Title

Use of mathematical derivatives (time-domain differentiation) on chromatographic data to enhance the detection and quantification of an unknown 'rider' peak.

Authors

S.J. Ford, , M.A. Elliott, G.W. Halbert

Cancer Research UK Formulation Unit,
Department of Pharmaceutical Sciences,
University of Strathclyde,
204 George Street,
Glasgow, G1 1XW, UK

Correspondence:

S.J. Ford

Cancer Research UK Formulation Unit,
Department of Pharmaceutical Sciences,
University of Strathclyde,
204 George Street,
Glasgow, G1 1XW, UK

Abstract.

Two drug samples submitted to this lab for HPLC assay showed an unidentified impurity which eluted as a 'rider' on the tail of the main peak. The use of mathematical derivation of the chromatograms offered several advantages over conventional skimmed integration: enhancement of sensitivity, excellent linear response, elimination of the subjective nature of skimming and a reliable estimate of the true area of the impurity peak.

Keywords

Tangent integration, Skimmed integration, Resolution enhancement, Baseline prediction.

Introduction.

Rider peaks are those which are incompletely resolved from larger peaks within a chromatogram. Their accurate characterisation and quantification are difficult and several alternatives have been investigated. Triangular/perpendicular methods and valley-to-baseline for rider integration are known to be prone to errors [1,2]. With the popular tangential (skimming) method careful selection of the baseline end points are required [3], but significant errors can still occur [4, 5] particularly with situations where the rider peak elutes after the main peak. The use of mathematical

deconvolution of the chromatographic data has been studied [6] as have two dimensional calibration techniques [7].

The technique of chromatogram derivatization (or time domain differentiation) is used to separate the signal of a narrow peak from an unresolved neighbour with a larger bandwidth. The technique is used in UV spectroscopy, polarography [8] and thermogravimetric analysis [9]. Chromatographic data is derivatized as part of the peak integration processing [1, 3] and the use of derivative chromatography for the quantification of closely eluting peaks of similar magnitude has been shown previously [10, 11]. Theoretical calculations have showed good results for peak area ratios between 100% and 10% [12]. This study investigated the use of derivatization of the chromatography data to assess an unknown rider peak where the peak areas of the two eluants are significantly different using experimental data and theoretical calculations. (The authors are unaware of any literature documenting the use of derivatives for the quantification of 'rider' peaks in chromatographic data.)

The Cancer Research UK Formulation Unit was provided with two samples of a new anti-cancer prodrug, AQ4N [13], for the purposes of formulation and analytical development. During HPLC assay development, six relevant peaks were observed; the parent drug, four well separated impurities and a fifth impurity (nominally labelled Impurity D) eluting as a rider on the tail of the main peak. The level of each impurity varied slightly between the samples. Samples of the observed impurities were not available.

Additional HPLC method development work was carried out in an attempt to improve the separation of parent compound and Impurity D. Modifications of aqueous phase pH, organic mobile phase, column length, stationary phase and column temperature were all attempted, but with no success: the developed HPLC assay method appeared to give the optimal separation. These investigations were hampered by the lack of an appropriate Impurity D sample and the low Impurity D content of the available AQ4N samples was very small. Additionally, the other well resolved impurities (which were also present at similar levels and for which no authentic standards existed) acted to obscure any change in elution characteristics that the experimental modifications generated.

While there is no substitute for good chromatographic separation, our HPLC investigations had yielded no additional benefit, and so the possibility of obtaining content information about Impurity D by the mathematical processing of the chromatographic data was examined. In the absence of a suitable standard a percentage (w/w) ratio content of Impurity D cannot be established, however this work concentrates on identifying a reliable indicator of the quantity of Impurity D using the second order derivative of chromatogram and using this information to estimate peak area.

Experimental

Chemicals

HPLC grade reagents and solvents were used throughout. Two samples of

AQ4N.2HCl (nominally titled samples 'A' and 'B') were provided by Prof. W. Denny (Auckland Cancer Society) [14].

