Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Pharmaceutical and Biomedical Analysis
34 (2004) 531–541

JOURNAL OF
PHARMACEUTICAL
AND BIOMEDICAL
ANALYSIS

www.elsevier.com/locate/jpba

Data preprocessing by wavelets and genetic algorithms for enhanced multivariate analysis of LC peptide mapping

Fredrik O. Andersson, Rudolf Kaiser, Sven P. Jacobsson*

Analytical Development, Pharmaceutical and Analytical R&D, AstraZeneca R&D, SE-151 85 Södertälje, Sweden

Received 25 June 2003; received in revised form 17 October 2003; accepted 20 October 2003

Abstract

Peptide mapping by means of liquid chromatography is a powerful technique used for the characterisation and analysis of the primary structure of proteins. Subtle changes in the covalent structure of the protein can be detected by means of the chromatographic profile (fingerprint). Chromatographic methods, however, display variations in the chromatographic profile even at identical instrumental settings and sample conditions. These variations may be due to changes of the chromatographic conditions, e.g. slight shifts in column temperature, and degradation or alterations of the stationary phase or small changes in the trifluoroacetic acid (TFA) concentration. Such variations may result in varying retention times and peak shapes of the analytes and differences in the chromatographic baseline, thereby having a detrimental impact on the results obtained on multivariate analysis of peptide maps. In order to reduce the non-sample-related variations and to be able to more fully extract the information in peptide mapping, approaches for achieving this objective are outlined in the present study. These methods are denoising and data compression of the chromatograms by wavelets, baseline corrections by linear interpolation, and peak shift alignments towards a target chromatogram by means of a genetic algorithm. Visual inspections of preprocessed chromatograms and principal component analysis (PCA) score plots demonstrate the efficiency of the methodology used. Furthermore, deliberately added changes, e.g. insertions of small Gaussian peaks (outliers), are more easily detected by the proposed methods than from the original chromatograms by multivariate analysis.

Keywords: Peptide mapping; Wavelets; Genetic algorithms; Principal component analysis; Compression; Denoising; Baseline correction; Peak shift alignment

1. Introduction

Peptide mapping is a crucial analytical procedure for protein characterisation and analysis, which involves chemical or enzymatic cleavage of the protein into peptide fragments [1–3]. Typically, separation and identification of these peptide fragments is performed by means of liquid chromatography and through the retention time of the resulting fragments [1,4–6]. Recently, mass spectrometric detection has become increasingly used [7,8], even replacing the separation step, e.g. matrix-assisted laser desorption mass spectrometry (MALDI-MS) [9] and nano electrospray ionisation-MS [10].

* Corresponding author. Tel.: +46-8-553-289-68; fax: +46-8-553-289-68.

E-mail address: sven.jacobsson@astrazeneca.com (S.P. Jacobsson).

0731-7085/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0731-7085(03)00583-1

The basic information gained from peptide mapping is the primary structure of the protein. Since the technique is sensitive to slight modifications of the protein that can occur as a consequence of post-translation, mistranslation, and crossover, point modifications at the single amino acid level can be identified [1,11]. The method gives a further opportunity to check batch consistency by comparison with a reference standard, thereby serving a purpose in quality control testing [3,12,13]. In order to meet the requirements of a quality-control method, it needs to be validated. Extensive reports of the validation of peptide mapping can be found in the literature [1,2,14,15], where various critical steps in the peptide mapping procedures are pinpointed.

Information on different kinds of modifications often has as a starting point a visual comparison of the chromatographic profile with that of a reference chromatogram, this traditionally being the sole means of information extraction. This approach to the detection of alterations in the chromatographic profile thus relies heavily on the expertise of the scientist involved. The chromatographic profile sometimes consists of up to hundreds of peaks and may undergo changes in the number of peaks, peak shapes, peak shifts and baseline patterns, depending on the type of alteration of the protein. Variations in the chromatographic conditions and in sample workup and digestion conditions may also alter the chromatographic profile.

Complex patterns, such as peptide maps, are generally considered to be interesting candidates for multivariate analysis aimed at clustering and classification. For instance, principal component analysis has been used for statistical validation of HPLC peptide mapping reproducibility and in a ruggedness study [14,15]. Woodward and Geiser [16] have used principal component multivariate visualisation for minimisation of baseline disturbance. Furthermore, in the papers of Malmquist [17] and Malmquist and Danielsson [18], the importance of data preprocessing in conjunction with HPLC peptide mapping and multivariate analysis has been pointed out. Malmquist and Danielsson [18] have shown that their approach of retention alignment by a cross-correlation function and a four-step procedure makes the principal component analysis feasible for peptide mapping.

It this study we address issues such as how to deal with variations in retention times, noise and baseline in order to facilitate multivariate analysis for detection and characterisation of alteration of generated peptide maps generated by reversed-phase liquid chromatography with single UV wavelength detection, without compromising the ability to identify modifications exposed by the protein. Techniques comprising, denoising and data compression, baseline correction and peak shift alignment, and multivariate visualisation were employed.

