Journal of Chromatography A, 661 (1994) 61–75
Elsevier Science B.V., Amsterdam

CHROMSYMP. 2908

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

CXXXIII☆. Peak tracking of peptides in reversed-phase high-performance liquid chromatography

A.J. Round, M.I. Aguilar and M.T.W. Hearn*

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

ABSTRACT

A peak tracking algorithm for peptide analysis has been developed based on a normalised spectral overlay method which directly compares the UV spectra of any two chromatographic peaks. Additionally, the algorithm compares the spectrum of each peak in the first chromatogram with the spectra of every peak in the second chromatogram to determine the best cross-match. The sensitivity of the technique was further enhanced by incorporation of the primary and secondary derivative spectra for cross-match normalisation. The utility of the software was demonstrated by its application to the analysis of tryptic digests of porcine growth hormone. Peptide solutes could be identified and tracked in chromatograms generated with various column types, gradient times, mobile phase types and temperatures. These results therefore constitute the initial stages of development of a more robust approach to the optimisation of the resolution, detection and characterisation of peptides and proteins separated by HPLC techniques.

INTRODUCTION

Over the past 15 years high-performance liquid chromatography (HPLC) has emerged as a powerful tool for the analysis and purification of peptides and proteins [1]. However, it has only been during the last several years that strategies for systematic optimisation of HPLC separations of peptides have begun to be systematically investigated [2,3]. Reversed-phase HPLC (RP-HPLC) is the mode of chromatography which has been most extensively studied, and now represents the dominant technique for resolution of peptide samples. Various different strategies have been proposed to permit optimisation of the mobile phase composition for this mode of chromatography. These optimisation strategies can generally be grouped into two categories. The first category consists of interpretive optimisation methods, which base their predictions of the optimum mobile phase conditions on a model (or map) of the retention behaviour of the individual components in a mixture. In this approach, a limited set of "scouting" experiments are performed on a given sample under various chromatographic conditions, the resolution data fitted to a mathematical function by linear or non-linear regression techniques and a retention "map" is then generated to encompass the behaviour of the solutes under conditions not explicitly tested during the initial experiments [2–6]. The second category includes optimisation methods whereby an iterative search is per-

* Corresponding author.
☆ For Part CXXXII, see ref. 25.

0021-9673/94/$07.00
SSDI 0021-9673(93)E0874-T

formed. In these methods, the results from one experiment are used to predict the conditions for the next experiment (often using algorithms such as modified sequential simplex methods) [2,3,7–10]. The process is repeated until the optimum chromatographic conditions have been determined. These methods make no assumptions concerning the nature of the mathematical function used to interrogate the retention data and have the advantage that generally fewer experiments are required to locate the optimum chromatographic conditions.

In order to effectively utilise any of these optimisation methods, the location (but not necessarily the structural identity) of the solute components in successive chromatograms must be known. Thus, a fundamental requirement in the application of chromatographic optimisation methods to the characterisation of unknown solute mixtures is peak recognition. In order to quantitatively describe the influence of changes in the experimental chromatographic parameters on the retention of the individual solutes in a sample, the investigator must be able to identify and follow the relative movement of individual solute peaks as the experimental chromatographic conditions are varied. Since the actual identification of the solutes is not necessarily of immediate interest, but rather the determination of the relative location of individual peaks corresponding to the same solutes in two (or more) different chromatograms of the same mixture (as the chromatographic conditions are varied), this process is referred to as peak tracking.

With the advent of rapid-scanning photodiode array UV–Vis detectors, complete spectral information for any or all peaks in a chromatogram can now be acquired. Comparison of the spectral data from peaks in different chromatograms has great potential to facilitate peak identification and hence the development of HPLC optimisation systems. Algorithms for the numerical comparison of spectra have been successfully used in the past to distinguish between very similar compounds [11]. The present study is based on a modification of one such algorithm, utilising the spectral information from each peak in a chromatogram to perform normalised spectral overlay comparisons (NSOC) for all of the normal (zero-order) UV spectra as well as the first-order and second-order derivatives of these spectra. The algorithm developed has been validated by the matching between any two chromatograms (derived under different chromatographic conditions) the spectral absorbance of peaks derived from samples of a tryptic digest of recombinant porcine growth hormone, regardless of whether or not the compositional identity of the peaks is known. The software compares the spectra from each peak in the first chromatogram with every peak in the second chromatogram to determine the best match. In this way each peak from the first chromatogram can be assigned to its best matching peak in the second chromatogram.

EXPERIMENTAL/MATERIALS AND METHODS

Chemicals and solvents

Acetonitrile (MeCN), methanol (MeOH) and isopropanol (*i*-PrOH) were ChromAR HPLC grade from Mallinckrodt Australia (Melbourne, Australia); trifluoroacetic acid (TFA) was obtained from Pierce (Rockford, IL, USA). Water was quartz-distilled and deionised by passage through a Milli-Q water purification system (Millipore, Bedford, MA, USA).

Recombinant porcine growth hormone (Met-Asp-Gln-pGH, r-pGH) was obtained from American Cyanamid (Princeton, NJ, USA).

Dithiothreitol (DTT), iodoacetic acid (IAA) and N-tosyl-L-phenylalanine chloromethyl ketone (TPCK) trypsin were purchased from Sigma (St. Louis, MO, USA), Fluka (Buchs, Switzerland) and Worthington (Freehold, NJ, USA), respectively. All other reagents were analytical grade or the best available grade.

Tryptic digest of r-pGH

The tryptic digest of recombinant porcine growth hormone (r-pGH) was performed using the following method: 1 mg growth hormone was dissolved in 250 μl guanidine hydrochloride (GdHCl) buffer [6*M* GdHCl, 200 m*M* Tris, 2 m*M* ethylenediaminetetraacetic acid (EDTA), pH 8.0] and incubated at 37°C for 30 min. After cooling, DTT (1 mg/30 μl GdHCl buffer) was added and the solution flushed with a stream of

nitrogen gas. Following incubation at 37°C for 3 h the solution was allowed to cool, and IAA (1.9 mg/70 μl 1 *M* Tris–HCl, pH 8.0) was added. The mixture was then incubated in the dark at room temperature for 15 min, after which 10 μl mercaptoethanol was added, followed by 5 μl TPCK trypsin (10 μg/5 μl 1 *M* HCl). The protein was recovered by adding 3.15 ml methanol (chilled at −20°C), stored overnight at −20°C and centrifuged at 2000 *g* for 20 min at 4°C. The supernatant was poured off and the pellet resuspended in 400 μl chilled methanol. The suspension was recentrifuged as above and the pellet resuspended in 400 μl of fresh 100 m*M* NH_4HCO_3–2 mM $CaCl_2$ solution. An additional aliquot of TPCK trypsin was then added to the solution, which was then incubated for 24 h at room temperature. Digestion was stopped by acidification with 2 *M* HCl.

