobile phase, namely, acetonitrile and 2-propanol-acetonitrile. It was found that the C_{18}TA[C_{18} stationary phase with trifluoroacetic acid (TFA)–acetonitrile mobile phase] and C_8(C_8 stationary phase with TFA-acetonitrile mobile phase) sets of GRCs had a high degree of correlation while the C_4 (C_4 stationary phase with TFA–acetonitrile mobile phase) and the C_{18}TPA(C_{18} stationary phase with TFA-acetonitrile-2-propanol mobile phase) scales were significantly different to all three scales. Principal component analysis of these four scales in conjunction with fourteen physicochemical descriptors of the amino acid side-chains was also carried out [10]. The results demonstrated a negative correlation between the C_{18}TA GRCs and amino acid parameters that describe electronic characteristics. Conversely, there was a positive correlation between the C_{18}TPA scale and amino acid parameters related to steric and volume characteristics. Derivation of GRCs using multiple linear regression procedures provides a general approach to quantitating the relative propensity of a particular amino acid to interact with a surface of defined ligand structure in different solvent environments. In order to further validate the use of the GRCs in the prediction of peptide retention behaviour, five new sets of amino acid GRCs were derived from the experimental retention data of a series of peptides related to the amino acid sequence of myohemerythrin. The peptides represent overlapping heptamers of the entire sequence of myohemerythrin. The new amino acid GRCs were then compared statistically with our previously published GRCs. The present study provides further validation of the general utility of these literature GRCs.

EXPERIMENTAL

*Calculation of **amino** acid retention coefficients*

The amino acid LIT GRCs were derived from the retention data of over 2000 peptides as previously described using multiple linear regression analyses [6]. An additional data base was also created using the experimental retention data of 118 synthetic peptides from which the EXP GRCs were derived. These peptides comprised all of the overlapping heptamers from the sequence of the protein myohemerythrin from the marine worm *Themiste zosteriola* shown in Fig. 1. Six reference peptides were also included in the analysis. All peptides were synthesised by the pin method of Maeji *et al.* [11] and were supplied by Chiron Mimotopes (Clayton, Australia). The peptides were synthesised using a cleavable linker that forms an N-terminal diketopiperazine derivative with a free carboxylic acid at the C-terminus. All peptides were analysed by positive-ion fast atom bombardment mass spectrometry and amino acid analysis. The derivation of the EXP GRCs employed the matrix method with Gaussian pivoting as described previously [6]. All statistical analyses were performed with the SPSS package on the Monash University VAX computer.

Reagents

Acetonitrile, methanol and 2-propanol were ChromAR HPLC grade from Mallinckrodt Australia (Melbourne, Australia). TFA was obtained from Pierce (Rockford, IL, USA). Water was quartz-distilled and deionised using a Milli-Q water purification system (Millipore, Bedford, MA, USA). Potassium dihydrogenphosphate (AnalaR grade) was

```
RESIDUE  NUMBER

1        10        20        30        40        50
|________|_________|_________|_________|_________|
GWEIPEPYVWDESFRVFYEQLDEEHKKIFKGIFDCIRDNSAPNLATLVKV

TTNHFTHEEAMMDAAKYSEVVPHKKMHKDFLEKIGGLSAPVDAKNVDYCK

EWLVNHIKGTDFKYKGKL
```

Fig. 1. Amino acid sequence of myohemerythrin.

purchased from BDH (Kilsyth. Australia). All mobile phases were filtered prior to use with a Millipore Durapore filter and then further degassed by helium sparging for 10 min.

Apparatus

Peptide retention times were measured using a Hewlett-Packard 1090 liquid chromatograph consisting of a DR5 solvent delivery system, a thermostatically controlled oven set at 37°C, an autosampler and a diode-array detector with an HP79994A workstation coupled to a Thinkjet printer and an HP7470 plotter. All data were stored on the HP computer before transfer to either an IBM PC or the Monash University VAX computer for further analysis.

