JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

Processing of Chromatographic Data for Chemometric Analysis of Peptide Profiles from Cheese Extracts: A Novel Approach

PAOLO PIRAINO,*,† EUGENIO PARENTE,† AND PAUL L. H. MCSWEENEY‡

Dipartimento di Biologia D.B.A.F., Università Basilicata, 85100 Potenza, Italy, and Department of Food and Nutritional Sciences, University College, Cork, Ireland

Chemometric analysis of chromatograms plays a fundamental role in characterization of foods or in detection of adulteration. Data for multivariate analysis of chromatographic profiles are usually obtained by visual matching (VM) of peaks, the identities of which, as for peptide profiles from cheese extracts, are often unknown. To avoid the main disadvantages of VM, which is subjective and time-consuming, a novel approach was developed. Fuzzy logic was employed to handle in a systematic way uncertainty in the position of peptide peaks, and chromatograms were processed by a rule-based membership function. Processed data consisted of classes of retention time wherein peak heights were accumulated by using the distance from the center of the class as a weight. The novel approach (fuzzy approach, FA) was compared with VM by using a real data set and by performing multivariate descriptive statistical techniques (principal component analysis, multidimensional scaling, and nonhierarchical cluster analysis). FA provided a fast, reliable, and objective alternative to VM and could be successfully applied for chemometric analysis of chromatographic profiles whenever knowledge of the identity of peaks is lacking or unnecessary.

KEYWORDS: Cheese; proteolysis; peptide profiles; fuzzy sets; chemometric analysis; PCA; MDS

INTRODUCTION

Separation techniques [high-performance liquid chromatography (HPLC), sodium dodecyl sulfate−polyacrylamide gel electrophoresis (SDS-PAGE), urea-PAGE, etc.) and chemometric analysis of digitalized profiles (chromatograms, electrophoretograms) are particularly suitable for the characterization of foods from a molecular point of view without the necessity of identifying all compounds detected (*1*). This is often the case for the characterization of cheese and, in particular, of proteolysis, one of the most important and complex events occurring during ripening. Due to its complexity, proteolysis in cheese is described by analyzing cheese extracts with separation techniques (*2*), some of which have been improved or developed for routine analysis (*3*). Using these techniques, many authors have supplied information on patterns of proteolysis in a large number of cheese varieties (*4*). Peptide profile by HPLC and SDS-PAGE are promising methods for the determination of the geographic origin, as for Emmental cheese (*5*), but their main disadvantage is that they are time-consuming.

Chromatography of cheese extracts, by separating breakdown products from caseins by their molecular mass, charge, size, hydrophobicity, etc., gives proteolytic profiles that contain information on the proteolytic process occurring in the cheese during ripening. Chemometrics may be defined as "how to get chemically relevant information out of measured data, how to represent and display this information, and how to get such information into data" (*6*). Chemometric treatment of analytical data plays a fundamental role in the characterization of foods or in the detection of adulteration. For instance, the information contained in peptide profiles of cheese extracts can be used to understand the biochemical pathways involved in proteolysis or simply to discriminate between cheeses of different varieties or between cheeses of the same variety made by different treatments.

Prior to chemometric analysis, chromatographic data (raw signal) are usually transformed and reduced to extract information. Signal treatment is designed to transform raw data in such a way that the results are more suitable for a specific application than the original signal. The purpose of data reduction, in turn, is the replacement of a large number of measurements by a few characteristic data in which all relevant information has been preserved (*7*).

If the identities of peaks in chromatograms are known, data processing of chromatograms is simplified by considering the peak itself as a variable, which can be used directly for chemometric analysis. For instance, gas chromatography−olfactometry data were used directly to perform principal component analysis (PCA) in chemometric analysis of Ragusano cheese flavor (*8*). When the identities of peaks are unknown, as in fingerprinting techniques, chromatographic data have to be processed to obtain variables, and this step can be time-consuming or can represent a source of error. The literature on

* Author to whom correspondence should be addressed (telephone +39-0971-205561; fax +39-0971-205503; e-mail piraino@unibas.it).
† Università della Basilicata.
‡ University College, Cork.