Reversed-phase high-performance liquid chromatography

Reversed-phase chromatographic analyses of r-pGH tryptic digests were performed using a Hewlett-Packard HP1090M HPLC system consisting of a DR5 solvent delivery system, a thermostatically controlled column compartment, an automated injection and sampling system and a HP1090 diode-array detector. This instrument was connected to a HP79994A ChemStation Analytical Workstation computer coupled to a ThinkJet printer and a HP7470 plotter.

Table I summarises the chromatographic conditions utilised in this study. Three linear gradient mobile phase systems were employed for the separation of the tryptic fragments (as listed in Table I). With each of these mobile phase systems, two types of ligands chemically bonded to the stationary phase packed into steel columns were used. The first was a 25 cm × 0.46 cm Bakerbond Analytical WidePore C_{18} reversed-phase column, and the second column was a 25 cm × 0.46 cm Bakerbond Analytical WidePore C_4 reversed-phase column (J.T. Baker, Phillipsburg, NJ, USA). Certain combinations of mobile and stationary phases created high back-pressures, necessitating the use of a range of solvent flow-rates, as illustrated in Table I.

Data processing

For all analyses, spectra were acquired at time intervals of 0.320 s over a wavelength range from 200 to 350 nm. Chromatographic peak spectral absorbances were also recorded at both 215 nm and 274 nm, with a reference wavelength of 350 nm in both cases. Raw data was stored on both a 20 MByte Hard and 1.44 MByte Floppy disks by the ChemStation for subsequent processing by the peak tracking software.

The peak tracking software was written using

TABLE I

CHROMATOGRAPHIC CONDITIONS USED TO ESTABLISH DATABASE OF CHROMATOGRAMS

Solvent system	Mobile phase solvents	Stationary phase[a] ligand	Flow-rate (ml/min)
1	(A) 0.1% TFA in water	C_4	1.0
	(B) 0.09% TFA 90% acetonitrile	C_{18}	1.0
2	(A) 0.1% TFA in water	C_4	0.8
	(B) 0.09% TFA 90% methanol	C_{18}	0.6
3	(A) 0.1% TFA in water	C_4	0.6
	(B) 0.09% TFA 90% *i*-propanol	C_{18}	0.4

[a] A linear gradient from mobile phase A to mobile phase B at different gradient times (30, 45, 60, 90, 120 min) and temperatures (25, 37, 50, 65, 80°C) was employed with the two RP-sorbents.

the high-level PASCAL interpreter command language available on the ChemStation. The software functions as a "stand-alone" program which allows access to, and manipulation of, the chromatographic and spectroscopic data previously stored on either the floppy or the hard disk media.

Chromatographic database

The initial task in the development of our new optimisation procedures and peak tracking algorithms was the creation of a large database of chromatograms and their associated spectra. This database consists of chromatograms of tryptic digests of r-pGH run under a wide variety of chromatographic conditions. These conditions consisted of the mobile phase and stationary phase ligand systems listed in Table I, with separations performed at 5 different linear gradient times (30, 45, 60, 90, 120 min) and 5 different temperatures (25, 37, 50, 65, 80°C). Together these combinations create 150 different chromatographic conditions. As each tryptic digest produces at least 20 peptide fragments, this represents a spectral database of several thousand solute spectra. This database was used to extensively test and validate the peak tracking software.

RESULTS AND DISCUSSION

(A) Development of the peak tracking software

Peak tracking software description. A number of methods have been previously described which perform peak tracking based on analysis of relative peak areas [5,6,12], wavelength ratios [13] or retention values [14]. These methods have limitations, especially when peaks overlap or samples vary in the relative concentration of the individual components. The introduction of linear photodiode array detection (PDAD) for HPLC instrumentation increased the potential for peak tracking considerably, since full spectra for each solute peak in a chromatogram can be compared. Drouen *et al.* [15] performed comparisons of spectra "by eye", and reported difficulties in distinguishing between similar spectra. This is not surprising since a visual interpretation of pattern similarities in spectra would be a highly subjective exercise. More recently, peak tracking methods based on combinations of spectral recognition and peak areas have been investigated [16,17]. These methods require prior knowledge of the spectrum of an individual solute obtained under analogous conditions and/or the use of sophisticated computer software such as neural networks.

We describe here a strategy based on an objective comparison of spectra by computer software. The UV spectra of peptides and proteins are characteristic of their constituent amino acids, especially with regard to their aromatic amino acids (phenylalanine, tyrosine and tryptophan). These amino acids have absorption maxima between 250 to 300 nm, but the spectra are rather broad and overlapping. It is therefore difficult to distinguish between peptides containing these aromatic amino acids simply from their zero-order derivative spectra. However, these problems can be overcome by derivatisation of the zero-order spectra which increases the resolution between spectral differences. Second-order derivative spectral analysis as a static method (derivative spectroscopy) has been widely used to assess solvent accessibility and conformational information for peptides and proteins containing aromatic amino acids [18]. In particular, the second-order derivative of a spectrum transforms peaks and shoulders of the corresponding zero-order derivative spectrum into well defined maxima/minima. The enhanced resolution between different spectra after derivatisation forms the basis of our new peak tracking software.