2. Experimental

2.1. Chemistry

Lactate dehydrogenase (LDH) variants were purchased from Boehringer Mannheim. Lysyl endopeptidase from *Achromobacter lyticus*M497-1 was obtained from Waco Chemicals. The LDH species, from porcine heart, bovine heart, porcine muscle and rabbit muscle, were delivered in 3.2 M ammonium sulphate. Prior to digestion, the buffer was changed to 6 M guanidine hydrochloride using a micro concentrator with a cut-off of 10 kDa. In an experiment, 100 μl of sample with a concentration of 0.5 mg/ml was centrifuged for 40 min at 7000 × g. The sample was washed twice with water and eluted with 6 M guanidine HCl by turning the concentrator upside down. LDH samples were preincubated for 30 min at 37 °C, diluted with 15 mM Tris–Buffer to give 2 M guanidine HCl and the pH was adjusted to 8.2. For the cleavage reaction, lysyl peptidase was added in an enzyme:protein ratio of 1:40 and the sample was incubated for 4 h at 37 °C. Adding a solution of 0.1% trifluoroacetic acid (TFA) stopped the reaction.

2.2. Analytical instrumentation

Fractionation of peptides from the digest was performed using a Merck-Hitachi HPLC with a diode array UV detector. The column was a Waters Symmetry C18, 4.6 mm × 50 mm, thermostatted at 33 °C. The gradient (with eluate A, 0.12% TFA and eluate B, 0.1 % TFA in acetonitrile) was optimised using an experimental design and run according to Table 1 to give an optimal separation. Single absorption measurement was performed at 214 nm at a flow rate of 1.0 ml/min.

Table 1

Time (min)	Acetonitrile (%)
0	7
6.5	13
17	25
28	46
29.5	80
32	80
33.5	7

For the experiment, four sets of samples of LDH from bovine heart, rabbit muscle, porcine heart and muscle, with nine samples of each species, were used and all samples were treated identically see Fig. 1.

3. Computational methods

All data processing was carried out using Matlab® (Mathworks, Inc., USA) either by routines written "in-house" and/or by the use of m-files in the toolboxes, Wavelets (Mathworks, Inc.), GA Toolbox (Department of Automatic Control and Systems Engineering, the University of Sheffield, UK) and PLS-Toolbox (Eigenvector, Inc., USA)

3.1. Denoising and data compression by wavelets

In order to minimise the influence of noise and to facilitate the optimisation procedure with the genetic algorithm, it was found advantageous to reduce