Four different mobile phase systems were used for the derivation of the experimental retention data base. These included: solvent 1 :A, 0.1% TFA; B, 0.09% TFA-50% acetonitrile; solvent 2: A, 0.1% TFA; B, 0.09% TFA-50% methanol; solvent 3: A, 0.1% TFA; B, 0.09% TFA-50% 2-propanol; solvent 4: A, 25 m*M* KH_2PO_4; B, 35 m*M* KH_2PO_4–50% acetonitrile. The flow-rate was 1 ml/min except for solvent 3 where 0.7 ml/min was used due to a high pressure drop. Two different alkyl ligand types were used: column 1: n-octadecylsilane (RP-C_{18}, 15 cm × 4.6 mm I.D., 5 μm, 75 nm pore size. DuPont Zorbax); column 2: phenylsilane (RP-phenyl, 15 cm × 4.6 mm I.D.. 5 μm, 75 nm pore size, DuPont Zorbax). Table I lists the codes and chromatographic conditions used for the derivation of the GRCs.

TABLE I

CODES AND CHROMATOGRAPHIC CONDITIONS USED TO DERIVE THE EXP COEFFICIENT SET

Code	Chromatographic conditions
C_{18}TA	RP-C,, sorbent, TFA-acetonitrile mobile phase
C_{18}KP	RP-C, $_8$ sorbent, KH_2PO_4–acetonitrile mobile phase
C_{18}TM	RP-C, $_8$ sorbent, TFA-methanol mobile phase
C_{18}TP	RP-C,, sorbent, TFA-2-propanol mobile phase
PHETA	RP-phenyl sorbent, TFA-acetonitrile mobile phase

RESULTS AND DISCUSSION

Influence of ligand type and mobile phase on jive sets of experimental retention coefficients

According to the solvophobic theory [12], peptide retention in RP-HPLC is largely caused by entropic expulsion of the peptide from the polar mobile phase, accompanied by their adsorption onto the non-polar stationary phase. A practical consequence of this phenomenon is that, in the absence of significant conformational effects, peptide retention can be correlated with the hydrophobicity of the peptide contact area as revealed by the GRCs of the constituent amino acids. These observations thus have important consequences in the development of protocols for the optimisation of retention times for peptides of known composition. To extend the observations based on our recently derived literature retention coefficient (LIT) sets [6,10], 112 heptapeptides which represent the entire group of overlapping peptide heptamers derived from the amino acid sequence of myohemerythrin were eluted under five different chromatographic conditions as described in the Experimental section. Myohemerythrin is the subject of related studies on epitope mapping and protein surface analysis and seven amino acid residues represent the minimum sequence required for epitope recognition by antibodies. Table II lists the values for each of the five experimental (EXP) GRCs and the correlation coefficients for the statistical comparisons of these scales are listed in Table III. It is evident that all five of the scale exhibited a high degree of similarity. The highest correlation (R = 1 .OO) was observed for the comparison between the C_{18}TA and the C_{18}TP sets while the poorest correlation (R=0.80) was observed between the C_{18}KP and the C_{18}TM sets. With 18 degrees of freedom, the probability of observing $R >$ 0.80 is less than 0.05% (P < 0.001). The similarity of the five EXP GRC scales is in contrast to the results observed for the previously published LIT GRCs [6]. It was found that a C_{18} or a C_8 ligand and a TFA-acetonitrile mobile phase resulted in GRCs that were very similar but which were both significantly different to GRCs derived using a C_4 ligand with the same solvent or a C_{18} ligand with a TFA-acetonitrile-propanol-based mobile phase. In the present study the experimental parameters that were systematically changed were the nature of the

TABLE 11

EXPERIMENTALLY DERIVED AMINO ACID GRCs (mm)