The flowchart shown in Fig. 1 provides a simplified explanation of the steps carried out by this peak tracking algorithm. Basically, the algorithm performs a detailed analysis and cross-correlation comparison of the zeroth-order, first-order and second-order derivatives of the UV spectra obtained from each solute peak. The software takes a chromatogram and compares the spectra (*i.e.* zero-, first- and second-order derivative spectra) of the first peak in that chromatogram with the spectra (*i.e.* zero-, first- and second-order derivative spectra) of every peak in a second chromatogram to determine the best matching correlation. The software then takes the second peak in the first chromatogram and compares its spectra to the spectra of every

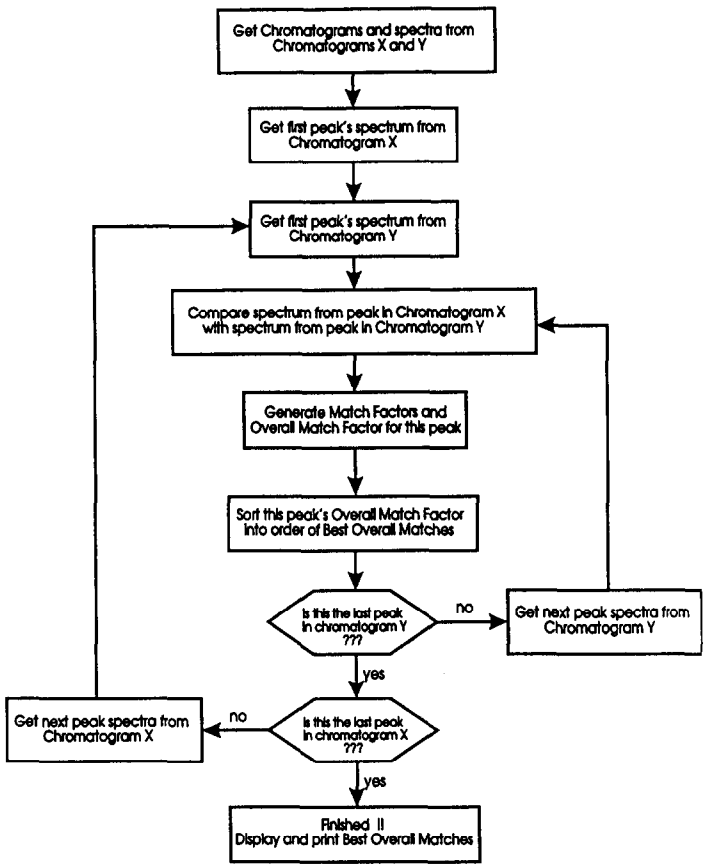

Fig. 1. Flowchart illustrating peak tracking procedure.

peak in the second chromatogram, again finding the best matching correlation. The process is repeated until each peak from the first chromatogram is assigned to its best matching peak in the second chromatogram.

To perform this matching procedure, the spectra are subjected to a process known as normalised spectral overlay comparison. The normalised spectral overlay comparison process is based on the numerical point-by-point comparison of two UV spectra by the COMPARE command implemented on the ChemStation [19]. The procedure is illustrated in Fig. 2 with the comparison of two typical spectra. Fig. 2a shows the two spectra to be compared. These spectra are first normalised at the point of maximum absorbance and then digitally superimposed. The absorbance values for spectrum 2 are then plotted against the corresponding absorbance values for spectrum 1 at each wavelength, as shown in Fig. 2b (solid line). A linear regression is then applied to the resulting scatterplot. The regression line calculated is shown in Fig. 2b (dashed line). The square of the correlation coefficient derived from this linear regression is defined as the match factor for the two spectra according to the following expression,

$$\text{Match factor} = 1000 \cdot r^2 = \frac{1000[\Sigma xy - (\Sigma x \, \Sigma y)]^2}{\left[\Sigma x^2 - \left(\frac{\Sigma x \, \Sigma x}{n}\right)\right]\left[\Sigma y^2 - \left(\frac{\Sigma y \, \Sigma y}{n}\right)\right]} \quad (1)$$

The x and y values are the measured absorbances in the first and second spectrum respectively at the same wavelength, n is the number of data points used in the comparison (typically >100), Σ is the sum of the data, and r^2 is the square of the linear regression correlation coefficient.

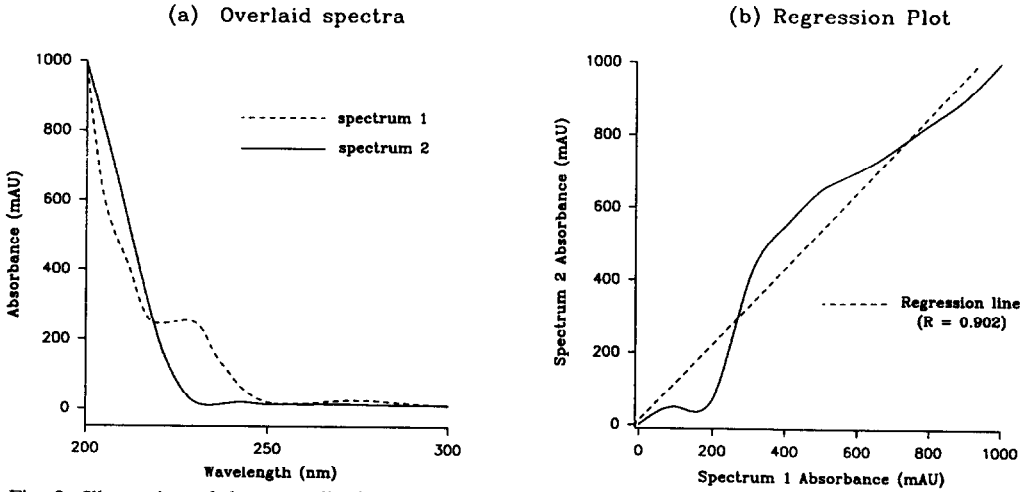

Fig. 2. Illustration of the normalised spectral overlay comparison procedure.

Individual match factors are determined by this overlay comparison process independently for each of the zero-order, first-order and second-order derivative spectra. An *overall match factor (O.M.F.)* concept was then applied to each peak comparison based on the optimum combination of each of the zero-, first- and second-order derivative UV spectra match factor scores. This combination enhanced the accuracy of the peak tracking procedure compared to the use of zero-order derivative spectra alone. The combination of match factors required to maximise the accuracy of peak tracking was initially established by statistical examination of the individual match factor scores for each of the zero-order, first-order and second order derivative spectra. Appropriate weighting factors for the zero-order, first-order and second-order derivative spectra were derived from this analysis.

Table II illustrates a typical example of the

TABLE II

TYPICAL EXAMPLE OF THE RANGE OF INDIVIDUAL MATCH FACTOR SCORES OBTAINED WHEN PEAK TRACKING IS PERFORMED BETWEEN TWO CHROMATOGRAMS DESIGNATED M AND N

The data below illustrate only a small portion of the total data analysed and show only five match factor scores for each of the two peaks investigated.