10.1021/jf049606n CCC: $27.50 © 2004 American Chemical Society
Published on Web 10/23/2004

processing of chromatographic data is extensive and covers different fields of research. In the field of cheese proteolysis, chemometric analysis of proteolytic profiles obtained by electrophoresis and chromatography has been reviewed (*9*). According to Pripp et al. (*9*), proteolytic profiles, in the form of reverse phase (RP)-HPLC chromatograms of cheese extracts, can be transformed into a multivariate data set by using peak height or peak area as variables and each chromatogram (sample) as an object. The same authors discussed mainly visual matching (VM) to obtain variables, but underscored the necessity for a more objective and efficient method of obtaining data. In VM, peaks are labeled manually and are visually matched among different proteolytic profiles by judging peaks as equivalent on the basis of their retention times. Peak or band identification by VM is still the most common approach used to obtain variables from chromatograms or electrophoretograms of cheese extracts (*10*−*13*). In addition, other approaches have been used, such as dividing chromatograms in sections and integrating each section (*14*).

Recently, a new procedure was presented by Piraino et al. (*15*) for processing and data reduction of electrophoretic profiles. Complex SDS-PAGE patterns of whole-cell proteins of lactic acid bacteria were transformed using a logistic weighting function into classes based on molecular weight wherein band intensities were accumulated. The procedure performed better than commercial software in clustering the patterns.

The goal of this work was to propose a novel data processing approach for chemometric analysis of chromatographic profiles, to avoid the main disadvantages of visual matching in obtaining variables. The approach was based on fuzzy logic and was developed to be objective, fast, and able to preserve all relevant information in a small set of variables. The FA was tested and compared with VM by using a data set as case study and by performing multivariate descriptive statistical techniques [PCA; multidimensional scaling (MDS), nonhierarchical cluster analysis (CA)] of proteolytic profiles.

MATERIALS AND METHODS

Data Set. The data set consisted of 24 RP-HPLC profiles from the ethanol-soluble fraction of Tilsit cheese, which were previously obtained by other authors (*16*) and used in this work as case study with permission from the authors. Objects in the data set were cheeses made with defined-strain surface starter mixes (labeled C or D) and reference cheeses made with a traditional old−young smear (R) at 1, 2, 4, and 8 weeks of ripening from both cheese core and surface.

Data Preprocessing and Data Reduction. For each sample, the chromatographic profile consisted of pairs of elution times and trace (absorbance at 214 nm detected at intervals of 1.5 s) values. Each profile was obtained over 70 min, and a matrix of $\sim 2.8 \times 10^3$ data was collected. Each profile was reduced first by integration and noise calculation. This step was performed by using the software (Varian Star Workstation 5) interfaced with the HPLC system (Varian Associates Inc., Walnut Creek, CA), in which all parameters (e.g., blank baseline subtraction, signal/noise ratio, peak width, tangent height %, etc.) were customized to transform signal into chromatograms composed of peaks height and retention times, so obtaining proteolytic profiles. Data reduction was of ∼1 order of magnitude. Data from each sample (chromatogram), which consisted of $m \times 2$ matrices with m retention time and m peak heights, were downloaded from the HPLC system (software) to be processed further.

Data Processing by Visual Matching. Peaks were identified by their retention time, labeled by peak number, and visually recognized in all chromatograms of the data set. A total of 125 peaks were visually matched by a single operator, and they were used as variables for multivariate statistical analysis. Peak height was zeroed for unmatched peaks. The final multivariate data set consisted of a 125 (peaks) × 24 (samples) matrix.