Amino acid	C_{18}TA	C_{18}KP	C_{18}TM	C_{18}TP	PHETA
Alanine	1.70	2.91	1.76	1.36	1.24
Cysteine	0.49	0.19	1.66	0.33	-1.41
Aspartic acid	1.10	-1.47	0.31	0.71	1.01
Glutamic acid	0.79	0.00	0.03	0.24	0.69
Phenylalanine	8.79	9.14	10.66	7.21	9.18
Glycine	0.39	0.36	0.00	0.00	0.00
Histidine	0.62	2.21	-0.46	0.12	0.23
Isoleucine	8.35	9.80	11.76	6.97	8.29
Lysine	0.00	3.36	-1.41	-0.54	-0.24
Leucine	9.51	7.88	14.69	7.56	8.88
Methionine	2.60	3.90	1.63	1.92	2.31
Asparagine	-0.02	1.31	-2.01	-0.40	-0.26
Proline	2.79	3.50	4.21	2.08	2.09
Glutamine	-0.66	3.09	-7.52	-0.76	-1.28
Arginine	2.36	4.61	1.62	1.66	3.19
Serine	0.27	1.66	2.64	-0.41	0.10
Threonine	1.80	2.40	2.10	0.97	1.40
Valine	4.93	4.97	6.03	3.80	4.72
Tryptophan	9.75	9.99	13.30	7.47	10.54
Tyrosine	6.14	4.93	9.01	4.06	5.95

organic modifier, the pH of the mobile phase and the nature of the stationary phase ligand which all exerted little effect on the experimentally measured GRCs for each amino acid residue.

There are a number of explanations for the differences in the statistical comparisons of the LIT GRCs and the similarity of the EXP GRCs. Firstly, large differences in the frequency distribution of the amino acids within the respective data base of peptides shown for the LIT GRC set (Fig. 2) and the EXP GRC set (Fig. 3) can strongly influence the outcome of the multiple linear regression used to calculate the GRCs. However, comparison of the relative frequency distribution for the LIT GRCs shown in Fig. 2 demonstrated that the frequency of each amino acid residue is not significantly different in the four data bases (Table IV). As a consequence, the amino acid frequency distribution would therefore not contribute to the differences between the LIT GRCs. The second parameter which may affect the calculation of the GRCs is the sample size of the experimental peptide data set. It has been previously established [6] that a minimum of 100 peptide retention times are required to establish statistically consistent sets of GRCs. The experimental peptide data set for the myohemerythrin heptamers consisted of 112 peptides which thus satisfies the sample size requirements. The third parameter which may also influence the resultant GRC values is the specific sequences of the respective peptide data bases.

TABLE III

CORRELATION COEFFICIENTS FOR LINEAR REGRESSION ANALYSIS OF 5 EXP GRC SETS ($n = 18, r_{95\%} = 0.44$)

	C_{18}KP	C_{18}TM	C_{18}TP	PHETA
C_{18}TA	0.90	0.97	1.oo	0.99
C_{18}KP		0.80	0.90	0.91
C_{18}TM			0.96	0.94
C_{18}TP				0.98

The origin of the peptides represents the most signicant differences between the 2 data sets. The LIT peptides were derived from enzymatic and chemical cleavage of a large range of proteins while the EXP peptides of the present study were derived from a single protein and were synthesised as overlapping peptides as depicted in Fig. 4. Peptide 1 consisted of the first seven residues of the myohemerythrin sequence while peptide 2 comprised residues 2-8. The remaining peptides were composed in a similar manner by proceeding along the sequence 1