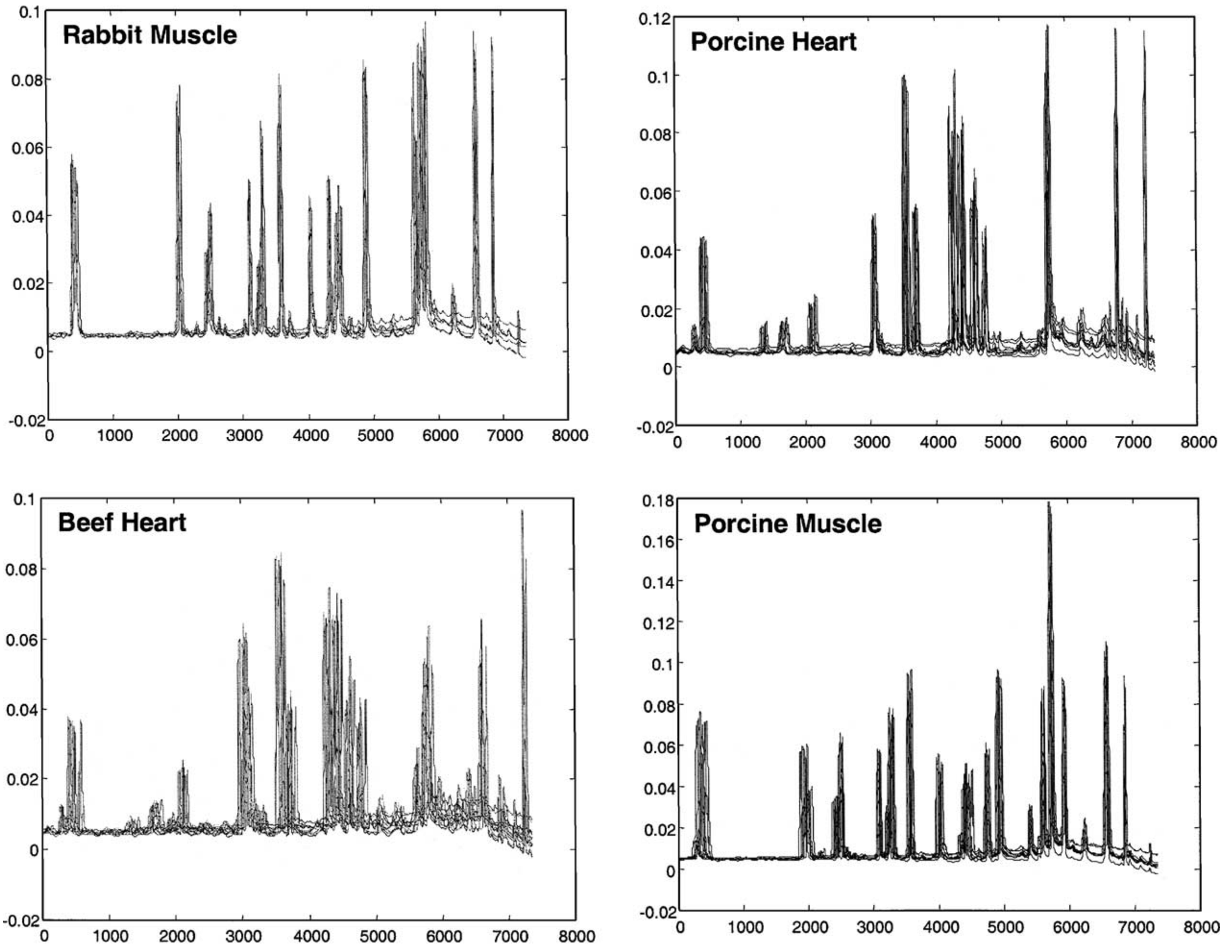

Fig. 1. Original chromatograms obtained by Lys-C digestion of LDH proteins. For each specific protein, chromatograms ($n = 9$) are superimposed on each other. x-Axis: chromatographic data points, y-axis: arbitrary UV response.

the data by wavelets. Wavelets are a relatively new technique in the field of chemometrics and consist of mathematical functions that divide the data into different frequency components, each component being studied with a resolution matched to its scale. Wavelets are analogous to the Fourier transform, but preserve the time information upon transformation and are more adapted to deal with non-stationary data [19].

The decomposition of the chromatogram was to a level 4 approximation with the ‘db1’ (Daubechies) wavelet. The data was compressed to a 1/16 of the original data (i.e. 7351data points to 460 data points).

3.2. Chromatographic baseline corrections

The differences in baselines, for instance slopes and levels were adjusted by the baseline correction algorithm, which divides the chromatogram into a number of equally spaced segments and finds the local minimum in each segment, followed by a baseline construction by linear interpolation of (segments + 2) data points. The chromatogram was divided into eight segments in this case. The calculated baseline was then subtracted from the chromatogram, respectively.

3.3. Normalization

Each chromatogram was divided with its maximum value, that is the peak with the largest peak area.

3.4. Peak shift alignments by genetic algorithms

Genetic algorithms are part of evolutionary programming, which is being increasingly used in a number of scientific fields [20,21]. In general, the genetic algorithm typically starts with randomly generated solutions to a defined problem. Each solution is evaluated according to its fitness, i.e. in our case being the difference between the actual chromatogram and a target chromatogram. The next step is to generate a new generation of solutions. The best solutions of the previous generation are moved to the new generation (selection). Some of the solutions are with a certain

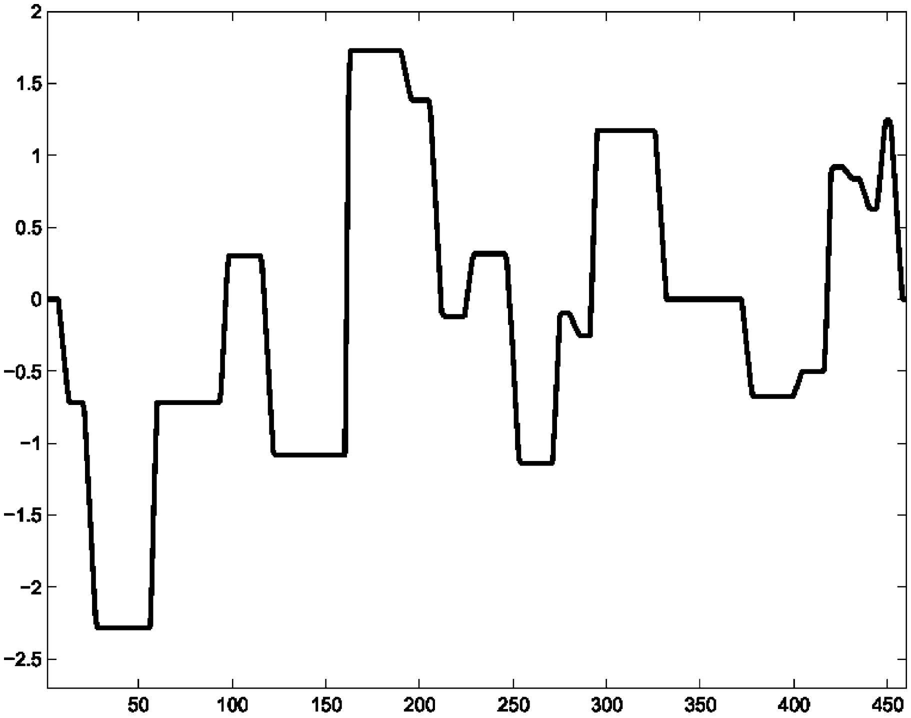

Fig. 2. Shift interpolation curve. *x*-Axis: chromatographic data points, *y*-axis: shifted chromatographic data points.

probability used as parents producing new children solutions by crossover. Finally, a number of mutations take place to generate an amount of randomness in the new solution, facilitate the algorithm to explore new areas of the solution space. Calculated for its fitness, the new generation is tested if the end condition is satisfied. If so, it is stopped and returns the best solution in the current generation, otherwise a new generation of solutions is generated.

In out approach, the peak alignment algorithm utilizes a shift interpolation curve to align the chromatogram toward a target chromatogram, see alignment scheme below:

3.5. Alignment scheme

Each chromatogram was aligned against a target chromatogram.

