ELSEVIER

Journal of Chromatography A, 687 (1994) 89–100

JOURNAL OF CHROMATOGRAPHY A

# Multivariate evaluation of peptide mapping using the entire chromatographic profile☆

Gunnar Malmquist[1]

*Institute of Chemistry, Department of Analytical Chemistry, Uppsala University, P.O. Box 531, S-751 21 Uppsala, Sweden*

First received 29 October 1993; revised manuscript received 2 August 1994

**Abstract**

Peptide mapping is an important analytical technique for quality control of rDNA-derived proteins. The evaluation in peptide mapping is complicated by variations in the digestion and the chromatographic separation. The variation sources in peptide mapping are briefly reviewed. A multivariate evaluation method that can account for the digestion variations is presented. The method utilizes the entire chromatographic profile as input data, eliminating the need for peak size and retention time determinations. The influence of chromatographic variations should be reduced by proper pretreatment of the chromatograms, in order to allow classification of protein samples. The method is intended to facilitate the evaluation in peptide mapping and is capable of handling numerous chromatograms in a data set.

## 1. Introduction

Many therapeutically important peptides and proteins, e.g., insulin and growth hormone, are currently produced by the recombinant DNA (rDNA) technique [1]. This sophisticated technology is based on insertion of foreign genetic material, coding for the substance of interest, into a host cell. The host cells will then produce the desired substance in addition to their natural production. Biotechnological production of pharmaceuticals requires very rigorous quality control owing to the risk of undesirable protein modifications and the numerous possibilities for contamination of the product. The integrity of the amino acid sequence of the protein has to be confirmed for each production batch. An introduction to the rDNA technique and the analytical aspects of quality control in biotechnology was given by Garnick et al. [2].

Peptide mapping is an indispensable analytical method in biotechnology for quality control of rDNA-derived proteins [2]. Peptide mapping consists of fragmentation of the protein by enzymatic digestion or chemical cleavage, with subsequent separation of the fragments. The fragmentation is in most instances performed by enzymatic digestion with trypsin [3], while the separation is usually performed by gradient elution reversed-phase liquid chromatography

☆ Parts of this material were previously presented at *Analysis of Peptides, Stockholm, 2–4 June, 1993*, and *Analysdagarna, Lund, 14–18 June 1993*.

[1] Present address: Pharmacia Biotech, R&D, S-751 82 Uppsala, Sweden.

0021-9673/94/$07.00 
*SSDI* 0021-9673(94)00727-6

(RPLC) [4]. Digestion with trypsin gives specific cleavage of the protein at the C-terminal side of arginine and lysine residues, resulting in a large number of fairly short fragments with average size 7–12 residues [3]. This requires a high peak capacity in the chromatographic separation and often necessitates the use of segmented gradients. The high resolving power obtained with shallow gradients in RPLC [5] is one of the main reasons for the popularity of RPLC in peptide mapping.

The chromatogram can be regarded as a fingerprint of the protein, where the overall appearance is used to assess the integrity of the amino acid sequence. Modification of one amino acid will alter the properties of one fragment, which may be detected as a change in retention for that fragment. For instance, substitution of a single amino acid in tissue-type plasminogen activator ($M_r \approx 64\,000$) leads to a significant change in the retention of one fragment [2]. The evaluation of peptide mapping is traditionally performed by visual comparison of the sample chromatogram with a reference chromatogram of a digested protein with the correct sequence.

Both mutations in the DNA sequence and translation errors may lead to incorporation of erroneous amino acids in the protein. Further variants can be formed by post-translational modifications of the protein, mainly by degradation processes. Proteolytic or chemical cleavage of the protein, oxidation of methionine residues and deamidation of asparagine residues all give rise to protein variants that may be detected by peptide mapping [1].

Other applications of the technique include the characterization of naturally occurring protein variants [6] and the identification of animal species [7]. One possible application is the detection of contaminants in the sample, but this approach has not yet been much employed, owing to the difficulty to detect small peaks in a complex chromatogram containing 20–150 peaks [8].

There are several sources of variations associated with the enzymatic digestion of protein samples, that may lead to variations in the fingerprint even for identical samples. For trypsin digestion, such deviations from the normal cleavage pattern have been reported to arise from chymotryptic cleavages, partial digestion and incomplete digestion [3,9]. The final fingerprint may also be influenced by incomplete cysteine reduction or alkylation [10]. Deamidation of asparagine or glutamine residues can be induced after the digestion, e.g. by improper storage of the digests in a non-refrigerated autosampler [10].

The commercial trypsin preparations are usually treated with L-1-tosylamide-2-phenylethyl chloromethyl ketone (TPCK) in order to reduce the chymotryptic activity in the preparation. Chymotryptic cleavages are nevertheless frequently observed, owing to the minute amounts of chymotrypsin that may remain despite the TPCK treatment [11]. Non-specific cleavage fragments can occasionally be caused by trypsin cleavages at less favourable sites, e.g., adjacent to proline residues [12]. The cleavage pattern may vary between trypsin preparations obtained from different manufacturers, and even between batches from the same vendor [11].

Chloupek et al. [9] showed an interesting example of variations in the amount of non-specific cleavage fragments. The additional peaks may reduce the possibility to detect contaminants in the sample, and should therefore be kept at a minimum. The amount of non-specific cleavages could unfortunately not be reduced by changes in the digestion conditions, e.g., buffer type, temperature and reaction time. Purification of TPCK-treated trypsin by RPLC did not reduce the chymotryptic activity either.