Chromatography

A ThermoSeparation HPLC system, consisting of a SM4000 four line vacuum degasser, P2000 binary gradient pump, A1000 autosampler and a UV1000 detector integrated via a SN4000 SpectraNet module with a PC (Dell Optiplec Gm) running HPLC acquisition software PC1000 (version 3.0.1). The HPLC system is calibrated on a biannual basis.

Mobile phase A consisted of 75% (v/v) of 0.1% (v/v) tri-fluoroacetic acid and 25% (v/v) acetonitrile. Mobile phase B was acetonitrile. The gradient used was 100% A for 10 minutes, with a linear gradient to 100%B at 20 minutes followed by a 16 minutes equilibration period. The column was a Phenomenex Luna C8(2) (5 µm, 150 x 4.6 mm) dedicated to AQ4N HPLC assay. The flow rate was 1ml/min, with an injection volume of 20µl. The UV detection wavelength was 245 nm. Each sample was injected in triplicate.

The integration of the Impurity D peak was carried out used the 'Rider' integration option in the PC1000 software. The start and end points of the baseline were chosen manually according to the guidelines suggested by Dyson [3]. Chromatograms were exported as comma separated (CSV) files under the PC1000 Data Maintenance program using start/stop times of 4.3/5.5 minutes and a data interval of 1.

Sample preparation.

Samples A and B were prepared at 0.2 mg/ml in sodium orthophosphate buffer (pH=7.0, 10mM). Mixtures of these two samples were prepared using calibrated Gilson variable volume pipettes in the following proportions: 100A:0B, 75A:25B, 50A:50B, 25A:75B and 0A:100B.

Differentiation calculations

The derivatization of the chromatography data was carried out using Microsoft Excel '98 (Macintosh Edition) by using the Savitzky-Golay series for the second derivative of a quadratic/cubic polynomial over eleven points [15]. The equation used to obtain the second derivative was as follows:

$$y'' = \frac{1}{429x\Delta t^2} \cdot (15y_{i-5} + 6y_{i-4} - y_{i-3} - 6y_{i-2} - 9y_{i-1} - 10y_i - 9y_{i+1} - 6y_{i+2} - y_{i+3} + 6y_{i+4} + 15y_{i+5})$$

where y" = second order derivative of the chromatogram at point y_i,
Δt = time interval,
and y_{i-5} to y_{i+5} = consecutive point on the chromatogram,

Prior to the application of the Savitzky-Golay algorithm, initial data bunching was carried out as an averaging process of six sequential datapoints, this generated a single time point every 0.005 minutes. This reduced the size of the subsequent spreadsheets and acted to smooth the data. A simpler, but more cumbersome, stepwise numerical

derivatization process of consecutively calculating the gradient of a slope between adjacent points was also assessed, but it generated equivalent results to the Savitzky-Golay process and required an additional averaging step and larger spreadsheets.

Theoretical calculations

The Gaussian equation which is used to describe a model chromatographic peak is as follows [2]:

$$h(t) = \frac{A}{\sigma.\sqrt{2\pi}} \cdot \exp\left[\frac{-(t - t_R)^2}{2\sigma^2}\right]$$

where h(t) = peak height at time t,
A = total peak area,
t_R = time at peak maximum (retention time),
and σ = standard deviation of the peak.

The theoretical model of the peak complex was generated by the addition of two Gaussian functions (Peak 1 and Peak 2). The total peak area values (A) were selected so that the ratio of the A values reflected the approximate peak area ratio determined experimentally. The total peak area values for peak 2 were varied to investigate the linearity of the second derivative method. The t_R values were set to the retention times of the main peak and Impurity D and the σ values were optimised to generate a peak complex which visually represented the experimental traces. The unitless Gaussian parameters are shown in Table I.