1. Rough alignment: a linear displacement of the chromatogram. Moving it forth or back a number of data points (≤abs(±4) dps).
2. Calculates the number of baseline points (*xp*), by finding the centre of regions in the chromatogram lacking peaks, that is peak areas are below a certain threshold. At the same time the *range* is set. In this study the *xp* corresponds to 16 positions and the *range* was ≤abs(±3 dps). The number of baseline point (*xp*) and its *range*, are then used to make the alignment curve. The number of *xp* decides how fine the alignment will be. More *xp* implies more detailed alignment. The *range* is set to avoid overlapping.
3. The genetic algorithm is then used to optimise the shape off the shift interpolation curve. (Fig. 2). The GA optimises the values of the *xp* within its allowed *range*. The shift interpolation curve is calculated by nearest neighbour interpolating the values in the *xp* to a vector with the same size as the chromatogram (*adc*). *adc* is smoothed with an average filter (5 dps) to remove the hard edges which can cause problem in following interpolation. A vector with the x positions of the chromatogram is added to *adc* (*adp* + 1:460). The

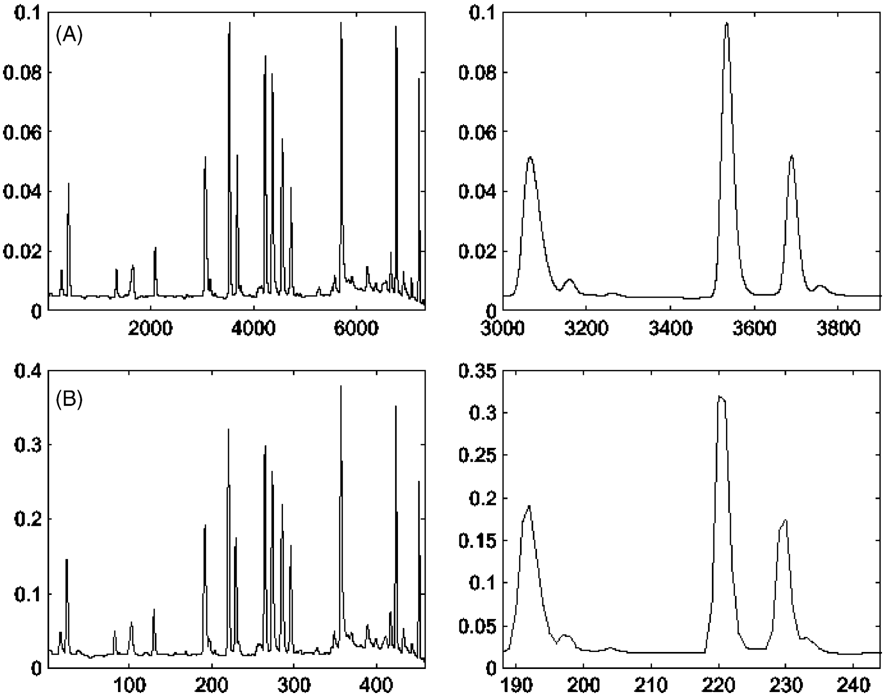

Fig. 3. A visual comparison between the original chromatographic data (A) and wavelet compressed data (B). The data were compressed to 1/16 of their original size.

chromatogram is peak shifted by linear interpolation with *adc*.

The GA used 60 individuals in 200 generations.

4. Chromatograms are fine adjusted in *y*-dimension (peak height) towards the target chromatogram, using GA. Maximum allowed adjustment is 5% of the total height. Implies that an additional value is given to the chromatogram.

   The GA used 20 individuals in 30 generations.

To summarize, the best shift interpolation curve, for each chromatogram, is generated (optimized) by the genetic algorithm, and subsequently used to align the chromatograms used in the following multivariate data analysis.

3.6. Principal component analysis

Principal component analysis (PCA), see for example reference [14,15,18], has been used throughout this study to visualise and to evaluate the results obtained on preprocessing of chromatographic data. Data were mean-centred prior to PCA. The lack of fit, Q, of the different PCA models was also calculated. Q is the