Partial digestion will take place if the protein contains a series of two to four adjacent basic amino acids, all potential cleavage sites. When cleavage at a random position within this series has occurred, trypsin will not cleave the terminal amino acids. This will lead to the formation of varying amounts of overlapping fragments, differing in the first and last positions [3]. Variations in the yield of the digestion leads to different amounts of the resulting fragments, and a variable amount of undigested protein remaining in

the sample. Incomplete digestion may also lead to formation of partially digested fragments.

Other important sources of variation in peptide mapping are connected with the chromatographic separation. Preparation of fresh mobile phases will inevitably introduce small differences in the pH and possibly in the amount of organic modifier. The retention of peptides in RPLC is very sensitive to the composition of the mobile phase [13]. Changes in the amount of mobile phase additives, e.g. ion-pairing agents, may influence the peak retention [14]. The reproducible generation of shallow gradients is difficult even with modern LC instrumentation [15]. This may lead to slight retention shifts, especially at the beginning of the gradient [2]. Temperature variations [14], the gradual degradation of column performance [11,16] and column to column differences [17] are additional possible sources of variations in the profile. Dong and Tran [4,8] have provided recommendations for reproducible chromatographic separations of tryptic digests.

The possible deviations from the expected cleavage pattern, together with the retention variations caused by the chromatographic process, implies that peptide mapping is a very demanding analytical technique. Development and validation of a successful peptide mapping method require great effort, where the expertise of both biochemists and analytical chemists is necessary. Despite these problems, peptide mapping is the most important technique for assessment of the amino acid sequence integrity in proteins.

The visual comparison of peptide mapping chromatograms is complicated by variations in the digestion and the chromatographic separation. The evaluation will be more or less subjective and requires great experience. A more unbiased evaluation can be made by multivariate pattern recognition methods capable of handling the experimental variations. Pattern recognition [18] is a category of chemometric methods suited for the characterization of complex data sets. A multivariate evaluation method for peptide mapping is proposed in this paper, where test samples are classified by SIMCA [19], a multivariate classification method based on principal component analysis (PCA).

## 2. Multivariate evaluation of peptide mapping

A set of reference chromatograms, obtained for digests of samples with the correct sequence, is accumulated. The data set should cover the normal variations encountered in both the digestion and the chromatographic separation. The chromatograms are represented by the entire profile, i.e., the digitalized detector signal where each sampled data point corresponds to one variable in the data set. This is advantageous for evaluation of peptide mapping, as discussed in the accompanying paper [20]. This data set can be characterized by PCA [21], expressing the main variations in the data set. Chromatograms of test sample digests can subsequently be classified by the pattern recognition method SIMCA [19]. A brief introduction to SIMCA is given below to facilitate the discussion on the evaluation method.

Multivariate analysis of chromatographic profiles requires that the chromatographic variations are reduced. The possible differences between the samples will be obscured by the chromatographic variations, unless proper pretreatment is performed. In the accompanying paper [20], this was illustrated by simulated data and a peptide mapping data set. A method for pretreatment of chromatographic profiles, intended to remove slight retention shifts caused by the chromatographic process, was developed. Compensation for variations in the injected amount was made by a selective normalization procedure that also allowed correction for baseline differences.

### *2.1. Multivariate classification with SIMCA*

SIMCA, an acronym for soft independent modelling of class analogy, is a multivariate classification method based on PCA [19]. The classification is based on a model of the similarities between the known members of a class in

a training set. The training set in this particular instance is composed of the accumulated reference maps, all of which belong to the same class. Each class is described by a few principal components calculated for the members of the class in the training set only. Objects, i.e., chromatograms, that belong to a class will be situated close to the hyperplane spanned by the principal components. A tolerance region is established for the class, in order to quantify the similarity between a test sample and the members of the class. The tolerance region defines the boundaries for the residuals and the normal range of scores along the principal components. This is illustrated in Fig. 1, where a hypothetical class described by two principal components is depicted. Outliers and non-members will be situated outside the tolerance region.

Classification of tryptic digests based on their chromatographic profiles constitutes a special case of SIMCA, where only the acceptable reference samples in the training set form a proper class of similar objects. Deviating test samples can be different in many disparate ways, leading to the formation of an asymmetric class

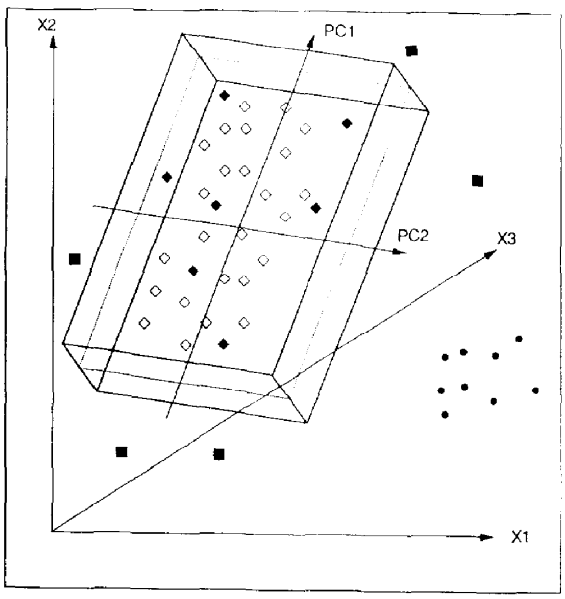


Fig. 1. SIMCA class model and tolerance region. Unfilled symbols denote objects in the training set and filled symbols refer to test set objects.

[22]. SIMCA can cope with this situation by assigning test objects to the acceptable class only if they fall within the tolerance region. All objects falling outside the region are referred to the asymmetric class of unacceptable samples.