Table I: Gaussian parameters for the theoretical curves.

Parameter	Values for Peak 1	Values for Peak 2
A	10000	10, 20, 30, 40, 50
t_R	4.0	4.7
σ	0.16	0.15

The theoretical data was processed the same way as the experimental data described above without the initial bunching operation.

Skimmed integration of the theoretical peak complex was carried out in the Excel spreadsheet by selecting two 'baseline' points on the curve and fitting a straight line. Peak areas were calculated by subtracting the baseline from the curve, multiplying by time interval and summing over the relevant range.

Results and Discussion

Initial chromatographic results

Skimming integration could be used for both samples, however, no 'valley' could be observed between the two peaks in Sample Band so the Impurity D peak would be below the 'shoulder level' but above the 'detection level' as defined by Westerberg [1].

Typical chromatograms are shown in Figure 1.

The mean and %RSD values for peak areas (as given by the skimmed integration within the HPLC acquisition program, PC1000) of Impurity D are shown Figure 2a. The best fit linear equation is y = 258x - 1337 with standard deviations of the slope and intercept, standard error and number of data points as 9.49, 577, 1290 and 15 respectively. The correlation co-efficient for a linear regression between the Impurity D content (expressed as solution content of Sample B) and the mean peak area is 0.983. The peak area gives a distinct curve and an aberrant negative intercept. The non-linear response suggests that even if an authenticated Impurity D standard were available a standard additions approach would still give erroneous results. The peak area data suggest that sample A has less than 1 % of the Impurity D content of sample B.

The poor result from the peak area data originates from the invalid use of straight 'baseline' in a situation where the peak clearly sits on top of a 'curved' background originating from the adjacent peak. It would be possible to utilize a curved baseline, however in the absence of a drug sample completely devoid of Impurity D such a baseline could only be drawn as an 'analytical' best guess. One possible option would be to attempt to fit the peaks to known equations (refs here). The use of the actual experimental data to determine the correct curved baseline by a linear regression method is discussed below.

Derivative results

The mean and %RSD values for the amplitude of the second derivative of Impurity D are graphed in Figure 2b.

The best fit linear equation for a linear regression between the Impurity D content (expressed as solution content of Sample B) and the second derivative amplitude is y = 0.0219x + 0.473 with standard deviations of the slope and intercept, standard error and number of data points as 0.000526, 0.0322, 0.072 and 15 respectively. The correlation co-efficient is greater than 0.99. The second derivative data gives a higher %RSD than the peak area data, but this is to be expected since the increased sensitivity of derivatized experimental data is at the expense of exaggerated experimental noise. However, the 2nd derivative data gives a straight line, with a positive intercept in contrast to the peak area data. The 2nd derivative data suggest that sample A has 17 % of the Impurity D content of sample B: a substantial difference from the equivalent figure derived from the skimmed peak area measurements.

In addition to this the skimmed integration is an operation that is difficult to reproduce since the start and end points are rather non-specific: different software packages or human chromatographers, may choose different points on the trace. The second derivative trace is more specific since amplitude (like peak height) is an easier parameter to define.

The good linear fit for the 2nd derivative data experimental is mirrored by a similar result for the theoretical results discussed below and this indicates that amplitude of

the second derivative is proportion to Impurity D.

Theoretical results

The values for peak areas as given by the skimmed integration and the amplitude of the second derivative of Peak 2 against the theoretical peak area (A) are shown in Figures 3a and 3b.

The results are very similar to those with the equivalent processing of the experimental data. The best fit linear equation for a linear regression between the skimmed peak area data against the total peak area is y = 0.173x - 1.85 with standard deviations of the slope and intercept, standard error and number of data points as 0.0130, 0.431, 0.411 and 5 respectively. The correlation coefficient is 0.983. Figure 3a shows a slight concave curve with linear fitting giving a significant negative intercept. There is a significant difference between the skimmed peak area values and theoretical peak area (A) for peak 2. This confirms that under the conditions of this experiment the skimmed integration method severely under estimates the actual peak area [3, 4, 5].