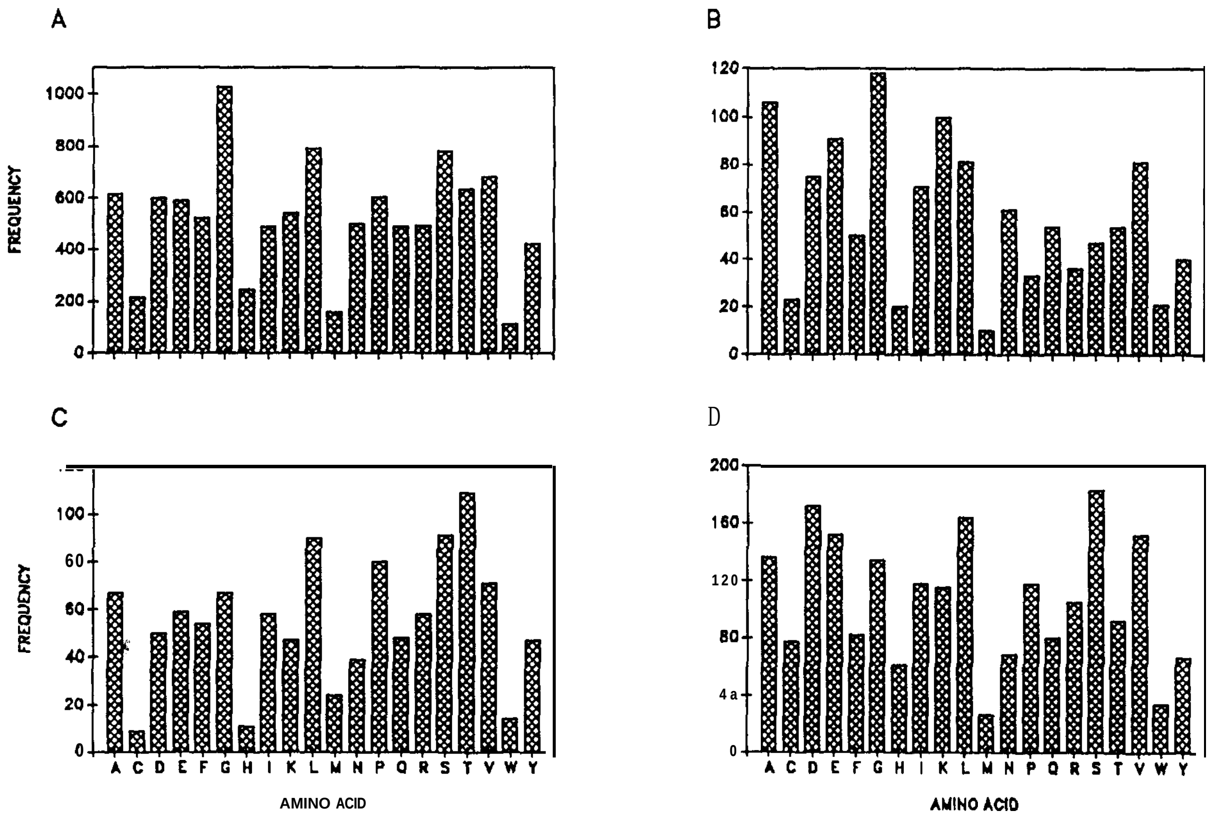

Fig. 2. Amino acid frequency distribution in each of the four LIT peptide data sets. The correlation coefficients for linear comparison between each data set are listed in Table IV.

residue at a time. While there are a large number of peptides generated with this procedure the variations in the molecular environment surrounding each amino acid residue is greatly reduced. For example, as illustrated in Fig. 4, glutamic acid-6 is flanked by 2 proline residues in 6 peptides. Thus, while there are more than 100 peptides in the data set, the number of amino acid environmental categories sampled by the non-polar ligands is reduced by a factor of 6 compared to the peptides selected from random sequences. The chromatographic results suggest that the range of amino acid environments associated with the synthetic heptapeptides does not allow differentiation of the molecular dynamics of each peptide during interaction with the stationary phase ligands in each of the 5 different chromatographic conditions.

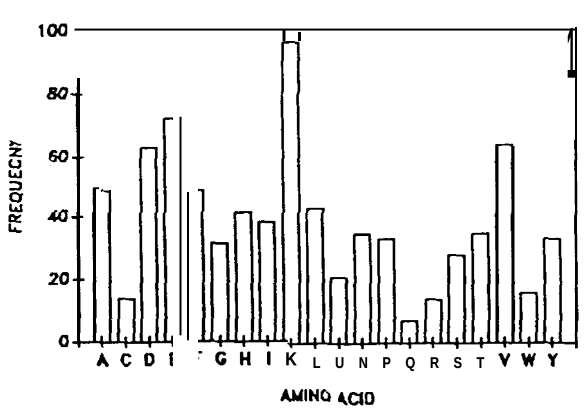

Fig. 3. Amino acid frequency distribution in the EXP peptide data sets.