The classification of a test sample by SIMCA is performed with an $F$-test according to Wold et al. [22] and Sharaf et al. [23]:

$$F = \frac{s_i^2}{s_0^2} \tag{1}$$

where $s_i^2$ is the class distance for test object $i$, composed of the residual variance and the deviation in scores, and $s_0^2$ is the residual variance of the class members in the training set. The $F$-value is compared with a critical $F$-value where the degrees of freedom can be calculated from the number of variables, $M$, the number of objects in the training set, $N$, and the number of principal components used to describe the class, $A$. If the $F$-value for the test sample exceeds the critical $F$-value, the sample is classified as an outlier or non-member of the class.

The critical $F$-value can be chosen with $(M - A)$ and $(N - A - 1)(M - A)$ degrees of freedom in the numerator and the denominator, respectively [23]. In data sets with many variables this will lead to a very low critical $F$-value that will reject many acceptable samples. Gemperline et al. [24] addressed this problem and proposed another method to calculate the degrees of freedom for classification of test samples. When a single test sample is classified, the appropriate degrees of freedom should be one in the numerator and $(N - A - 1)$ in the denominator. The critical $F$-value is calculated for one-tailed tests both at the 0.10 and 0.05 significance levels. If the $F$-value for the test sample exceeds the critical value at the 0.05 level it is rejected, i.e. classified as unacceptable. The sample will be assigned as an outlier if the $F$-value is between the 0.10 and 0.05 levels. Test samples with $F$-values below the 0.10 level are regarded as acceptable. This approach was used in this paper for classification of the tryptic maps, together with a graphical presentation of classification results, as discussed below.

## 3. Experimental

### 3.1. Tryptic digests

The tryptic digests of equine and bovine cytochrome *c* (Sigma, St. Louis, MO, USA) were prepared according to the procedure described by Renlund et al. [25], with the exception that the concentration of trypsin (Sigma) was decreased to 0.2 $\mu$g/$\mu$l [26]. The procedure was also scaled up fivefold, by increasing the volumes in all steps. Four replicate preparations of the reagents for denaturation and cysteine reduction, desalting buffer and the trypsin solution were used.

### 3.2. Chromatographic procedure

The tryptic digests were injected by a CMA 200/240 refrigerated (4°C) autosampler (CMA Microdialysis, Stockholm, Sweden), and separated on a SuperPac Pep-S $C_2/C_{18}$ (5 $\mu$m, 100 Å) column (250 × 4 mm I.D.) using a precolumn (10 × 4 mm I.D.) packed with the same material. The separations were performed with a Model 2249 gradient pump with detection at 215 nm with a Model 2141 variable-wavelength monitor. The chromatographic system was controlled by HPLCmanager software, also used to store the chromatograms prior to the multivariate analysis. All chromatographic columns and instruments were from Pharmacia Biotech (Uppsala, Sweden) except where indicated.

The separations were performed by gradient elution (flow-rate 1 ml/min). The mobile phases were consistently prepared by weighing instead of volumetric measurements. The aqueous phase (A) consisted of 50 m*M* phosphate buffer (pH 2.5), prepared by mixing fixed amounts of stock solutions of phosphoric acid and sodium dihydrogenphosphate (both from Merck, Darmstadt, Germany). The organic phase (B) consisted of acetonitrile–A (80:20). The acetonitrile was of gradient grade (Merck). The mobile phases were degassed by sparging with helium for 5 min (A) and 10 min (B). The samples (125 $\mu$l) were eluted with a linear gradient from 0 to 60% B in 96 min, corresponding to a gradient slope of 0.5% acetonitrile/ml.

All calculations were implemented in the programming environment ASYST (Macmillan Software, New York, USA).

## 4. Results and discussion

The chromatographic separation in this paper was not optimized with respect to the resolution of the fragments. The composition of the mobile phase and the shape of the gradient were merely chosen to give acceptable resolution within a reasonable analysis time. The proposed evaluation method is intended to facilitate the interpretation of the peptide mapping results, possibly without extreme requirements regarding the resolution. However, special precautions were taken to minimize the variations in the mobile phase composition. The critical aspect of the mobile phase composition in this context is reproducibility, not accuracy. It is not so important whether the pH of the aqueous phase is 2.50 or 2.55, as long as it is consistent throughout the data set. The pH of the aqueous buffer in the mobile phase is often established by titration of the acid with a base until the desired pH is achieved. Higher precision in the pH may be obtained by instead mixing stock solutions of the acid and the corresponding salt. The mobile phase preparation in this study was entirely based on weighing instead of volumetric measurements, in order to increase the reproducibility.

### 4.1. Description of the data sets

The training set of reference tryptic maps consisted of 50 objects (chromatograms), each described by 4900 variables, i.e., sampled data points 0.8 s apart. The training set originates from the set of 54 reference chromatograms (27 digests, each chromatographed twice) that had been characterized by PCA [20]. Two digests (corresponding to chromatograms 9, 23, 29 and 42 in the original set) were excluded from the training set.

The evaluation method was tested by simulated spiking, where one of the excluded chromatograms (29) was used as a template. Test chromatograms were obtained by addition of one Gaussian peak to the original chromatogram. The peak width of the added peak was set approximately equal to the width of the original peaks, i.e., with $\sigma = 0.1$ ml. Twenty-seven distinct peaks in the template were selected for the simulated spiking (see Fig. 2). Three positions for each selected template peak, corresponding to complete co-elution and shoulder peaks on the leading and trailing edge of the template peak, were independently used as the retention volume of the added Gaussian peak. The theoretical resolution between the template peak and the added shoulder peak was 0.5. Twelve baseline positions were also selected to estimate the level of detection for well resolved peaks. The height of the added peak was in all instances set at 3, 5, 7 and 10% of the largest peak in the template (hereafter refered to as 3–10% FS, respectively). The use of 91 peak positions and four peak heights at each position led to a test set consisting of 364 chromatograms, each containing one added Gaussian peak. The chromatograms were pretreated according to the previously described procedure [20] and subsequently classified by SIMCA.