The best fit linear equation for a linear regression between the 2nd derivative data against the total peak area is area is y = 15.0x - 45.4 with standard deviations of the slope and intercept, standard error and number of data points as 0.130, 4.31, 4.10 and 5 respectively. The correlation coefficient is 1.000 and indicates a good basis for a linear relationship. The non-ideal linear regression values concur with previous work [12] and show that even under theoretical conditions the main peak will distort the 2nd derivative signal. Linearity in these experiments cannot be assumed.

Extrapolation of experimental data to obtain 'pure' drug curve.

Since sample A and B have different contents of Impurity D, it should be possible to extrapolate from the available chromatograms to a point where the impurity D content is zero. This would provide an Impurity D free baseline, where chromatographic subtraction methods could be applied. The two key issues to be addressed are: firstly the correct alignment of the chromatograms and, secondly, assigning the Impurity D content of the samples mixtures. The latter problem is addressed by equation 3, where, although the exact concentration cannot be deduced, the relative concentration is expressed by the amplitude of the second derivative.

The alignment of the chromatograms was carried out using two procedures: use of the second derivative maximum for the impurity D peak and use of visual alignment of the steep tail gradient of the main drug. There was no restriction on the retention time/datapoint displacement of the chromatograms during alignment by second derivative maximum, but visual alignment was considered as 'fine tuning' and so was restricted to displacement by three data points (the equivalent of 0.015 minutes).

The extrapolation of the chromatograms to the condition where the Impurity D concentration is zero was carried out by linear regression of the absorbance of the averaged traces for the five sample mixtures at a fixed time point (as y values) against the amplitude of the second derivative (as the corresponding x value). The intercept of

this line gives the absorbance at the fixed time point where the Impurity D content is zero. Repeating this procedure for all the relevant time points gives a chromatogram where the Impurity D content is zero.

The results for the regression are shown in Figure 5 together with the scaled correlation coefficients of the analyses at the different time points. The correlation coefficients values show that the linear fit for the Impurity D peak to the 2nd derivative amplitude is good. On either side of the impurity D peak the scaled correlation coefficients value drops, since the relationship between the chromatograms is random. Subtraction of this experimentally determined baseline from the sample mixture chromatograms gives the appropriate Impurity D peak areas.

The algorithm within the HPLC software used to calculate peak areas from skimmed integration is unknown to the authors. For the purposes of comparison with the Impurity D peak areas from the curve subtraction process above the skimmed integration was replicated on a spreadsheet. The peak areas of Impurity D as given by the spreadsheet/skimmed integration process and spreadsheet/curve subtraction are shown in Table II.

Table II: Peak area from the three different estimation alogrithims

Sample mixture	Peak area from curve subtraction	Peak area from spreadsheet skim integration	Peak area from HPLC software skim integration
100B:0A	77433	24858	25712
75B:25A	61236	-	17463
50B:50A	48462	-	10613
25B:75A	30122	-	3845
0B:100A	12880	300	242

Calculations correlating the magnitude of the second derivative to the 'trace' of the theoretical peaks showed the 'baseline' of the main peak could be constructed using the same process as discussed for the experimental chromatographs above. The data is shown in the Table III.

Table III: Peak area from curve extrapolation on theoretical data. (The figures in parenthesis show the percentage of the actual peak area, A.)

Rider total peak area	Peak area from skim integration	Peak area from curve subtraction
50	7.15 (14%)	46.8 (94%)
40	5.03 (13%)	36.8 (92%)
30	2.91 (10%)	26.9 (90%)
20	1.41 (7.1%)	16.9 (84%)
10	0.28 (2.8%)	6.9 (69%)

The errors in the fitting process originate from the use of the second derivative

amplitude as the x co-ordinate in the regression. As shown above the main peak will distort the actual amplitude and lead to error at low rider peak areas. However comparison of the figures in parenthesis in Table II show that the curve fitting/regression approach is several time more accurate that the skimming integration method.