Peak number from chromatogram M	Peak number from chromatogram N	Zero-order derivative spectra match factor	First-order derivative spectra match factor	Second-order derivative spectra match factor	Correct matching peak
1	3	999.473	986.344	690.558	yes
1	6	999.245	987.802	615.727	
1	5	988.271	862.140	239.036	
1	2	976.217	962.382	642.526	
1	4	986.276	862.656	223.382	
2	2	998.889	982.509	742.898	yes
2	9	997.421	970.649	623.631	
2	8	992.723	967.411	658.823	
2	7	991.461	977.483	773.512	
2	10	993.149	929.182	680.812	

range of individual match factor scores obtained when peak matching is performed between two chromatograms (designated as chromatograms M and N) using unweighted zero-order, first-order and second-order derivative spectra. It can be seen in Table II that peaks 3 and 6 of chromatogram M have very similar zero-order derivative spectra match factor scores (999.473 and 999.245, respectively) when their spectra are compared to the spectrum of peak 1 from chromatogram N. Using just the zero-order derivative spectra match factors, it could be concluded that peak 3 is a slightly better match for peak 1 from chromatogram M than peak 6, although the confidence limits for this assignment would be low since the two scores are so close. It can be noted at this point that the correct matching peak for peak 1 from chromatogram M, as determined by independent structural analysis, is peak 3 from chromatogram N. An O.M.F. score based on just the first-order derivative spectra match factor scores would have incorrectly matched peak 6 to peak 1. Similarly, an unweighted combination of both the zero-order and first-order derivative spectra match factors would also have resulted in an incorrect match. Examination of the second-order derivative spectra match factors for these peaks, reveals that peak 3 has a significantly higher score than peak 6 and an O.M.F. score based solely on the second-order derivative spectra match factor score would yield a correct match. However, this situation does not always arise, as can be seen in the analysis of peak 2 from chromatogram M in Table II. In this case, an O.M.F. score based solely on the second-order derivative spectra match factor score would have resulted in an incorrect match.

While in general, the zero order match factors usually provided good peak matching, as illustrated for peaks 1 and 2 in Table II, the accuracy of the peak tracking procedure can be enhanced by incorporation of a weighted contribution of the first- and second-order derivative spectra match factors. Detailed analysis of several spectral comparisons of this sort demonstrated that a weighted combination of each of the individual match factors is needed to create an accurate O.M.F. score. The largest weighting was assigned to the zero-order derivative spectra match factor score (0th DSMF) with smaller contributions from the first- (1st DSMF) and second-order (2nd DSMF) derivative spectra match factor scores. The weighting of the individual match factors also allowed increased baseline noise levels associated with derivative spectra to be taken into account, since taking the derivative of a spectrum also multiplies the noise inherent in the baseline of the spectrum.

Considering the factors outlined above, the relative contribution of the first derivative spectra ($dA/d\lambda$) match factor was given 1/10th the weight of the zero-order derivative spectra match factor; the relative contribution of the second derivative spectra ($d^2A/d^2\lambda$) was then given $(1/10\text{th})^2 = 1/100\text{th}$ the weight of the first derivative spectra match factor. Thus, the final equation for the overall match factor (O.M.F.) score has the form:

$$\text{O.M.F.} = \frac{\text{0th DSMF} + 10\%\ \text{1st DSMF} + 0.1\%\ \text{2nd DSMF}}{1.101} \quad (2)$$

The peaks with the highest overall match factor are selected as the best matching peaks and should thus represent the same solute in both chromatograms.

Reproducibility of peak tracking software and chromatographic equipment. The reproducibility of individual match factors determines the statistical limits for similarity between any two spectra, and thus defines the sensitivity of the spectral matching. According to classical linear regression theory, a match factor of 1000 would characterise a perfect match according to eqn. 2, whereas a value of 0 would indicate the spectra are totally dissimilar. Values >990 (*i.e.* $r^2 > 0.99$) would indicate *statistical* identity, values between 900 to 990 would indicate *statistical* similarity, whilst values <900 (*i.e.* $r^2 < 0.90$) would indicate the spectra are *statistically* different. Since the characteristic UV spectra of peptides between 200–300 nm arises from the peptide backbone carbonyl bond and the aromatic side-chain residues, a fairly high degree of spectral similarity is expected for peptide solutes (*i.e.* match factors >900). However, additional factors will also

influence the degree of spectroscopic identity. For example, experimental errors in chromatographic equipment such as pump flow-rates, temperature instabilities, UV lamp deterioration, electrical interference and even mechanical vibrations all contribute to levels of detector baseline spectral noise above those envisaged in the ideal theoretical models upon which the *statistical* values for match factors are based. Two spectra and their corresponding solute peaks can only be considered different when the mean and standard deviation for the O.M.F. between them differs significantly from those obtained by repeatedly matching identical spectra. Thus, reproducibility of the peak tracking method (with respect to the software, detector and chromatographic hardware) was determined in order to obtain more appropriate cut-off values for matching criteria than the purely statistical values quoted above.

One system to determine such cut-off scores is by repetitive matching of identical chromatograms. Thus, by repetitively injecting the tryptic digest sample of r-pGH, recording the chromatograms under identical conditions and applying the peak tracking procedure, information was acquired on the reproducibility of spectra and the minimum cut-off scores needed to determine whether two spectra are associated with the same solute peak or different solute peaks. The r-pGH tryptic digest mix was therefore chromatographed three times under identical chromatographic conditions (*i.e.* same mobile phases, stationary phase, gradient time, flow-rate, and temperature). The chromatograms recorded were designated A, B and C, and are shown in Fig. 3. Chromatograms A, B and C were then subjected to the peak tracking procedure whereby each chromatogram was compared to the other two chromatograms (*i.e.* A to B and C, B to C and A, and C to A and B). Table III shows part of the output from one of those comparisons. For each of the six pairs of chromatograms compared, the mean of the highest O.M.F. scores was calculated and used to construct Table IV. These values in turn were used to calculate the overall mean of the highest O.M.F scores (999.79 ± 0.27 as shown in Table IV). Note that in this case, all the highest O.M.F. scores were obtained from peaks which are