### 4.2. Multivariate classification

The training set had been characterized by PCA [20], and cross-validation [27] indicated that five principal components was optimum for the description of the training set. Inclusion of additional components in the class model may improve the classifications by SIMCA, however. Nevertheless, it is very important not to include too many components in the class model, in order to avoid bad classification results for new samples. Gemperline et al. [24] suggested that the number of components could be determined by the optimum classification accuracy. Two types of classification errors are possible which are conceptually related to the Type 1 and Type

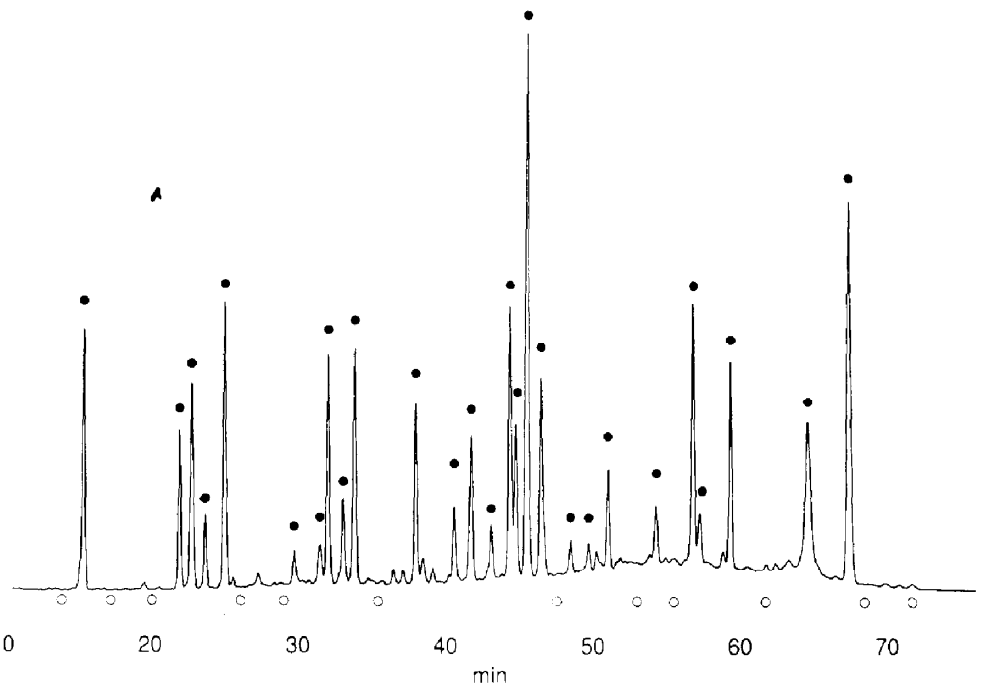


Fig. 2. Simulated spiking. A Gaussian peak is added at one position in each test set chromatogram. ● = Peak positions; ○ = baseline positions. Chromatographic conditions as in Experimental.

2 errors in significance testing that are discussed in textbooks on statistics, e.g., by Miller and Miller [28]. Error of Type 1 refers to rejection of a true null hypothesis. In the present case, this corresponds to rejection of reference chromatograms, i.e., $F$-values above the critical $F$-value at the 0.05 significance level. Inclusion of additional principal components in the class model will increase the risk of Type 1 errors as the residual variance of the class members in the training set will decrease (cf., Eq. 1). Error of Type 2, on the other hand, refers to a failure to reject a false null hypothesis, corresponding to acceptance of deviant chromatograms ($F$-value below the critical $F$-value at the 0.10 significance level). The risk for Type 2 errors will be reduced by increasing the number of principal components in the class model.

An extended cross-validation, or jack-knife, procedure was carried out to assess the classification accuracy, and to find the optimum balance between the risks for Type 1 and Type 2 errors. Independent reference chromatograms were obtained by successively excluding two acceptable digests, i.e., four chromatograms, from the training set (for this procedure, the original set of 54 reference chromatograms was used). The principal components were calculated for the remaining objects, and the excluded objects were classified according to the classification rules outlined above. This was repeated until all chromatograms in the training set had been excluded once. Fig. 3 shows the percentage of Type 1 errors from the extended cross-validation as a function of the number of components in the class model. Class models with five or six components resulted in 100% acceptance of the reference chromatograms, i.e., no Type 1 errors. No reference chromatograms were rejected if up to nine components were used in the class model; however, three chromatograms were classified as outliers.

A test set of deviant chromatograms were obtained by the simulated spiking, where one Gaussian peak had been added to the template. Fig. 4 shows the percentage of Type 2 errors for test set chromatograms with different peak heights for the Gaussian peak added as complete

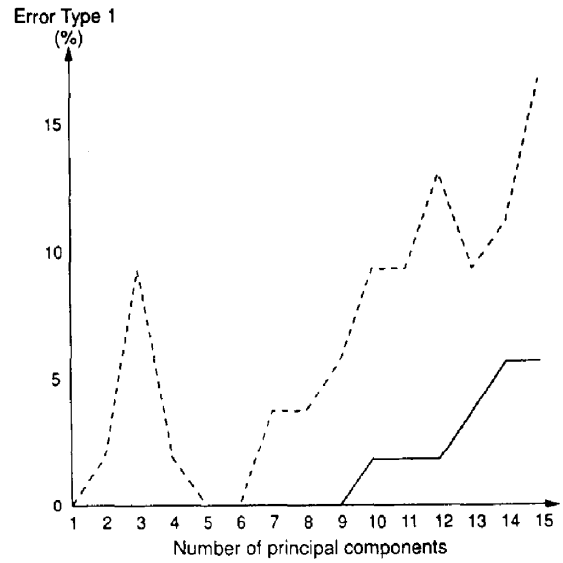


Fig. 3. Percentage of Type 1 errors, i.e., erroneous rejection of reference chromatograms. The dashed line indicates the percentage of the chromatograms that were either rejected or classified as outliers. The solid line refers to rejected reference chromatograms.