Additional studies were carried out to examine the possibility of reconstructing the Impurity D peak using a Gaussian function and σ values derived from the second order chromatogram. For the theoretical models the correct peak area could be calculated for all five rider peak areas to within 99.5% of the actual value. However, for the experimental data the values of σ predicted varied widely with relative standard deviations on the average values of up to 30%. In the equations used, the actual peak area is proportional to σ^3 and so any errors in the latter are magnified during the subsequent mathematical processing. Consequently the results from the experimental peak area predictions are meaningless.

References:

[1]. A.W. Westerberg, Anal. Chem. 41 (1969) 1770
[2]. N. Dyson, J Chromatogr A 842 (1999) 321-340
[3] N. Dyson, Chromatographic Integration Methods, Roy Soc. of Chemistry Monographs, Cambridge (1990)
[4] V.R. Meyer, 33, J Chroma Sci, 26-33
[5] V.R. Meyer, 40, Chromtagraphia, 15-22
[6] M. Johansson, M. Berglund, D.C. Baxter, 48B, Spectrochimica Acta, pp1393-1409
[7] (Jurt et al Journal Chrom A WoS printout!)
[8] Beckett and Stenlake, Prac. Pharm Chem 4th ed p 298, p223
[9] Treatise on Analytical Chemistry, Part 1, Vol 13, 2nd ed'n Kolthoff and Winefordner (Eds) Authors Jeffrey G Dunn and John H Sharp pp127-266 John Wiley and Son ISBN 0-471-80647-1
[10] A. A. Fasanmade, A.F. Fell, 61, (1989), pp720-728
[11] A.M. Delapena, F. Salinas, T. Galeano, A. Guiberteau, 234, (1990), 263-267 (WoS printout)
[12] E. Grushka, G.C. Monacelli, Anal. Chem. 44 (1972) 484
[13] Patterson, LH and McKeown, SR, (2000). Br. J. Cancer 83, 1589-1593.
[14] H.H. Lee, W.A. Denny, J. Chem. Soc. Perkin Trans 1 (1999) 2755-2758
[15] A. Savitzky, M.J.E. Golay, Anal. Chem. 36 (1964) 1627-1639

Figure 1: Chromatograms for the Impurity D peak from Samples A and B.