Influence of peptide length on experimental coefficients

It is well established that peptides larger than 15–20 residues are often eluted with retention times

TABLE IV

CORRELATION COEFFICIENTS FOR THE LINEAR COMPARISON OF THE AMINO ACID FREQUENCIES IN EACH OF THE LIT PEPTIDE DATA SETS (n = 18. $r_{95\%}$ = 0.44)

	$C_{18}TA$	C_8TA	C_4TA
C_8TA	0.65		
C_4TA	0.75	0.44	
$C_{18}TPA$	0.69	0.66	0.66

which are different to the predicted times based on the summation of their GRCs. The differences in the chain lengths of the peptides used for the LIT and EXP data bases may therefore represent an additional factor which contributes to the similarity of the 5 EXP coefficient sets. The EXP peptides were all 7 residues in length while the LIT peptides ranged in length from 2 to 30 amino acids. It is generally considered that as peptide length increases, there is a greater probability for chromatographic ligands to induce transitions from a disorganised random coil structure to a stabilised secondary structure. Conformational rearrangements of peptides have been monitored in RP-HPLC through measurements of the change in chromatographic contact areas and affinity and also through changes in the enthalpy and entropy associated with the binding of the peptide to the stationary phase ligands [1,2]. Thus, the hydrophobic nature of the reversed-phase ligands can induce the formation of secondary structure, such as an amphipathic cc-helix [13] and this process will prevent the amino acid residues which are oriented away from the stationary phase surface to contribute to the peptide retention. The extent of conformational flexibility of the peptides used for the retention data base will thus strongly influence the calculation of the GRCs.

The effect of peptide length on the GRCs was therefore assessed by performing additional linear regression analysis between the EXP-C_{18}TA and LIT-C_{18}TA GRC scales according to various peptide length criteria. In the first instance. 8 groups of peptides of different chain lengths were established using the LIT-C_{18}TA GRC scale which included peptides containing 2–30, 2 20. 2-15, 2–10, 2- 8, 2–7, 2–5 and 2-4 amino acid residues in length. The correlation coefficient from the linear regression analysis of these restricted LIT GRCs with the EXP-C_{18}TA scales are shown in Fig. 5. The highest correlation for these comparisons was obtained for the LIT sub-group containing peptides up to 15 amino acid residues. The average length of the LIT peptides in the group of peptides with 2- 15 residues was 7.2 residues. It can therefore be anticipated that the LIT peptide group with 2215 amino acid residues should exhibit a high degree of correlation with the EXP-C_{18}TA scale which were derived from peptides of 7 amino acid residues. The LIT peptide groups consisting of shorter peptides, i.e. 2-10. 2–8, 2 -7. 2- 5 and 224 amino acid residues, exhibited lower correlation with the EXP-C_{18}TA scale. This divergence may be due to exaggerated N- and C-terminal effects associated with the LIT peptide sets due to their origin. For example. many peptides were tryptic peptides which contain a basic amino acid residue (arginine or lysine) at the C-terminal

```
    MYOHEMERYTHRIN SEQUENCE

    G W E I P E P Y V W D E S F R V F Y ........ F K Y K G K L

    SYNTHETIC PEPTIDES

1.  G W E I P E P
2.    W E I P E P Y
3 .     E I P E P Y V
4.        I P E P Y V W
5.          P E P Y V W D

                          .
111.                                        D F K Y K G K
112.                                          F K Y K G K L
```

Fig. 4. Diagram illustrating the method used to generate overlapping heptamers of the amino acid sequence of myohemerythrin.