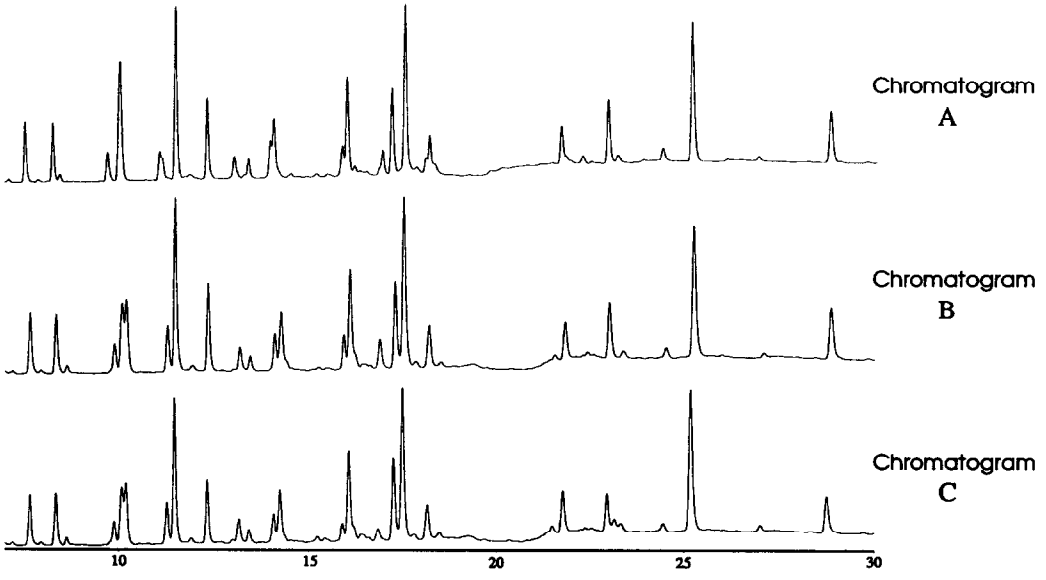

Fig. 3. Chromatograms of r-pGH tryptic digest used to determine the reproducibility of the peak tracking procedure. Each chromatogram was recorded using the same chromatographic conditions.

TABLE III

HIGHEST OVERALL MATCH FACTOR (O.M.F.) SCORES FOR THE PEAKS IN CHROMATOGRAM B WHEN COMPARED TO THE PEAKS IN CHROMATOGRAM C BY THE PEAK TRACKING METHOD

These scores give an indication of the reproducibility of O.M.F. scores, since the two chromatograms (B and C) were recorded under identical chromatographic conditions.

Peaks matched[a] by highest O.M.F. score	Score obtained (*i.e.* highest O.M.F. score)
B1 to C1	999.98
B2 to C2	999.97
B3 to C3	999.75
B4 to C4	999.99
B5 to C5	999.97
B6 to C6	999.66
B7 to C7	999.91
B8 to C8	999.98
B9 to C9	999.96
B10 to C10	999.99
B11 to C11	999.68
B12 to C12	999.31
B13 to C13	999.25
B14 to C14	999.98
B15 to C15	999.58
Mean ± S.D. of highest O.M.F. score	999.80 ± 0.25

[a] Since chromatographic conditions are identical between chromatograms B and C, all peaks are correctly matched.

correctly matched. The overall mean of the highest O.M.F. scores gives a minimum cut-off score above which two peaks from different chromatograms are very likely to arise from the same solute.

These cut-off values for O.M.F. scores should be very reliable since they were obtained under near-ideal circumstances in which the chromatographic conditions were identical. Actual applications of this peak tracking method will involve variation in at least one of the chromatographic conditions, which would be expected to lead to deterioration in match factor stabilities. That is, when chromatograms are recorded under different chromatographic conditions, lower match factor scores would be anticipated. Thus, using the data summarised in Table IV, it can be concluded that peak comparisons having O.M.F.

TABLE IV

VARIATION OF O.M.F. SCORES UNDER INVARIANT CHROMATOGRAPHIC CONDITIONS

Chromatograms compared	Highest O.M.F.[a] scores (mean ± S.D.)	Range of highest O.M.F. scores
A→B	999.77 ± 0.31	998.95–999.97
B→A	999.77 ± 0.31	998.95–999.97
A→C	999.81 ± 0.26	999.14–999.98
C→A	999.81 ± 0.26	999.14–999.98
B→C	999.80 ± 0.25	999.25–999.99
C→B	999.80 ± 0.25	999.25–999.99
Overall average	999.79 ± 0.27	

[a] Represents the average O.M.F. scores derived from individual peak pairs with highest O.M.F. scores, representing correctly matched peaks between the two chromatograms (*i.e.* corresponding to the same peptide solutes in both chromatograms).

scores greater than 999.79 ± 0.27 can be considered to be correctly matched with a high degree of confidence.

The results obtained also allow the limits of the reproducibility of the peak tracking method to be defined with respect to the software and hardware used. Any peaks with O.M.F. scores above 999.79 ± 0.27 can be considered to be identical. Therefore if any one peak is matched with two or more peaks with an O.M.F. score above 999.79 ± 0.27 then as far as the sensitivity of the equipment and the software is concerned, those peaks are identical and an unambiguous identification of the correct matching peak cannot be made.

Special features of peak tracking software. The peak tracking software has a number of special features which allow the user to selectively manipulate the way the software is applied to different problems. The software has been designed to incorporate the following features:

(1) The user can select which section of a chromatogram they wish to search for matching peaks. That is, selected portions of one chromatogram can be compared with selected portions of another chromatogram. This feature is especially useful in cases where only a small area of the chromatogram is of interest, or if the

identity of only a few of the peaks is uncertain. This feature allows the user to selectively exclude solvent breakthrough peaks.

(2) The wavelength range over which spectra are compared can be selected by the user. Peptides (such as those derived from growth hormone) often lack chromophores with UV absorbance above 300 nm and hence the operator might select a wavelength range from 200 to 300 nm. Alternatively, for peptides or proteins containing a heme group or some other chromophore, the operator might select an extended wavelength range for comparisons, anywhere from 200 to 600 nm.

(3) The integrator threshold value (the absorbance value above which a peak is detected) can be adjusted to exclude small "noise" peaks from the comparison process.