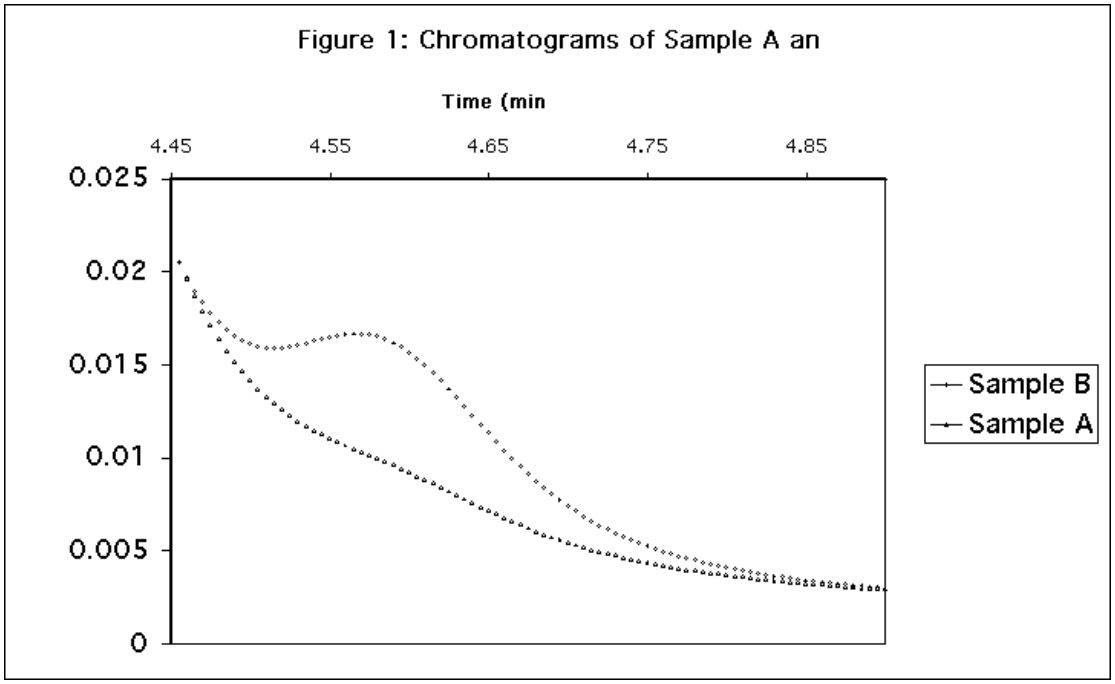

Figure 2a: Peak area of Impurity D as given by 'skimming' integration.

AQ4N integration paper Chart 1

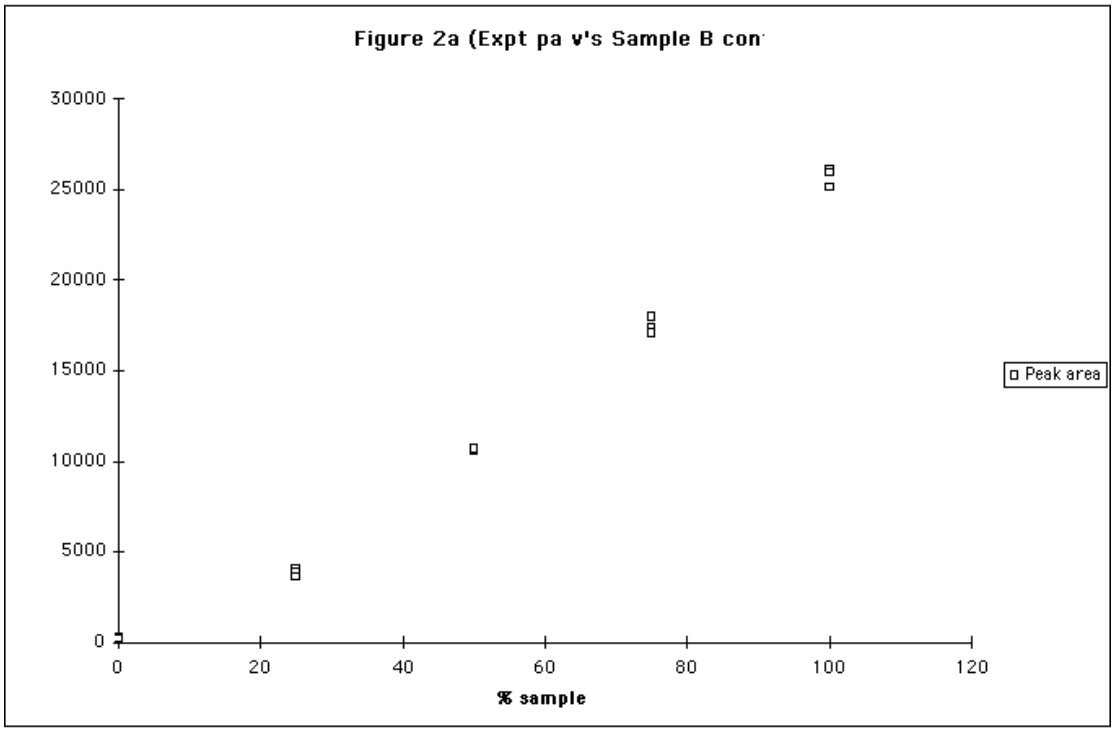

Figure 2b:

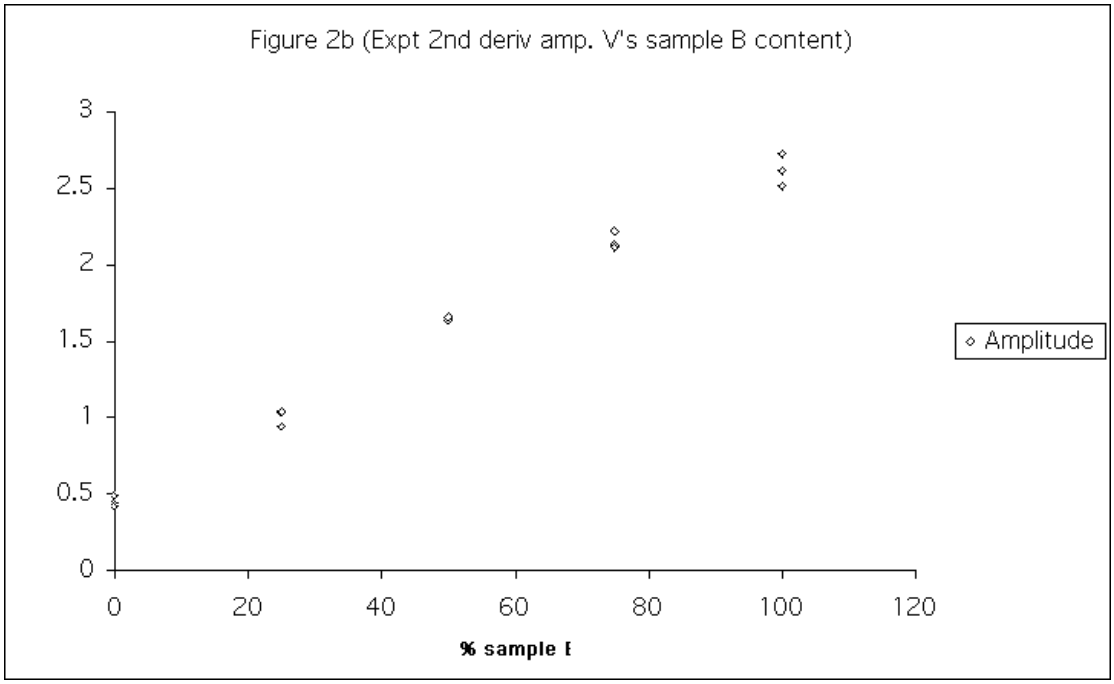

Figure 3a:

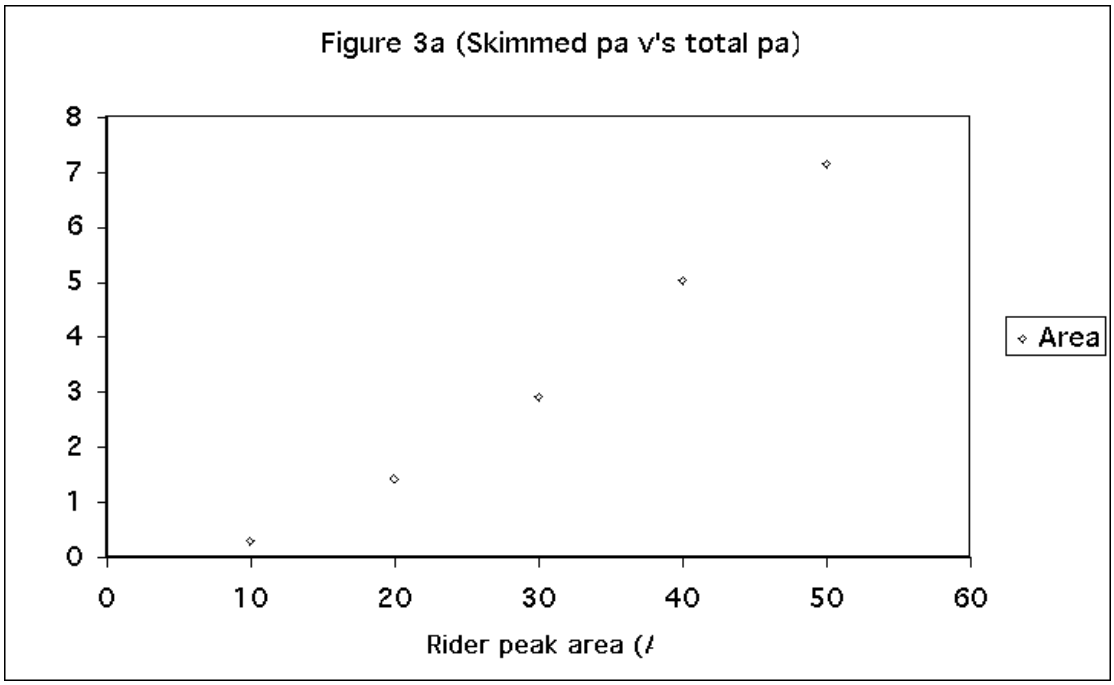

Figure 3b:

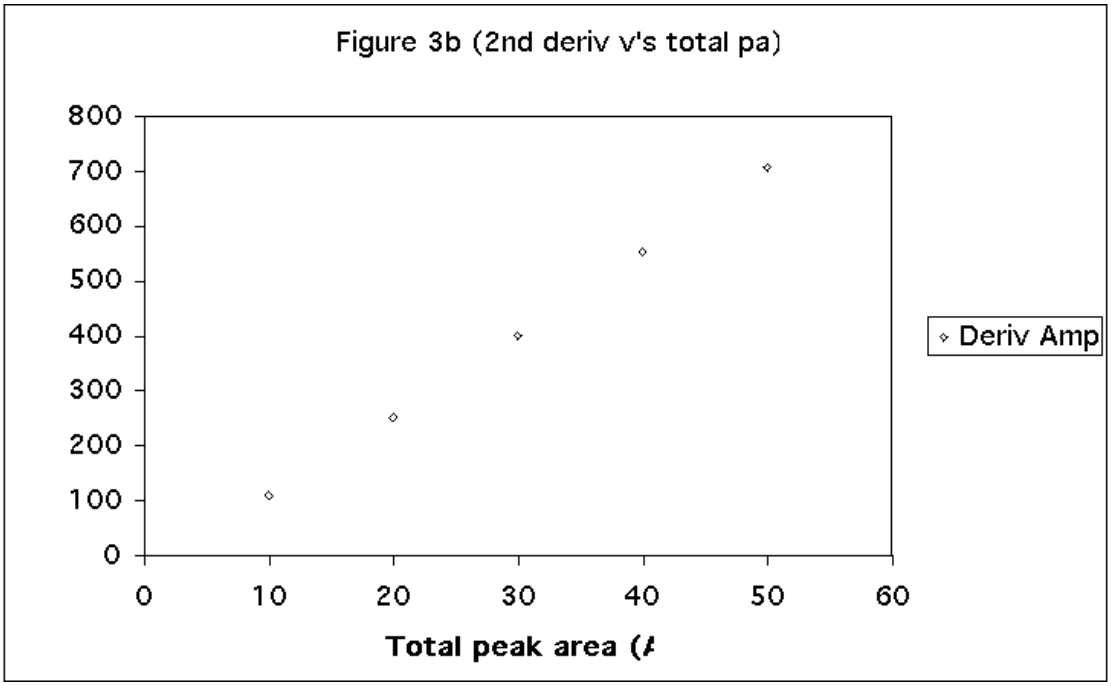

Figure 5: Chromatograms for linear regression of baseline.