Logistic Function. Variables were transformed in classes (c) by a logistic weighting function defined by the formula

$$C_c = \sum_{j=1}^{n} \mathrm{ph}_j \times \frac{1}{1 = \mathrm{e}^{a(|\mathrm{rt}_j - \mathrm{rt}_c|(w_c/2))}} \quad (1)$$

where C is the value for variable c (class, or interval of time defined over the elution time axis); $\mathrm{ph_j}$ is peak height for peak j in the chromatogram; n is the number of peaks in the chromatogram; a is a shape parameter of the function; rt_c and rt_j are retention time for class c and for peak j, respectively; and w_c is class width. The second term of eq 1 is the weight with which ph is attributed to class c and is expressed as percentage. To assist the choice of a, two further parameters were defined: flat range (FR) and membership in the flat range (MFR). MFR was the minimum weight (in percent) for ph when peak position was within a specific distance d from the class center ($d = |\mathrm{rt}_c - \mathrm{rt}_x|$). FR was defined as the range around the class center given by $\mathrm{rt}_c \pm d$ (expressed as percent of w_c). Thus, a is a function of MRF and FR:

$$a = \ln\left(\frac{100}{\mathrm{MFR}} - 1\right) \times \frac{2}{w_c\left(\frac{\mathrm{FR}}{100} - 1\right)} \quad (2)$$

Calculations were carried out using a datasheet (Microsoft Excel format) with embedded macros. The file can be downloaded at http://www.unibas.it/utenti/parente/fuzzy.html. The final multivariate data set obtained by using the logistic function consisted of a c (classes) × 24 (samples) matrix.

Settings and Data Subsets. The following parameters were set: retention time of the first (rt_1) and of the last (rt_n) class (the range of elution time in the chromatogram to be processed); the number of intervals (I) over the profile, which given the number of classes c is $c = I + 1$; MFR; FR. Class width (w_c) resulted from the formula

$$w_c = \frac{\mathrm{rt}_n - \mathrm{rt}_1}{I} \quad (3)$$

Considering that the peptide profiles were obtained by a chromatographic run of 73 min and that the injection peak appeared after 2 min, the first and last classes, rt_1 and rt_n, were set at 3 and 73 min, respectively. To set the maximum number of intervals, the variance associated with peak position was assessed by computing statistics for shifts in retention time. The shift corresponded to the difference between the maximum and minimum retention times of visually matched peaks that were present in all chromatograms of the data set and were judged to be equivalent by the operator who performed visual matching. The number of intervals over the profile was set at 35, 70, or 100, so obtaining three subsets of data with numbers of classes $c = 36$ (LOW), $c = 71$ (MED), and $c = 101$ (HI) with class widths (w_c) of 2, 1, and 0.7 min, respectively. MFR was 95% in all cases; FR was 50% of class width for subsets LOW and MED or 75% of class width for subset HI.

Statistical Analysis. All chromatograms were scaled in percent (by dividing peak height by the sum of heights for each profile) before data processing. The information from chromatographic data, which were processed by either VM or FA, was extracted by three techniques: PCA of the covariance matrix; MDS (the Kruskal loss function was used for scaling) of the similarity matrix of Pearson product-moment correlation coefficients; CA with K-means clustering using Pearson correlations (clusters were represented as convex hulls of the samples that were members of a cluster on the MDS plots). Calculations and graphics were carried out by using Systat 10 for Windows (SPSS, Chicago, IL).

RESULTS

Logistic Function. The logistic weighting function transforms chromatograms consisting of two vectors of data (retention time and peak height) with varying lengths (i.e., different numbers

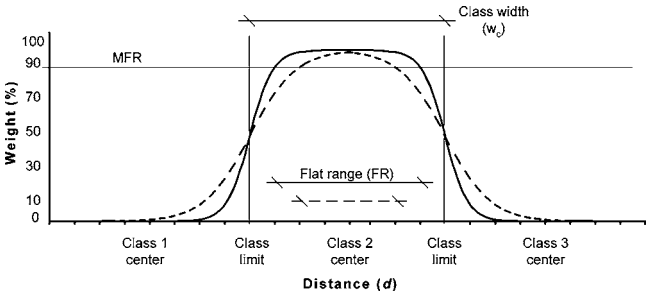

Figure 1. Weighting term of the logistic function. Class 2 weight distribution (in percent) as a function of the distance (*d*) from class center, with MFR = 90% (membership in the flat range) and FR = 50 (- - -) or FR = 75 (—). FR (flat range) is expressed as percent of w_c (class width).