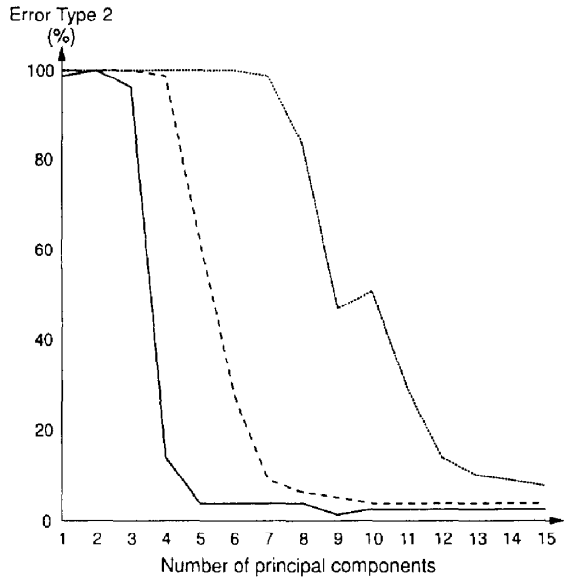


Fig. 4. Percentage of Type 2 errors, i.e., erroneous acceptance of spiked chromatograms in the test set. The solid line refers to simulated spiking with a peak height corresponding to 10% FS. The dashed and dotted lines refer to peak heights of 7 and 5% FS, respectively.

co-elution or as a shoulder on an existing peak. Chromatograms with a Gaussian peak corresponding to 10% of the largest peak in the template chromatogram are in most instances detected as deviant already with five components in the class model, while additional components are necessary to detect the smaller peaks. A class model with nine components will detect the majority of the test chromatograms where a 7% FS peak was added and about half of the chromatograms with a 5% FS peak.

Combination of these results indicates that a class model with nine principal components provides the best balance between the two types of classification errors, and is therefore considered optimum for classification purposes.

All 364 objects in the test set were classified, using nine principal components in the class model. The results are summarized in Table 1.

Table 1
Classification by SIMCA with nine principal components in the class model

| Type of object[a] | Classification results[b] | | |
|---|---|---|---|
| | Rejected[c] | Outliers | Accepted |
| Reference | 0 | 5.6 (3) | 94.4 (51) |
| Non-resolved, 10% FS | 96.2 (76) | 2.5 (2) | 1.3 (1) |
| Non-resolved, 7% FS | 77.2 (61) | 17.7 (14) | 5.1 (4) |
| Non-resolved, 5% FS | 0 | 53.2 (42) | 46.8 (37) |
| Non-resolved, 3% FS | 0 | 0 | 100 |
| Baseline, 10% FS | 100 | 0 | 0 |
| Baseline, 7% FS | 100 | 0 | 0 |
| Baseline, 5% FS | 0 | 100 | 0 |
| Baseline, 3% FS | 0 | 0 | 100 |

[a] Reference refers to the 54 training set samples where the classifications were made by the extended cross-validation procedure. Non-resolved refers to chromatograms where the Gaussian peak was added as co-eluting or shoulder. Baseline refers to simulated spiking at baseline positions.

[b] The classification results are expressed in percent and the corresponding number of objects are indicated in parentheses where appropriate.

[c] Rejected refers to objects with $F$-values larger than the critical $F$-value at the 0.05 significance level. Outliers are objects with $F$-values between the critical $F$-value at the 0.05 and 0.10 significance levels. Objects with $F$-values below the critical $F$-value at the 0.05 significance level are accepted.

The peak-height level necessary for detection in case of co-elution or low resolution is about 7% relative to the largest peak in the chromatogram (see Fig. 5). Co-elution with a peak that has a large variation in the training set will reduce the possibility for detection. The majority of the objects on the 7% FS level that were erroneously accepted were spiked at the broad peak eluted at approximately 64 min. This peak exposed the largest variation in the training set, as revealed by its dominant role in the loadings on the first principal component [20]. New peaks at the 5% FS level may be detected if they are baseline resolved, although only as outliers. The detection limit can thus be improved by optimization of the chromatographic separation. Segmented gradients could be utilized to increase the resolution between the fragments in some parts of the chromatogram. This may increase the probability that a new peak will be well resolved from the original peaks.

The method was also tested with mixtures of bovine and equine cytochrome *c*. The two protein forms are phylogenetically related [29], but differ in three amino acid positions out of 104 [30]. The pure forms of the proteins are easily distinguishable by visual comparison of tryptic maps (see the two upper chromatograms in Fig. 6). The ability of the method to detect amino acid sequence modifications was assessed by adding a small amount of bovine cytochrome *c* (1–25%) as an impurity to the equine protein. The bottom chromatogram in Fig. 6 shows a peptide map of a sample where about 7% bovine protein had been added. This sample was classified as unacceptable by the proposed method, as were samples with larger amounts of the impurity. Smaller amounts could not be distinguished from the normal variability of the peptide mapping method.