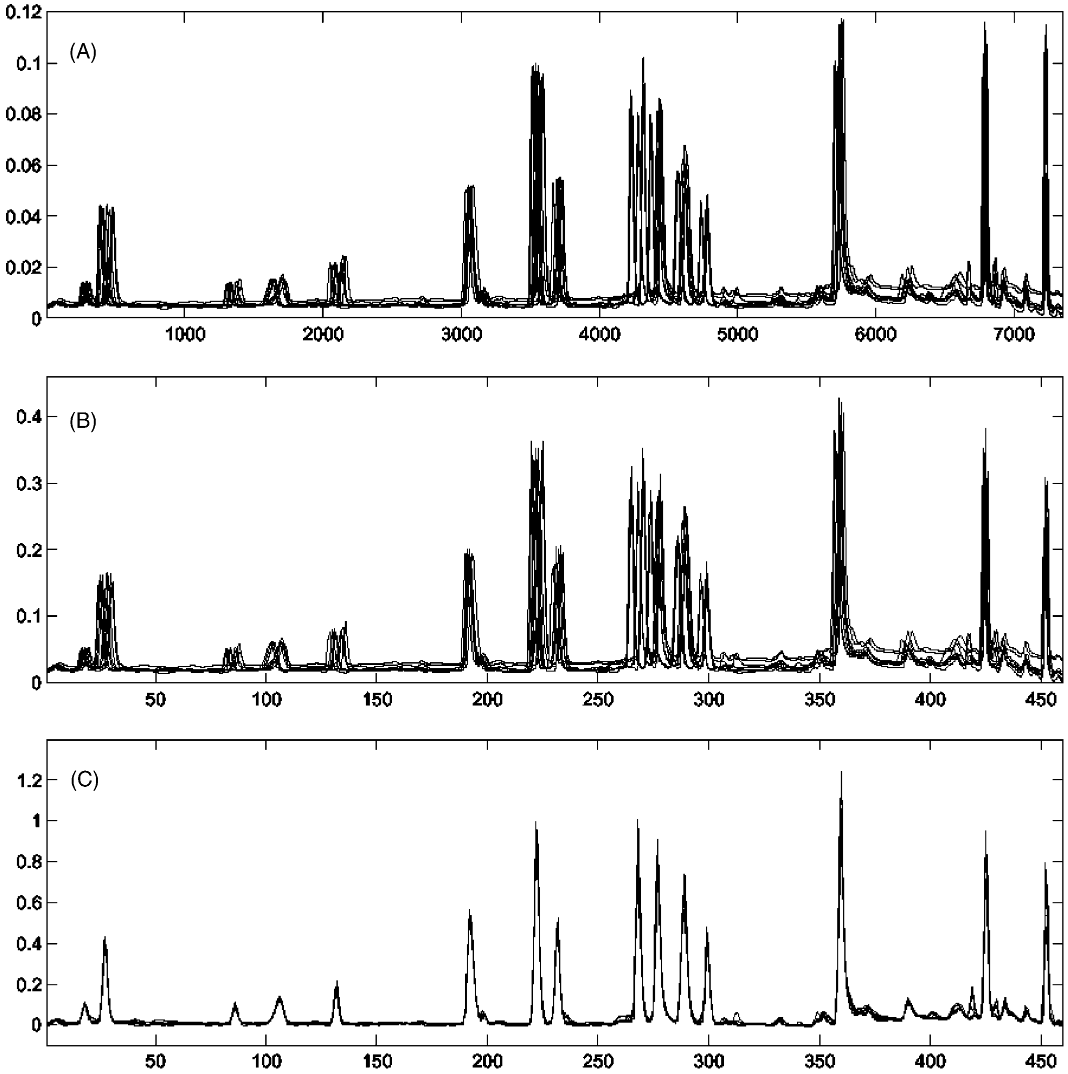

Fig. 4. Original chromatograms (A). Wavelet compressed chromatograms (B). Baseline-corrected and peak shift aligned chromatograms (C).