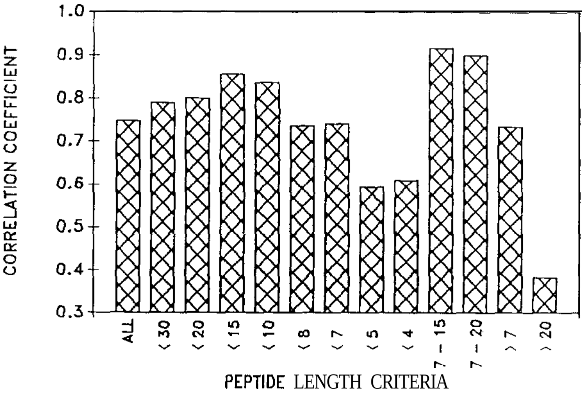

Fig. 5. Correlation coefficients generated for the comparison of the EXP C_{18}TA GRCs and LIT GRCs with specified peptide length criteria.

position. Similar sequence-dependent retention behaviour has been noted by Guo *et al.* [7]. However, in the derivation of both the LIT and EXP GRCs, the influence of the end-groups was minimised through the calculation of separate coefficients for the N- and C-terminal end-groups. Comparison of the correlation coefficients for the peptide groups consisting of longer peptides, *i.e.* peptides containing 2-20 or 2-30 residues, revealed that the *r* value for these groups was similar to that of the peptide group containing only l-15 amino acid residues. In these cases, a large number of smaller peptides which are common to the sub-groups with 2-15, 2-20 and 2-30 amino acid residues will exert the dominant influence on the final GRC values.

Three further cluster groups were also defined with the LIT peptides comprising 7-l 5 residues and all peptides with more than 7 or more than 20 amino acid residues. The maximum correlation with the EXP GRC sets was found for the LIT group derived from peptides between 7 and 15 residues in length and thus represented the highest correlation observed for the comparisons of all peptide subgroups. The low correlation coefficient (0.38) for the comparison of the LIT group containing peptides greater than 19 residues in length with the EXP-C$_{18}$TA set clearly demonstrates the influence of the peptide length on the GRCs. Overall, the results of the linear regressions of the peptide subgroups indicates that peptide length is an important factor controlling the differences within LIT GRCs and the constancy of the EXP GRCs. It should thus be possible to selectively improve the degree of linear correlation by restricting the length of the peptides in the original calculation of the LIT GRCs. This was performed using the LIT-C$_{18}$TA and the LIT-C$_{18}$TPA GRC sales. The original correlation coefficient between both scales has been calculated as 0.44 [6]. If the LIT-C$_{18}$TA data base is restricted to include only peptides containing up to 7 residues in length the correlation increases to 0.82. It can be concluded from these results that selectivity changes and reversals which are observed in the elution of peptides in RP-HPLC under different chromatographic conditions are not only due to changes in the intrinsic hydrophobicity of each amino acid. The relative retention of a peptide will also be strongly influenced by the conformational flexibility of the peptide which in turn controls the portion of the peptide that interacts with the stationary phase surface.

Prediction of peptide retention times

The ability of the EXP and the LIT scales to predict the retention times of the 112 peptides was examined. Fig. 6 shows the retention time profile predicted according to the summation of the LIT-C$_{18}$TA coefficient scale (A), the EXP-C$_{18}$TA coefficient scale (B) and the experimentally observed retention times for these peptides under the same experimental conditions *(i.e.* TFA-acetonitrile) (C). A correlation coefficient of 0.98 was observed from linear regression comparison between the EXP-C$_{18}$TA GRCs and the experimentally observed retention times for the 118 synthetic peptides. Comparison of the retention times predicted by the LIT-C$_{18}$TA coefficient scale and the experimentally ob-