of peaks in different profiles) in a single vector of fixed length (the number of classes, *c*, is constant for all profiles). Peak heights are accumulated in classes using a logistic weighting function, which provides the fuzzy membership of each peak for each class. When $rt_c = rt_j$, the retention time of the peak matches the retention time of the class, the weight is 1 (eq 1), and peak height is completely assigned to class *c*. When $rt_j = rt_c \pm (w_c/2)$, the retention time of peak *j* matches the class limit, and the weight is 0.5 (eq 1), so peak height is equally divided between the two neighboring classes. For other cases, the weight depends on the *a* parameter and the distance of the peak from the class center. The shape of the logistic weighting function used in the fuzzy approach is shown in **Figure 1**. The parameter *a* regulates the fuzzy feature of the output. With increasing values of *a*, the function becomes steeper and a more defined membership is obtained; that is, the weight with which a given peak is attributed to the nearest class increases. The parameter, *a*, can be calculated as a function of FR, MFR, and class width (w_c), and these parameters can in turn be adjusted to keep into account uncertainty in peak position.

The effect of data processing and reduction using the fuzzy approach on the matrix of data for a chromatogram (as downloaded from the HPLC system, with *m* retention times and *m* peak heights) is shown in **Figure 2** for a peptide profile randomly taken as an example. Variables (peaks, shown in the preprocessed data chart) were weighted and accumulated in 36, 71, and 101 classes as shown in **Figure 2** for subsets LOW, MED, and HI, respectively. In subset HI, class width was narrow (0.7 min), number of classes was high, most peaks contributed individually to the height of the corresponding class, and only a small number of peaks were either partitioned in two classes or summed in a single class. As a result, preprocessed data and subset HI (**Figure 2a**) were rather similar. Using a lower number of classes (subsets MED and LOW), class width was larger (1 and 2 min, respectively) and more often groups of peaks were weighted and summed into the same class. In the last case, few variables (classes) had a zero value, and class heights were proportionally higher than peak heights. In each case, because chromatograms were scaled in percent before data processing, the sums of heights in all processed peptide profiles were equal.

Settings and Data Subsets. The variance associated with peak position was computed using retention time of visually matched peaks that were present in all chromatograms of the data set and were judged to be equivalent by the operator who performed visual matching. A total of 32 peaks of 124 met these conditions. Shifts in retention time (the difference between the maximum retention time and minimum retention time for each peak) and associated statistics are shown in **Table 1**. The maximum and average shifts in retention time were 6.67 and