### 4.3. Practical considerations

A serious problem with all chromatographic fingerprinting methods is the influence of the gradual degradation of column performance. This has been observed to affect the outcome of multivariate characterization methods in

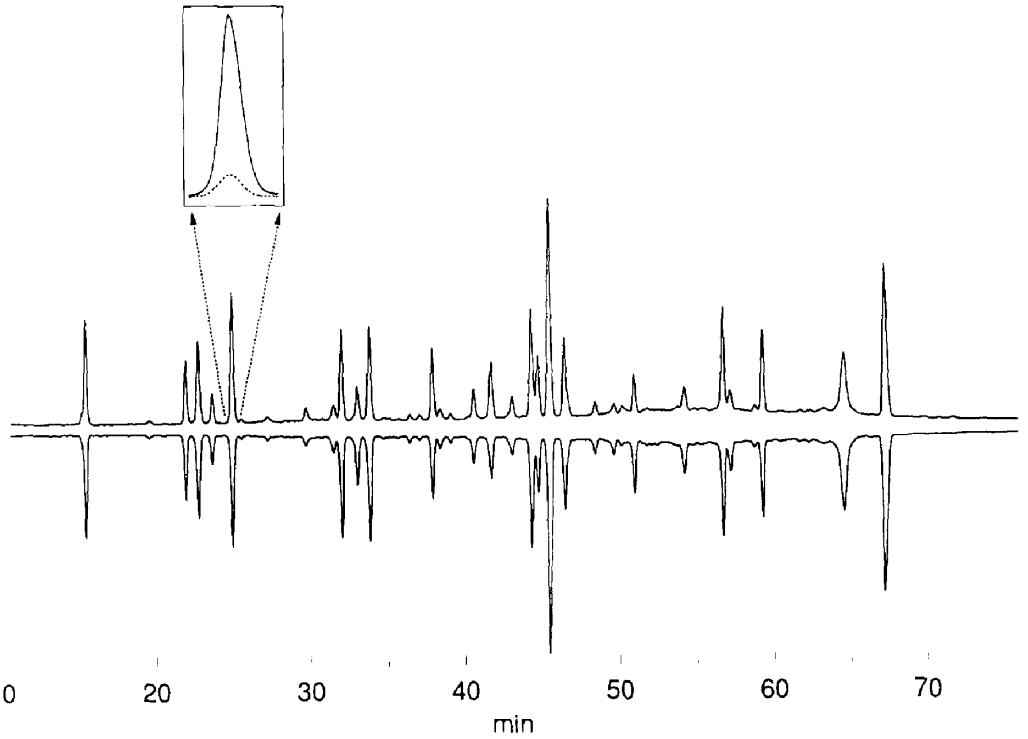


Fig. 5. Top: simulated spiking with addition of a Gaussian peak corresponding to 7% of the largest peak in the chromatogram. Bottom: reference chromatogram. Both chromatograms are shown after the pretreatment to facilitate the comparison. The inset shows the position and size of the added peak. Chromatographic conditions as in Experimental.

pyrolysis–GC [31] and classifications in GC [32]. The effect of stationary phase degradation on predictions in experimental design has recently been demonstrated [33]. The normal chromatographic pattern in peptide mapping is also influenced by column degradation [11,16]. Proper experimental precautions should be taken to maximize the stability of the column, e.g. by using high-purity mobile phase additives and regular cleaning of the column to remove any adsorbed contaminants. The acidic mobile phases commonly used in peptide mapping may cause column degradation introduced by cleavage of the bonded groups or end caps, exposing the silanol groups [8]. The most common mobile phase system in peptide mapping is based on trifluoroacetic acid (TFA) added both to the aqueous and organic components of the mobile phase. The volatility of the TFA system is a distinct advantage, facilitating mobile phase removal [4]. The aqueous part of the mobile phase in this study consisted of a phosphate buffer (pH 2.5), known to give different selectivity for tryptic fragments compared with TFA-based mobile phases [8,12]. The phosphate buffer is less acidic than TFA-based mobile phases (typical pH 2.0), which might reduce the column degradation, and thus possibly give better long-term reproducibility. It may also be worthwhile to explore the possibilities of polymer-based or zirconium-based reversed-phase columns, which have been claimed to have better pH stability than silica columns [4,34].

The traditional approach in peptide mapping is to produce a reference map together with each new test sample map, and make a visual comparison of the two chromatograms. The multivariate evaluation method is not intended to replace the visual inspection of the chromatograms, but to produce a less subjective criterion