(4) Automatic spectral baseline subtraction. Concern has been expressed in the past [20,21] that spectroscopic peak matching methods assume that the spectral characteristics of the solute components do not change significantly with varying experimental conditions. In light of the known background absorbance of TFA at relatively high concentrations [22], these concerns seem well founded. For example, if the differences between the UV spectra for a given solute induced by variations in the background absorbance of the mobile phase are larger than the differences between the UV spectra of different solutes recorded under identical conditions, then clearly the application of multi-channel PDAD UV detection will be of limited use. This concern has been directly addressed in the software by automatically subtracting baseline spectra from each peak spectra to obtain a "pure" peak spectra free from baseline (background) absorbances. This procedure results in peak spectra which are independent of the mobile phase absorbances arising in the particular chromatographic system used to obtain the chromatogram.

(5) The overall match factor is based not simply on the UV spectra, but also on the first- and second-order derivatives of these spectra, adding a further level of sensitivity to matching when the underivatised UV spectra may seem similar.

(6) The software automatically ranks the 5 best matching peaks so that if a mismatch occurs, the next best candidate for a matching peak can be found in the 5 best matching peaks. This feature also allows a quick analysis of the degree of similarity between peptide spectra being compared.

(7) Visual presentation of spectra can be made either to the computer screen or to hard-copy devices. This option may prove useful in cases when human judgement is desired or required. The human eye and brain still remains the unsurpassed instrument for pattern recognition.

(8) Automatic retention time checking can be enabled for chromatograms recorded under identical chromatographic conditions. This option does not form part of the actual peak matching decision making process, rather it is a feature which indicates in the final report peaks which fall within a 10% time window of the peaks with which they are being compared. This feature should prove especially useful in applications such as quality control testing of different batches of synthetic peptides separated by RP-HPLC.

(9) Spectral data is "smoothed" using a 7-point Savitzki-Golay smoothing algorithm [23] to reduce the influence of baseline noise in spectra. The use of "weightings" in the calculation of the O.M.F. score also attempts to reduce noise influences in the comparison process.

(B) Application of the peak tracking software

The general application of the peak tracking procedure to match peaks between any two chromatograms will be described on the basis of the following selected example. The example was chosen to illustrate not only the ability of the software to correctly match peaks, but also to illustrate some of the shortcomings in the method and to discuss ways to avoid them.

Two typical examples of chromatograms of tryptic digest maps of r-pGH were selected from the database of more than 150 recorded chromatograms (see Fig. 4). The two chromatograms (designated as chromatograms X and Y in the following discussion) were generated under significantly different chromatographic conditions.

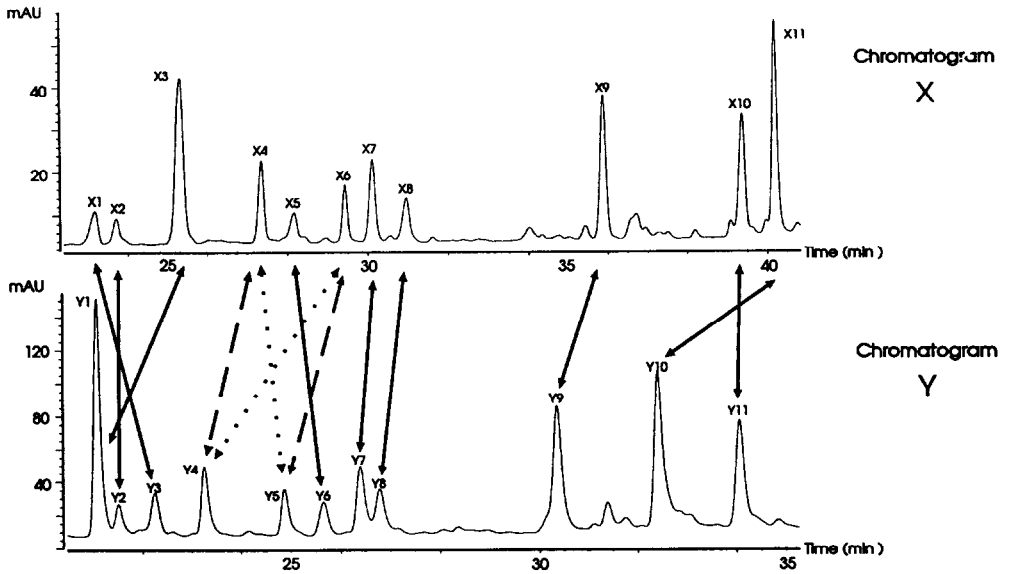

Fig. 4. Graphical representation of the peak tracking method applied to match peaks between two different chromatograms. ⟷ Peaks correctly matched by peak tracking method; ←– – –→ correct matching peaks (but incorrectly matched by peak tracking); ←· · ·→ peaks matched incorrectly by peak tracking method.

Chromatogram X was recorded using solvent system 1 (aqueous TFA–acetonitrile —see Table I) using a 120-min gradient elution (0–100% B) at a flow-rate of 1.0 ml/min with a C_{18} sorbent at 37°C. Chromatogram Y was recorded using solvent system 3 (aqueous TFA–propanol —see Table I) using a 90-min gradient elution (0–100% B) at a flow-rate of 0.6 ml/min with a C_{18} sorbent at 37°C.

The identity and position of the individual peaks in each chromatogram were confirmed by fraction collection of the separated r-pGH tryptic fragments, independently determining their composition by amino acid analysis, and re-injection of purified isolated fragments under the chromatographic conditions used here. However, the peak tracking strategy as such assumes no prior knowledge of the number of components nor their identities, hence chromatograms of "unknown" mixtures can be analysed in an analogous manner.

Peak self-matching. In part A of this discussion, the reproducibility of the O.M.F. scores were established (that is, the reproducibility of the O.M.F. scores with respect to the software and hardware used). Another factor which needs to be considered is the selectivity of the peak matching process; that is, the stability of the method against false positive matches. To establish the selectivity of the method, each solute peak in a chromatogram was matched against every other peak in that same chromatogram to determine how similar the peak spectra are within a chromatogram, and hence estimate the number of potential mismatches. A mismatch in this context is defined as a solute peak spectra for which more than one match candidate (*i.e.* other than itself) was found with a O.M.F. score within the cut-off limits established in part A. That is, any peaks with O.M.F. scores of 999.79 ± 0.27 are considered to be identical within the reproducibility limits of the peak tracking method.