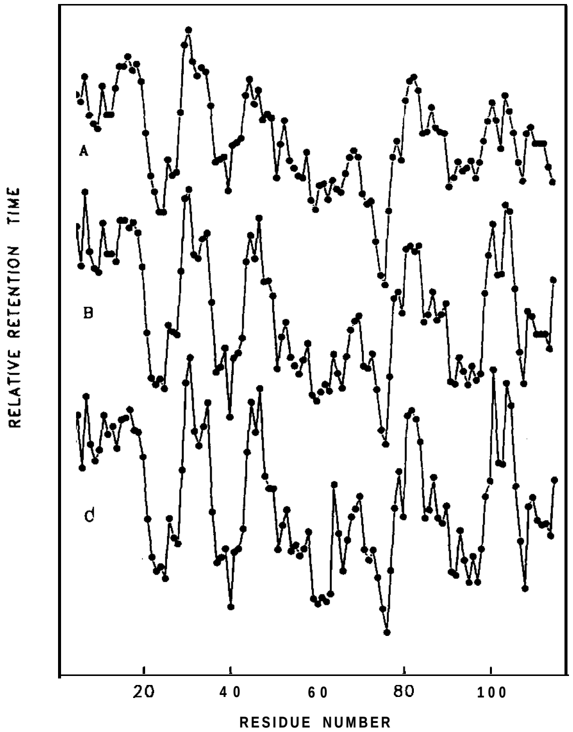

Fig. 6. Retention profile of overlapping heptamers derived from the amino acid sequence of myohemerythrin. (A) Retention times (min) predicted with the LIT GRCs; (B) retention times predicted with the EXP GRCs; (C) experimentally observed retention times, which ranged between 10 and 45 min. The linear relationship between panel A and panel C is given by eqn. 1.

served times gave a correlation coefficient of 0.91 for the same 118 synthetic peptides. In order to accommodate differences in column configuration and experimental parameters, the retention times predicted by the LIT-C_{18}TA scale required linear adjustment to correlate in magnitude with the experimentally observed retention times according to the following relationship:

observed retention time =
(LIT predicted retention time) × 1.46 + 14.02 (1)

The high degree of linear correlation between the retention times predicted using the LIT GRCs and the experimentally observed retention times demonstrates the general utility of the LIT GRCs for the prediction of relative retention times of peptide solutes not originally used as part of the data base for the derivation of the GRCs. The ability of both the LIT GRCs and the EXP GRCs to predict the hydrophobic surface accessibilities of unrelated sets of peptides and proteins is documented elsewhere [14].

ACKNOWLEDGEMENTS

These investigations were supported by grants from the Australian Research Council and the National Health and Medical Research Council.

REFERENCES

1 A. W. Purcell, M. 1. Aguilar and M. T. W. Hearn, ***J. Chromutogr.***, 476 (1989) 125.

2 A. W. Purcell, M. I. Aguilar and M. T. W. Hearn. ***J. Chromutogr.***, 593 (1992) 103.

3 M. Kunitani. D. Johnson and L. R. Snyder, ***J. Chromatogr.***, 371 (1986) 313.

4 M. L. Heinitz, E. Flanigan, R. C. Orlowski and F. E. Regnier, ***J. Chromatogr.***, 443 (1988) 229.

5 M. T. W. Hearn and M. I. Aguilar, ***J. Chromatogr.***, 392 (1987) 33.

6 M. C. J. Wilce, M. I. Aguilar and M. T. W. Hearn. ***J. Chromatogr.***, 536 (1991) 165.

7 D. Guo. C. T. Mant, A. K. Taneja, J. M. R. Parker and R. S. Hodges, ***J. Chromutogr.. 359 (1986) 499.***

8 J. L. Meek and Z. L. Rosetti, ***J. Chromatogr.***, ***211*** (1981) 15.

9 C. Chabanet and M. Yvon, ***J. Chromatogr.***, ***599*** (1992) 21 I.

10 M. C. J. Wilce, M. 1. Aguilar and M. T. W. Hearn. ***J. Chromatogr.***, ***548 (1991)*** 105.

11 N. J. Maeji, A. M. Bray and H. M. Geysen. ***J. Immunol. Methods.***, 134, (1990) 23.

12 Cs. Horvath, W. R. Melander and I. Molnar, ***J. Chromatogr.***, ***125 (1976) 129.***

13 ***N.*** E. Zhou, C. T. Mant and R. S. Hodges, *Peptide Res.*, 3 (1990) 8.

14 K. L. Spiers, M. I. Aguilar and M. T. W. Hearn, in preparation.

15 M. Zacheriou and M. T. W. Hearn, ***J. Chromatogr.***, ***599 (1992) 171.***