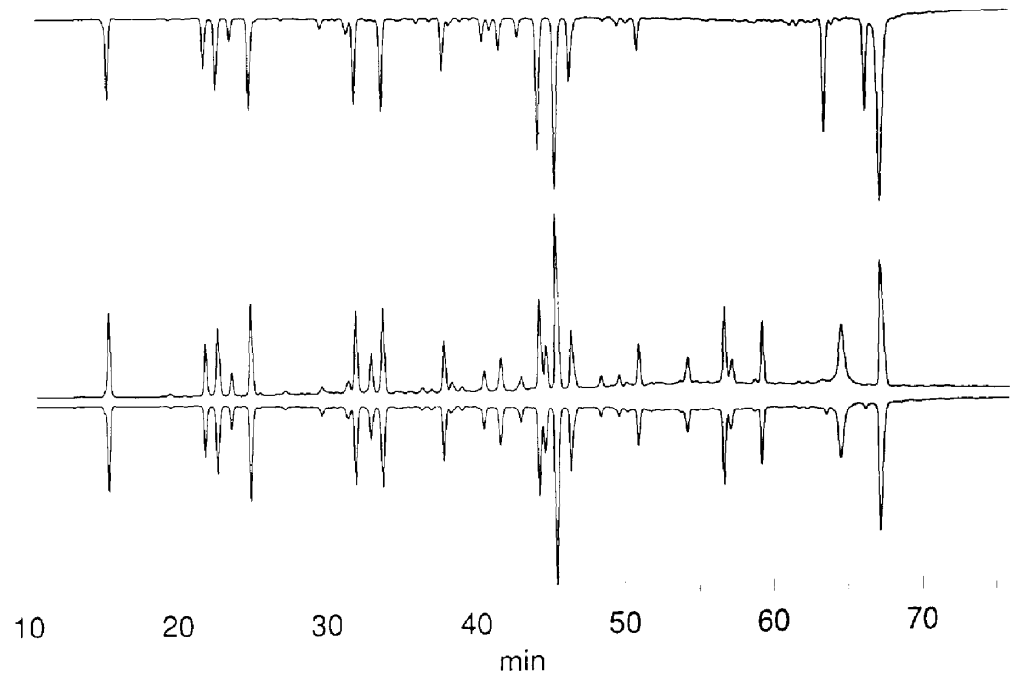


Fig. 6. Chromatograms of (top) a tryptic digest of bovine cytochrome *c*, (middle) equine cytochrome *c* and (bottom) a mixture in which 7% of bovine cytochrome *c* was added to the equine protein. All chromatograms are shown after the pretreatment to facilitate the comparison. Chromatographic conditions as in Experimental.

for classification. Each test sample should be digested and the fragments separated with the same mobile phase preparation as a new reference digest.

A system suitability test for the training set can be provided by a graphical evaluation procedure, where the *F*-values of the training set are plotted together with the test sample *F*-values. This is illustrated in Fig. 7, where some selected test samples are shown. The two critical *F*-values corresponding to outliers and rejected samples are indicated, thus allowing a graphical classification of the test samples. The graphical representation is used as a control chart where the *F*-value of the test sample is inserted together with the *F*-value of the corresponding reference chromatogram. The training set is valid as long as the new reference chromatograms are classified as acceptable. This is indicated by *F*-values for the new reference chromatograms below the critical *F*-value at the 0.10 significance level.

Visual inspection of the chromatograms in the training set reveals that the column had deteriorated throughout the study. This was also con-

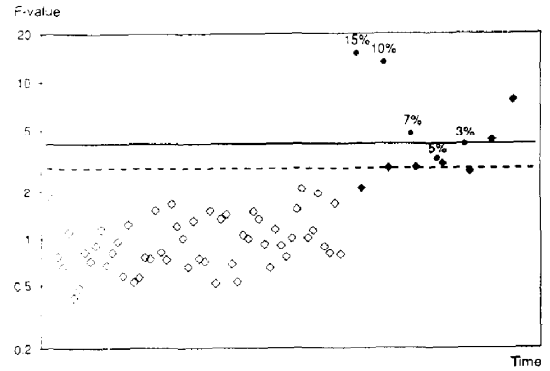


Fig. 7. System suitability test for the training set as a function of time. The solid line indicates the limit for rejection and the dashed line shows the limit for outliers. Note the logarithmic scale on the ordinate. ◇ = Training set objects; ◆ = reference chromatograms in the test set; ● = chromatograms with bovine cytochrome *c* added.

firmed by the characterization of the raw data made previously [20]. At the end of the study, the system suitability test indicated that the training set no longer is valid for classifications of test samples (see Fig. 7). One possible solution is the approach presented by Headley and Hardy [32]. They used a dynamic training set, where a set of new reference chromatograms was included in the training set after validation. The oldest chromatograms in the training set were simultaneously removed in order to keep the size of the set constant. A new class model has to be calculated by PCA each time the content of the training set is modified. The training set could be used for a prolonged period of time in their case, despite the observed changes in the experimental conditions, e.g., column degradation.

The visual detection of impurities or modified fragments is, of course, easier if the retention times for these fragments are known. The proposed classification method, on the other hand, treats the entire chromatographic profile and will detect additional peaks without prior knowledge of their position. An indication of the position of the deviant peaks can be obtained by inspection of the residual vector for the suspect chromatogram. Large residuals are expected for regions that deviate from the normal peak pattern in the training set.

General detection limits in peptide mapping are difficult to express, as the peak size necessary for detection is dependent on the elution position of the modified fragment. The general experience of Chloupek et al. [9] is that new, well resolved peaks must be greater than 5% (mol/mol) and co-eluting peaks greater than 15% for visual detection. Instrumental detection of 8% of a spiked contaminant eluted with baseline resolution has been reported [35]. Dougherty et al. [16] used an extensive characterization of the variability in the amount of individual fragments to achieve detection of spiked impurities in rDNA-derived somatotropin in 2–4.5% levels for some specific fragments [16]. The purpose of the method presented in this paper is not primarily to decrease the detection limit, but to facilitate the interpretation of the peptide maps. The automatic processing of numerous chromatograms in a data set, and the unbiased evaluation, are the main benefits.

## 5. Conclusions

The proposed multivariate evaluation method for fingerprinting techniques is one approach towards the full exploitation of the information contained in complex chromatograms. Numerous chromatograms in a data set can be automatically processed. The method is based on the entire chromatographic profile, thus eliminating the determination of retention time and peak area for the large number of peaks commonly encountered in peptide mapping. The differences between the samples are highlighted by reduction of the chromatographic variations in the data. The use of a training set to describe the normal variations in the cleavage pattern will improve the possibilities for detection of amino acid sequence modifications and contaminants in the protein sample. Multivariate classifications of protein samples according to their peptide maps are less subjective than the traditional visual inspection of two chromatograms.

This approach may also be useful in areas other than peptide mapping, as long as the profiles are relatively similar. In other cases where the profiles are very different, classification will be fairly straightforward from a direct visual comparison [7]. DNA fingerprinting [36] and pyrolysis–GC [37] are examples of other fingerprinting techniques that could be facilitated by the proposed method.