Table V summarises the results obtained when just such an analysis was performed on chromatogram Y (*i.e.* the peaks from Chromatogram Y have been compared to themselves). Table V shows that although some peaks have quite high O.M.F. scores with other peaks (*i.e.* other than themselves with the expected perfect match factors of 1000.00) within the same chromatogram, only two of the peaks, Y4 and Y5, have

TABLE V

SUMMARY OF THE O.M.F. SCORES OBTAINED FOR FIVE BEST MATCHING PEAKS WHEN THE PEAKS IN CHROMATOGRAM Y ARE COMPARED TO THEMSELVES BY THE PEAK TRACKING METHOD

Values in bold have exceptionally high overall match factor.

Peaks Y[a]	Peaks Y										
	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10	Y11
Y1	1000.00										
Y2		1000.00				966.56	993.42	995.14	998.89	991.27	
Y3			1000.00		983.95	998.86					
Y4	959.33		983.09	1000.00	**999.60**	976.82				985.11	996.18
Y5	955.30		983.95	**999.60**	1000.00	978.65					995.01
Y6			998.86			1000.00					
Y7		993.42					1000.00	999.12	995.74		
Y8		995.14					999.12	1000.00	997.60	987.43	
Y9		998.89					995.74	997.60	1000.00	990.88	
Y10		991.27	975.47	985.11			983.38	987.43	990.88	1000.00	978.93
Y11	975.26			996.18	995.01						1000.00

[a] Peaks Y: peak numbers from chromatogram Y.

O.M.F. values which fall within the previously established confidence limits of the reproducibility of the method (*i.e.* 999.79 ± 0.27). Thus, when the sensitivity of the peak tracking method is taken into account, these two peaks are basically indistinguishable, and we can can expect the peak tracking method to have problems identifying them accurately. Since these two peaks have such high O.M.F. scores, this must mean that they have very similar spectra.

Amino acid analysis of the chromatographic fractions corresponding to peaks Y4 and Y5 from chromatogram Y revealed that they have identical amino acid composition corresponding to the peptide pGH [141–152] (QTYDKFDTNLR) and would therefore be expected to exhibit identical spectra. As indicated in Fig. 4, peaks Y4 and Y5 from chromatogram Y and their respective corresponding peaks X4 and X6 from chromatogram X have significantly different elution positions. The cause of the different retention behaviour of these peptides with apparently identical amino acid composition is currently being investigated, but by analogy with other xxYDKxx containing peptides may correspond to the β-rearranged form of the aspartic acid at position xDx.

To summarise, because of the exceptionally high O.M.F. scores (which fall within the absolute sensitivity limits of the peak tracking method) and the corresponding spectral similarity between peaks Y4 and Y5, these peaks are unlikely to be able to be accurately distinguished by the peak tracking software. In other words, the software provides a pre-warning for matches assigned by the peak tracking method for peaks Y4 and Y5.

Peak cross-matching. Table VI summarises the results obtained after application of the peak tracking software to the two selected chromatograms. The peaks from chromatogram X were matched to peaks in chromatogram Y. The peaks from chromatogram Y with the 5 highest O.M.F. values for each of the peaks in chromatogram X are presented. Fig. 4 graphically illustrates the matching of peaks from chromatogram X to chromatogram Y as selected by the peak tracking algorithms. As indicated in Fig. 4 and Table VI, 9 of the 11 solute peaks in chromatogram Y were correctly matched to their corresponding solute peaks in chromatogram X. Fig. 4 and Table VI also indicate that 2 of the 11 peaks were not correctly matched. Peak X4 was incorrectly matched to peak Y5 when it should have been matched to peak Y4. Similarly, peak X6 was

TABLE VI

SUMMARY OF THE O.M.F. SCORES FOR THE FIVE BEST MATCHING PEAKS FROM CHROMATOGRAM Y WHEN COMPARED TO EACH PEAK IN CHROMATOGRAM X BY THE PEAK TRACKING METHOD

Values in italics: peaks correctly matched by the peak tracking method. Values in bold: peaks incorrectly matched by the peak tracking method. Value between parentheses: correct matching peaks (but incorrectly matched by the peak tracking method).

Peaks Y[a]	Peaks X[b]										
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
Y1			*999.54*								
Y2	974.66	*997.17*			965.67		991.46	994.23	998.45		991.25
Y3	*998.00*			983.38	997.20	982.75					
Y4	974.36		964.43	(999.22)	973.30	**997.81**				995.93	987.27
Y5	976.13		960.98	**999.33**	975.78	(997.63)				995.23	
Y6	997.86				*998.72*						
Y7		989.99					*999.58*	997.70	995.07		
Y8		990.12					998.21	*999.10*	996.76		986.25
Y9		994.65					993.68	996.96	*999.80*		990.54
Y10		987.06					980.45	986.39	990.14	975.16	*999.79*
Y11			979.72	994.85		993.60				*999.54*	

[a] Peaks Y: peak numbers from chromatogram Y.
[b] Peaks X: peak numbers from chromatogram X.

incorrectly matched to peak Y4 when it should have been matched to peak Y5.

Thus, as predicted, peaks Y4 and Y5 were indeed mismatched in this example. It is not surprising that peaks Y4 and Y5 were incorrectly matched, since even the slightest differences in chromatographic baseline noise will subtly alter the spectral characteristics of these peaks and make these peptide peaks especially susceptible to mismatching. It should be noted, however, that the correctly matched (but *not* best matching) peaks in these cases (values between parentheses in Table VI) were in fact the second-best matching peaks (*i.e.* had the second highest O.M.F. scores) and have very similar O.M.F. scores to the incorrectly matched (but best matching) peaks (values in bold in Table VI) in both cases.

This example has therefore illustrated an important feature point in this peak tracking method. When all components differ sufficiently in their spectral characteristics, a close match between the spectra, expressed by the O.M.F score, is adequate for unambiguous identification purposes (as illustrated by 9 of the 11 peaks correctly matched in the above example). However, when two or more components in the mixture being investigated have very similar spectral characteristics (O.M.F. scores within the reproducibility limits of the equipment and software), as was the case for the solutes corresponding to peaks Y4, Y5, X4 and X6, an additional source of peak matching information is required. This extra information could be supplied by comparing the relative areas or heights of the peaks. Close examination of Fig. 3 demonstrates that just such an analysis would easily have matched peaks X4 to Y4 and X6 to Y5, since they have different area ratios to each other.