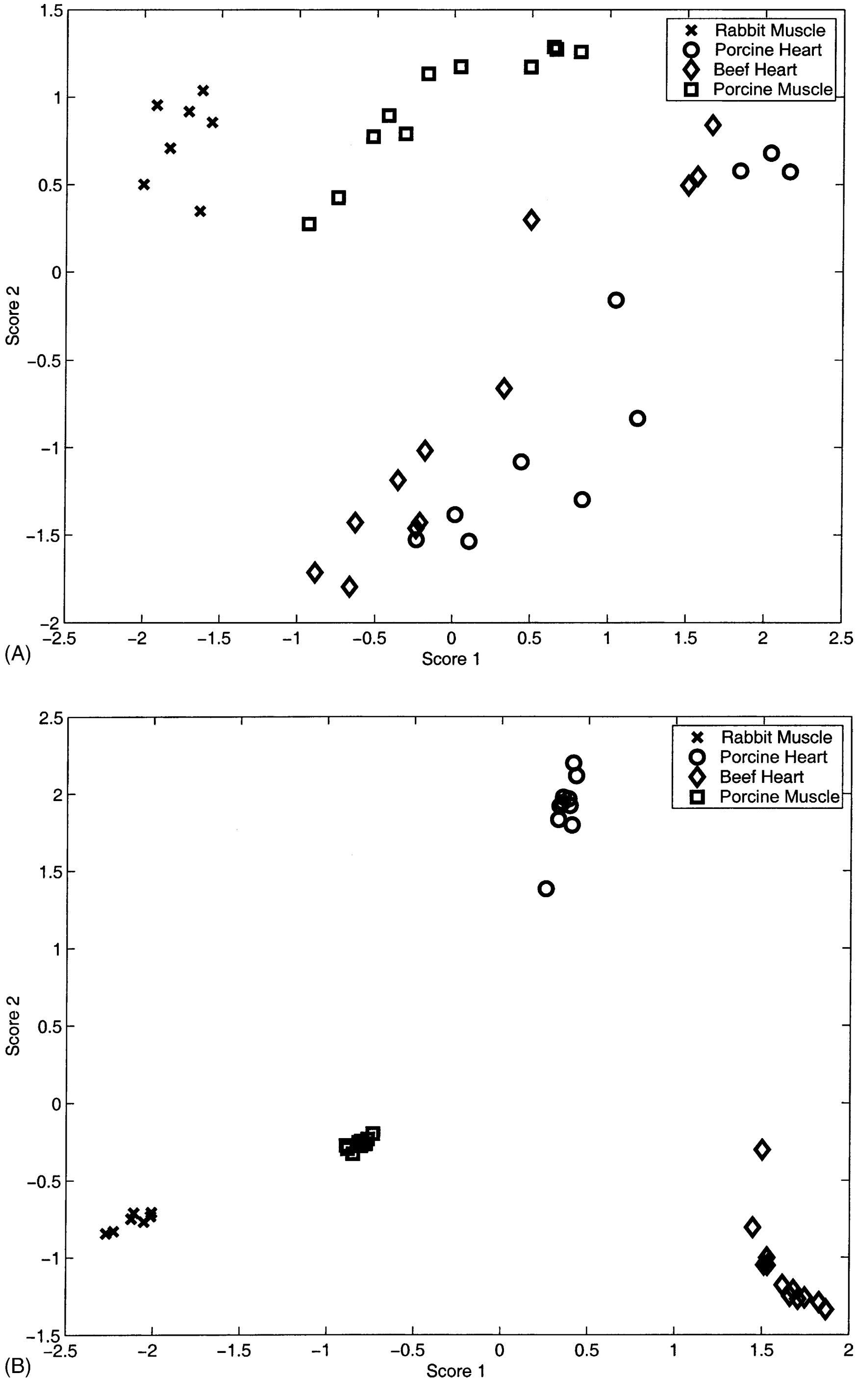

Fig. 5. Principal component score plots of protein samples before (A) and after preprocessing (B). In (A) the explained variance of PC1 is 22% and for PC2 19%, and for (B) the corresponding values are 41 and 3 %.

sum of squares of each sample (each row) of the residual matrix of the model, which allows for assignment of confidence limits for the overall residual Q of samples used to generate the PCA model as well as for new samples [22].

4. Results and discussion

The use of wavelets reduces the data to 1/16 of their original size, which was found to be advantageous when running the genetic algorithm as a result of reducing the search/optimisation space considerably and thus improving data processing times. The pay-off on data compression of approximately 94% is slight distortions of peak shapes and peak areas at the very beginning and ending part of the chromatograms (see Fig. 3) and loss of resolution. However, this effect was consistent and similar throughout the processed chromatograms and thus no extra variations were introduced by the data compression and denoising procedures.

Although simple in its approach, the baseline correction algorithm was found be highly efficient (see Fig. 4A and B) and deemed to be suitable for its intended use. The resulting chromatograms after peak alignment are shown in Fig. 4C.

The algorithms, including compression and baseline corrections, combined with the peak shift alignment, have been tested and evaluated on two basically different cases. The first case concerns the clustering ability for four different forms of the LDH proteins, which differ in either species origin or organ. As can be seen in Fig. 5, the score plots generated by a principal component analysis show a substantial improvement with regard to clustering properties for the preprocessed data in comparison to the original data. The score plots were chosen as the optimal ones for each condition, i.e. preprocessed or original data. Not only are the different LDH types completely separated

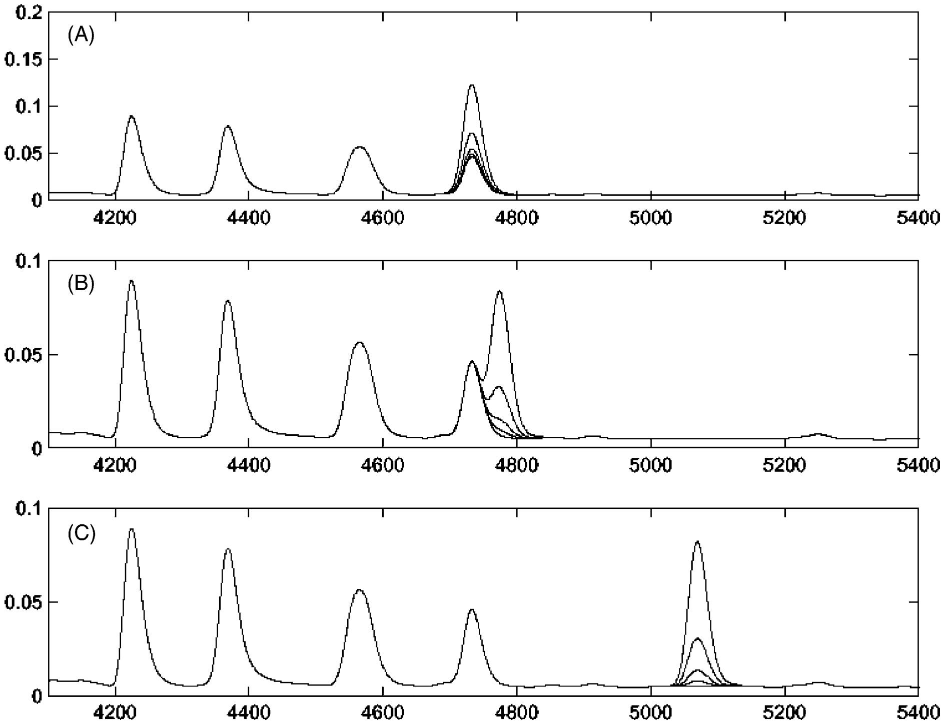

Fig. 6. Three cases of deliberately alternated chromatograms through the introduction of an extra peak. (A) Superimposing a peak, (B) doublet formation, and (C) an extra peak in a peak free region. Peaks were added corresponding to 0.15, 0.45, 1.34, and 4.03% of the total peak area of the original chromatogram.