## Acknowledgements

I am grateful to my colleague Rolf Danielsson for the collaboration regarding the pretreatment procedure. Niklas Lundell, now at Pharmacia Bioscience Center, Stockholm, Sweden, is sincerely thanked for constructive ideas during the entire project. Pharmacia Biotech (Uppsala, Sweden) is acknowledged for supplying the chromatographic equipment used throughout the project. Staffan Renlund at Pharmacia is

thanked for providing insight into the peptide mapping technique.

## References


[1] W.S. Hancock, *LC · GC Int.*, 5, No. 4 (1992) 30.
[2] R.L. Garnick, N.J. Solli and P.A. Papa, *Anal. Chem.*, 60 (1988) 2546.
[3] F.E. Regnier, *LC · GC*, 5 (1987) 392.
[4] M.W. Dong and A.D. Tran, *J. Chromatogr.*, 499 (1990) 125.
[5] L.R. Snyder, in C. Horváth (Editor), *High Performance Liquid Chromatography: Advances and Perspectives*, Vol. 1, Academic Press, New York, 1980, p. 208.
[6] G.A. Ross, P. Lorkin and D. Perret, *J. Chromatogr.*, 636 (1993) 69.
[7] H. Rehbein, *Electrophoresis*, 13 (1992) 805.
[8] M.W. Dong, *Adv. Chromatogr.*, 32 (1992) 21.
[9] R.C. Chloupek, J.E. Battersby and W.S. Hancock, in C.T. Mant and R.S. Hodges (Editors), *HPLC of Peptides and Proteins: Separation, Analysis, and Conformation*, CRC Press, Boca Raton, FL, 1991, p. 825.
[10] S. Borman, *Anal. Chem.*, 59 (1987) 969A.
[11] K.L. Stone and K.R. Williams, in D.H. Schlesinger (Editor), *Macromolecular Sequencing and Synthesis*, Alan R. Liss, New York, 1988, Ch. 2, p. 7.
[12] R.C. Chloupek, R.J. Harris, C.K. Leonard, R.G. Keck, R.G. Keyt, M.W. Spellman, A.J.S. Jones and W.S. Hancock, *J. Chromatogr.*, 463 (1989) 375.
[13] N. Lundell, *J. Chromatogr.*, 639 (1993) 97.
[14] M. Herold, D.N. Heiger and R. Grimm, *Am Lab.*, August (1993) 20J.
[15] E.R. Hoff, *LC · GC*, 2, *No.* 6, (1989) 28.
[16] J.J. Dougherty, Jr., L.M. Snyder, R.L. Sinclair and R.H. Robins, *Anal. Biochem.*, 190 (1990) 7.
[17] R.C. Chloupek, W.S. Hancock and L.R. Snyder, *J. Chromatogr.*, 594 (1992) 65.
[18] R.G. Brereton (Editor), *Multivariate Pattern Recognition in Chemometrics*, Elsevier, Amsterdam, 1992.
[19] S. Wold, C. Albano, W.J. Dunn, III, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Londberg and M. Sjöström, in B.R. Kowalski (Editor), *Chemometrics: Mathematics and Statistics in Chemistry*, Reidel, Dordrecht, 1984, p. 17.
[20] G. Malmquist and R. Danielsson, *J. Chromatogr.*, 687 (1994) 71.
[21] S. Wold, K. Esbensen and P. Geladi, *Chemometr. Intell. Lab. Syst.*, 2 (1987) 37.
[22] S. Wold, C. Albano, W.J. Dunn, III, K. Esbensen, S. Hellberg, E. Johansson and M. Sjöström, in H. Martens and H. Russwurm, Jr. (Editors), *Food Research and Data Analysis*, Applied Science, London, 1983, p. 147.
[23] M.A. Sharaf, D.L. Illman and B.R. Kowalski, *Chemometrics*, Wiley, New York, 1986, Ch. 6, p. 179.
[24] P.J. Gemperline, L.D. Webber and F.O. Cox, *Anal. Chem.*, 61 (1989) 138.
[25] S. Renlund, I.-M. Klintrot, M. Nunn, J.L. Schrimsher, C. Wernstedt and U. Hellman, *J. Chromatogr.*, 512 (1990) 325.
[26] S. Renlund, personal communication, 1992.
[27] S. Wold, *Technometrics*, 20 (1978) 397.
[28] J.C. Miller and J.N. Miller, *Statistics for Analytical Chemistry*, Ellis Horwood, Chichester, 3rd ed., 1993, Ch. 3, p. 75.
[29] M.A. Sharaf, B.R. Kowalski and B. Weinstein, *Z. Naturforsch. Teil C*, 35 (1980) 508.
[30] *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, Georgetown University Medical Center, Georgetown, Washington, DC, 1972.
[31] B.K. Lavine, *Chemometr. Intell. Lab. Syst.*, 15 (1992) 219.
[32] L.M. Headley and J.K. Hardy, *J. Food Sci.*, 57 (1992) 980.
[33] B. Bourguignon and D.L. Massart, *Anal. Chim. Acta*, 282 (1993) 33.
[34] H.-J. Wirth, K.-O. Eriksson, P. Holt, M.I. Aguilar and M.T.W. Hearn, *J. Chromatogr.*, 646 (1993) 129.
[35] D.G. Bursryn, T. Copmann, M. Dinowitz, R. Garnick, A. Losikoff, A. Lubiniecki, M.S. Rubino and M. Wiebe, *Biopharm*, 4 (1991) 22.
[36] J.S.C. Smith and O.S. Smith, *Adv. Agron.*, 47 (1992) 85.
[37] J.A. Pino, J.E. McMurry, P.C. Jurs, B.K. Lavine and A.M. Harper, *Anal. Chem.*, 57 (1985) 295.