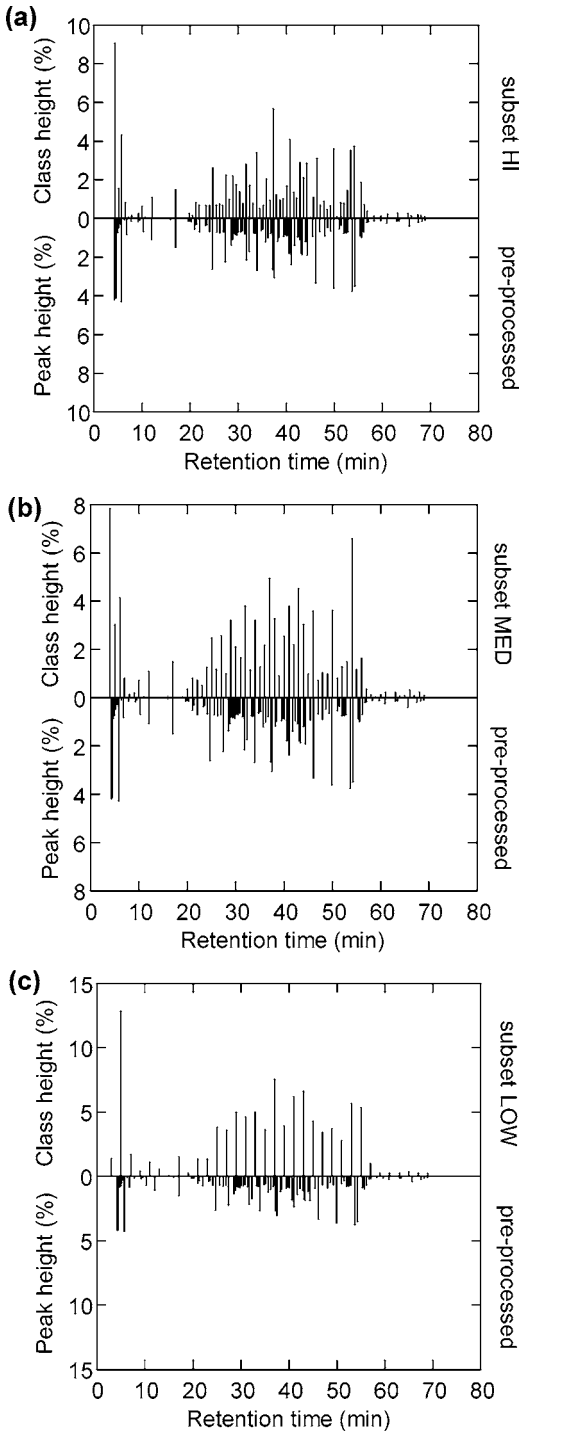

Figure 2. Comparison of a proteolytic profile as downloaded from HPLC system with the corresponding processed profile obtained by fuzzy approach for subsets HI (**a**), MED (**b**), and LOW (**c**). Variables were peak height and retention time in the preprocessed chromatogram or class height and class center in the processed profile; numbers of classes were 36, 71, and 101 for subsets LOW, MED, and HI, respectively.

1.22 min, respectively. When the normal probability plots of the retention times of matched peaks were analyzed, one or more outliers were identified (the 6.67 min shift, for example, was an outlier). This can be explained by the subjectivity of the VM. In fact, because the instrumental error should result in a normal distribution of peak positions, the presence of outliers may be a consequence of incorrect identification of peaks (or problems associated with the HPLC analysis). To exclude outliers and to

Table 1. Basic Statistics on Retention Time of Peaks Present in All Chromatograms of the Data Set and Judged Equivalent by Visual Matching (Matched Peaks)

no. of chromatograms	24
no. of matched peaks	32
max shift[a]	6.67
min shift[a]	0.22
av shift[a]	1.22
median shift[a]	0.99
max shift at 99% of confidence[a]	1.88
av shift at 99% of confidence[a]	0.28

[a] Shifts were calculated as the difference between the maximum retention time and minimum retention time for each peak and are expressed in minutes.

obtain an objective estimate of the range of the position of each peak, the difference between upper and lower 99% confidence limits was calculated for each peak. The maximum and average values of these differences should provide a better approximation for the variability in peak position. On the basis of these results, the maximum number of classes was set. For subset HI, which had the highest number of classes (101), w_c was 0.7 min, and FR was 75% of w_c. With these settings, all peaks in the range of 0.53 min around the class center were treated as equivalent by the logistic function (i.e., were weighted with a minimum weight of 95% in the flat range), so the same peak present in different chromatograms was always attributed with equal weight to the same class independently of its shift in retention time. In this way, the fuzzy approach grouped peaks into classes by applying a rule comparable to the VM. Peaks outside the flat range were weighted with different values according to the shape of the function.