from each other, but the information is also captured by the use of fewer principal components than in the case of the original data. This shows that the preprocessing algorithms substantially reduce variation in the data set while preserving information relevant to the different protein samples. Furthermore, outliers are more easily detected and diagnosed, as is evident in Fig. 5. As indicated in Fig. 5B, the "porcine heart" protein samples may contain one outlier. To examine this, principal component analysis was carried out for this particular protein sample in its original and in its preprocessed form. Using Q statistics ($\alpha = 0.05$) for outlier detection and a model of 2 PCs, it is possible to define one of the nine samples as being a true outlier in the case where the chromatograms have been subjected to preprocessing. A similar conclusion could not be drawn from the corresponding PCA model obtained for data not subjected to preprocessing. Careful visual inspection of that particular chromatogram substantiates the conclusion about it being an outlier. A likely explanation for one of the deviating samples is incomplete digestion of the protein, resulting in a slightly different pattern in the relative peak areas.

A more common case in a quality control situation was mimicked in the second case. Here one of the LDH proteins was subjected to deliberate changes, such as the addition of an extra peak. This was done in a systematic fashion by an increasing amount of variation introduced at each type of alteration. The idea was to arrive at some kind of threshold value for the peak detection/changes that could be introduced and allow a comparison to be made between preprocessed chromatograms and the original chromatograms. Three different cases were examined: (A) an extra peak was added to an existing peak in one of the chromatograms. The added peak had similar characteristics to that of the original peak in regard to peak shape and width. (B) An extra peak was added in close proximity to a peak in the original chromatogram. This resulted in a partial resolved doublet, with a resolution factor of 0.5 for equal sized peaks. (C) An extra peak was added in an area where the original chromatogram lacked peaks. The different cases are illustrated in Fig. 6. Peaks were added at the following values: 0.15% of the total integrated area of the original chromatogram, 0.45, 1.34 and 4.03%. In this study the outlier chromatogram (porcine heart) was also included to further validate the procedure for outlier detection. In the latter part of this study, PC models were generated for the different cases that accounted for approx. 96–97% of the explained variance (corresponding to seven PCs) or three PCs (corresponding to an explained variance of approximately 67–68%). Statistical analysis of unmodelled data, i.e. the residual (Q statistic) at $\alpha = 0.05$, were generated. Owing to the stochastic nature of the peak alignment procedure, a repeated number of

Table 2
Preprocessed chromatographic data

	Q 95% lim.[a]			
	Three PCs	Seven PCs		
A	0.0939	0.0176		
B	0.0925	0.0148		
C	0.0985	0.0137		
	3 PCs		7 PCs	
	Q mean	Q std	Q mean	Q std
Case A[b]				
0.00%	0.0480	0.0024	0.0171	0.0077
0.15%	0.0531	0.0024	**0.0205**	0.0097
0.45%	0.0789	0.0030	**0.0359**	0.0042
1.34%	**0.2416**	0.0071	**0.1615**	0.0050
4.03%	**1.5572**	0.0576	**1.3122**	0.0471
Outlier	**0.5356**	0.0035	**0.5358**	0.0149
Case B[b]				
0.00%	0.0631	0.0047	0.0091	0.0030
0.15%	0.0573	0.0147	0.0125	0.0045
0.45%	0.0743	0.0128	**0.0283**	0.0067
1.34%	**0.2065**	0.0157	**0.1776**	0.0212
4.03%	**1.4589**	0.0651	**1.4058**	0.1340
Outlier	**0.5378**	0.0137	**0.5267**	0.0037
Case C[b]				
0.00%	0.0534	0.0035	0.0153	0.0022
0.15%	0.0559	0.0059	0.0153	0.0052
0.45%	0.0844	0.0122	**0.0331**	0.0060
1.34%	**0.2255**	0.0078	**0.1747**	0.0023
4.03%	**1.5523**	0.0169	**1.4960**	0.0083
Outlier	**0.5279**	0.0077	**0.5135**	0.0115

Q statistics for PCA models, three and seven PCs, respectively. Case A, fully overlapped peak. Case B, partial overlap. Case C, fully resolved peak. See Fig. 6. Q statistics for the predicted results are generated from five separately preprocessed chromatographic data at each level of peak interference.

[a] Preprocessed chromatographic data. Generated PC models for Cases A, B, and C.

[b] Predicted results using generated PC models for preprocessed chromatographic data. Cases A, B, and C. Q means in bold represent values outside the Q-95% limit of the model considering one standard deviation.

Table 3
Original chromatographic data

	Q-95% lim.[a]	
	3 PCs	7 PCs
A	2.2446	0.2727
B	2.2446	0.2727
C	2.2446	0.2727
	Q	*Q*
Case A[b]		
0.00%	0.5241	0.1384
0.15%	0.5283	0.1390
0.45%	0.5488	0.1522
1.34%	0.7076	**0.2886**
4.03%	2.0568	**1.5669**
Outlier	**4.2648**	**3.0545**
Case B[b]		
0.00%	0.5241	0.1384
0.15%	0.5305	0.1433
0.45%	0.5555	0.1652
1.34%	0.7274	**0.3276**
4.03%	2.1165	**1.6850**
Outlier	**4.2648**	**3.0545**
Case C[b]		
0.00%	0.5241	0.1384
0.15%	0.5268	0.1409
0.45%	0.5446	0.1583
1.34%	0.6974	**0.3096**
4.03%	2.0480	**1.6561**
Outlier	**4.2648**	**3.0545**

Q statistics for PCA models, three and seven PCs, respectively. Case A, fully overlapped peak. Case B, partial overlap. Case C, fully resolved peak. See Fig. 6.