In the rare situation where two different components in a sample have identical spectral characteristics as well as equal peak areas/heights, the present version of the software will not distinguish between them and the outcome will involve three possible peak matching solutions. The first would be to assume that no cross-over of peak elution order occurs as the chromatographic conditions are changed. This seems the most reasonable since peptide homo-

logues will usually respond in a similar fashion to changes in the selectivity of the solvent system. The second option would be to assume that peak cross-over has occurred, and the third option would be to assume that the peaks co-eluted (this latter option would be obvious since there would be a "missing" peak in the chromatogram). Each of these three possible solutions can then be considered independently from analysis of the log $\bar{k}$ *versus* $\bar{\psi}$ plots, or from the slope of log $\bar{k}$ *versus* $1/T$ plots [24].

CONCLUSIONS

The described peak tracking procedure is capable of monitoring the positions of each peptide solute peak between any two chromatograms of a solute mixture recorded under different chromatographic conditions. The process is based on the analysis of each peak's characteristic UV–Vis spectrum to generate an overall match factor (O.M.F.) representing the similarity between any two peptide peaks from the different chromatograms.

Using this new peak tracking method, the peptide solutes derived from a tryptic map of r-pGH can be identified and tracked across various chromatographic conditions, including changes in stationary phases, mobile phase solvents, gradient times, temperatures and solvent flow-rates. As shown in the selected example, the peak tracking software can effectively deal with changes in peak elution orders and relative peak areas. The nine correctly matched peaks in the selected example illustrate that when the solute peaks in the mixture differ sufficiently in their spectral characteristics from each other, a close match between spectra, as expressed by an O.M.F. score, is sufficient for unambiguous identification. However, when components with very similar spectra are present in a mixture, additional information such as the relative areas/heights of the peaks must be used in order to perform an unambiguous identification. Further work is underway to incorporate into the peak tracking software an algorithm to perform peak area matching for these difficult cases. Another area which needs to be investigated is the problem of mixed component spectra due to poorly resolved peaks. This problem could be addressed by the use of principle component analysis to de-convolute the complex impure spectra into their component spectra. In the present method, no attempt is made to match peaks whose spectra reveal that they are impure (*i.e.* co-eluting or poorly resolved peaks).

The peak tracking method described in this paper is not limited to tracking peaks from tryptic digests of proteins such as the r-pGH tryptic digests. The method should be generally applicable to any peptide mixture, or indeed any mixture of organic molecules. In fact, peptide fragments, because of their rather nondescript and similar spectra, probably represent a more difficult scenario to deal with than many other types of chromatographic samples of comparable compositional complexity but with considerably greater spectral variety. This peak tracking procedure should thus assist in the development of new HPLC optimisation protocols, providing a basis for improved strategies for the monitoring and control of the analysis and purification of biological macromolecules, particularly peptides and proteins produced by chemical synthesis or recombinant DNA techniques, and from enzymatic and chemical digestions.

ACKNOWLEDGEMENT

The support of the Australian Research Council is gratefully acknowledged.

REFERENCES

1 M.T.W. Hearn (Editor), *HPLC of Proteins, Peptides and Polynucleotides —Contemporary Topics and Applications*, VCH, Deerfield, FL, 1991.
2 J.C. Berridge, *Techniques for the Automated Optimization of HPLC Separations*, Wiley-Interscience, New York, 1985.
3 P.J. Schoenmakers, *Optimization of Chromatographic Selectivity —A Guide to Method Development* (*Journal of Chromatography Library*, Vol. 35), Elsevier, Amsterdam, 1986.
4 S.D. Patterson, *J. Chromatogr.*, 592 (1992) 43.
5 A.G. Wright, A.F. Fell and J.C. Berridge, *J. Chromatogr.*, 458 (1988) 335.
6 H.J. Issaq and K. McNitt, *J. Liq. Chromatogr.*, 5 (1982) 1771.

7 J.C. Berridge and E.G. Morrissey, *J. Chromatogr.*, 316 (1984) 69.
8 A.S. Kester and R.E. Thompson, *J. Chromatogr.*, 310 (1984) 372.
9 J.C. Berridge, *J. Chromatogr.*, 244 (1982) 1.
10 A.G. Wright, A.F. Fell and J.C. Berridge, *Chromatographia*, 24 (1987) 533.
11 D.H. Hill, T.R. Kelly and K.J. Langner, *Anal. Chem.*, 59 (1987) 350.
12 M. Otto, W. Wegscheider and E.P. Lankmayr, *Anal. Chem.*, 60 (1988) 517.
13 A.C.J.H. Drouen, H.A.H. Billiet and L. de Galan, *Anal. Chem.*, 56 (1984) 971.
14 Y. Zhang, H. Zou and P. Lu, *J. Chromatogr.*, 515 (1990) 13.
15 A.C.J.H. Drouen, H.A.H. Billiet and L. de Galan, *Anal. Chem.*, 57 (1985) 962.
16 H-J.P. Sievert, S-L. Wu, R. Chloupek and W.S. Hancock, *J. Chromatogr.*, 499 (1990) 221.
17 P.J.M Coenegracht, H.J. Metting, E.M. van Loo, G.J. Snoeijer and D.A. Doornbos, *J. Chromatogr.*, 631 (1993) 145.
18 B. Grego, E. Nice and R.J. Simpson, *J. Chromatogr.*, 352 (1986) 359.
19 A. Drouen, *The COMPARE Command Information Note*, Hewlett-Packard, Waldbronn, Publication Number 12-5952-3725, 1987.
20 J.K. Strasters, H.A.H. Billiet, L. de Galan and B.G.M. Vandeginste, *J. Chromatogr.*, 499 (1990) 499.
21 J.K. Strasters, F. Coolsaet, A. Bartha, H.A.H. Billiet and L. de Galan, *J. Chromatogr.*, 499 (1990) 523.
22 G. Winkler, P. Wolschann, P. Briza, F. Heinz and C. Kunz, *J. Chromatogr.*, 347 (1985) 83.
23 A. Savitzky and M.J.E. Golay, *Anal. Chem.*, 36 (1964) 1627.
24 A.W. Purcell, M.I. Aguilar and M.T.W. Hearn, *J. Chromatogr.*, 593 (1992) 103.
25 Q.M. Mao, I.G. Prince and M.T.W. Hearn, *J. Chromatogr.* 646 (1993) 81.