[a] Original chromatographic data. Generated PC models for Cases A, B, and C.

[b] Predicted results using generated PC models for original chromatographic data. Cases A, B, and C. *Q* means in bold are values outside the *Q*-95% limit of the model.

preprocessings took place ($n = 5$). In Tables 2 and 3 the results for the different cases are summarised. For the model based on seven PCs, using *Q* statistics, indications of the difference between original chromatographic data and peak added data show up at the 0.45% level and are clearly evident at the 1.34% level for all preprocessed chromatographic data (Table 2). In comparison, for the original data (i.e. unprocessed data) the corresponding levels are 1.34% for indications and 4.03% for clear evidence (Table 3). For models based on three PCs the difference between preprocessed and original data is even more pronounced. Preprocessed data models picks up alterations for all of the three cases at the 1.34% level. In contrast, for the original chromatographic data, only the outlier sample is outside the confidence level of the three PCs model. These results indicate that through the proposed preprocessing steps it is possible to obtain an enhancement of the information extraction and a higher sensitivity to outlier detection.

5. Conclusion

By combining methodologies based on genetic algorithms for peak shift alignments, wavelets for denoising and data compression, a baseline correction algorithm and principal component analysis for clustering and classification, small and subtle changes in peptide maps could be detected. For the peptide mapping data, detection of subtle differences between chromatograms, either due to incomplete digestion or deliberately added changes, was possible due to the preprocessing procedures applied. It was also possible to detect these changes at a lower level than in the corresponding original chromatographic data. It should be noted that such an approach easily lends itself to automation of the entire data evaluation process and is non-parametric in nature.

References

[1] D. Allen, R. Baffi, J. Bausch, J. Bongers, M. Costello, J. Dougherty, M. Federici, R. Garnick, S. Peterson, R. Riggins, K. Sewerin, J. Tuls, Biologicals 24 (1996) 255–275.

[2] J. Bongers, J.J. Cummings, M.B. Ebert, M.M. Federici, L. Gledhill, D. Gulati, G.M. Hilliard, B.H. Jones, K.R. Lee, J. Mozdzanowski, M. Naimoli, S. Burman, J. Pharm. Biomed. Anal. 21 (2000) 1099–1128.

[3] R.L. Garnik, N.J. Solli, P.A. Papa, Anal. Chem. 60 (1988) 2546–2557.

[4] R.C. Judd, in: J.M. Walker (Ed.), The Protein Protocol Handbook, Humana Press, Totowa, NJ, 1996, pp. 453–455.

[5] E. Skrzydlewska, J. Ostrowska, Z. Figaszewski, Acta Chromatogr. 8 (1998) 70–83.

[6] M.J. Cox, R. Shapira, K.D. Wilkinson, Anal. Biochem. 154 (1986) 252–345.

[7] K. Hirayama, R. Yuji, N. Yamada, K. Kato, Y. Arata, I. Shimada, Anal. Chem. 70 (1998) 2718–2725.

[8] S. Hess, F.J. Cassels, L.K. Pannell, Anal. Biochem. 302 (2002) 123–130.

[9] V. Egelhofer, K. Bussow, C. Luebbert, H. Lehrach, E. Nordhoff, Anal. Chem. 72 (2000) 2741–2750.

[10] D.K. Bryant, S. Monte, W.J. Man, K. Kramer, P. Bugelski, W. Neville, I.R. White, P. Camilleri, Rapid Commun. Mass Spectrom. 15 (2001) 418–427.

[11] C.M. Smales, D.S. Pepper, D.C. James, Biotechnol. Bioeng. 67 (2000) 177–188.

[12] M. Wan, F.Y. Shiau, W. Gordon, G.Y. Wang, Biotechnol. Bioeng. 62 (1999) 485–488.

[13] N. Lundell, T. Schreitmuller, Anal. Biochem. 266 (1999) 31–47.

[14] K.R. Lee, J. Bongers, B.H. Jones, S. Burman, Drug Dev. Ind. Pharm. 26 (2000) 123–134.

[15] K.R. Lee, J. Bongers, D. Gulati, S. Burman, Drug Dev. Ind. Pharm. 26 (2000) 1045–1057.

[16] C. Woodward, F.O. Geiser, J. Chromatogr. A 699 (1995) 11–20.

[17] G. Malmquist, J. Chromatogr. A 687 (1994) 89–100.

[18] G. Malmquist, R. Danielsson, J. Chromatogr. A 687 (1994) 71–88.

[19] B.K. Alsberg, A.M. Woodward, D.B. Kell, Chemometr. Intell. Lab. Syst. 37 (1997) 215–239.

[20] R. Wehrens, L.M.C. Buydens, TRAC-Trends Anal. Chem. 17 (1998) 193–203.

[21] J. Forshed, I. Schuppe-Koistinen, S.P. Jacobsson, Anal. Chim. Acta 487 (2003) 189–199.

[22] J.E. Jackson, G.S. Mudholkar, Technometrics 21 (1979) 341–349.