Statistical Analysis. *PCA*. Results from PCA (first two components only) of chromatographic data from Tilsit cheese processed either by visual matching or by fuzzy approach are shown in **Figure 3**. The number of components was chosen on the basis of a predefined amount of variance to be explained (70%) and of the scree plots from each subset. For the purpose of this study the VM plot (**Figure 3a**) was considered to be the reference plot. Core and surface samples can be distinguished clearly. No separation occurred between samples from the core over the first two components. Within samples from the surface, cheeses made with reference smear (R) formed a single and separate group without differences between ripening times. Cheeses made with defined-strain surface starter (mix C or D) were separated over PC2 by ripening time and to a lesser extent by starter mix used in cheesemaking. The loading vectors showed that three main peaks (peptides or group of peptides eluted at 4.2, 5.7, and 31.9 min, respectively) explained most of the variance. Peaks in the range from 31.9 to 39.6 and hydrophilic peptides eluting at 4.3 and 5.3 min were correlated and were mainly associated with the first component. The second component was associated mainly with peaks eluting at 4.2 and 5.7 min and with groups of peaks in the ranges of 54.7−56.2 and 44.5−49.9.

Peptide profiles were transformed into classes by the fuzzy approach, and classes were used as variables for chemometric analysis. Results for subset HI (101 classes) are shown **Figure 3b**. Scores were comparable with those obtained using VM (**Figure 3a**), but some differences were evident. Core samples had a wider spread in subset HI (**Figure 3b**) compared to VM. In addition, samples from cheese surface at 1 or 2 weeks of ripening and samples of 4 or 8 weeks made with reference smear (R) were separated, whereas they were tightly grouped when VM was used. Surface samples of cheeses made with defined strain surface starters (mix C or D) were separated mainly over PC2 as for VM (**Figure 3a**). For the MED data subset (**Figure 3c**; 71 classes) the score plot obtained was very similar to the score plot obtained with subset HI (**Figure 3b**), and a slight increase of variance explained (76.3 versus 71.3%) was obtained. The score plot obtained from the subset LOW is shown in **Figure 3d**. Core and surface samples were clearly separated on PC1, as in **Figure 3a**−**c**, whereas positions of samples from the surface were comparable to those of VM, except for the position of group R (cheeses made with reference smear).

PCA is especially suitable for analysis of peptide profile if loadings are taken into account to provide information on the peaks responsible for the grouping. PCA transforms variables into principal components by computing linear combinations of the original variables. The importance of each original variable is expressed by the loadings. Loadings are shown as vectors in **Figure 3** and were analyzed to understand differences between score plots. Peaks with high loadings eluted at 4.2, 5.7, and 31.9 min, which explained most of the variance in VM, were also present in **Figure 3b** (HI, classes at 4.4, 5.8, and 31.7 min, respectively) and **Figure 3c** (MED, classes at 4, 6, and 32 min, respectively), although classes at 31 and 5 min also had high loadings in MED. In **Figure 3d** (LOW), the main variables that explained the variance in the data set were classes at 5, 31, and 55 min. In general, variables obtained by the fuzzy approach were classes including peaks that had high loadings in the VM approach. This is emphasized in **Figure 4**, in which the position on the retention time axis of the variables which had high loadings (higher than 0.5 in absolute value) is compared.

When a high number of classes was used, most peaks contributed individually to the height of the corresponding class. For a lower number of classes, heights of several peaks may be summed in the same class. This feature of fuzzy variables affected the differences observed in scores. Moreover, **Figure 4** highlights the presence of variables (classes at 28 or 42 min) that were not present in VM. This suggested that VM might hide differences due to small peaks (with low loadings), whereas the fuzzy approach may reveal these differences including small (or bordering) peaks in a class with high loadings.

MDS. **Figure 5** shows the bidimensional plots obtained by MDS of Pearson matrices that were computed on data processed by VM and by FA. Bidimensional plots obtained by using the Pearson matrices were more similar to the PCA score plots than bidimensional plots obtained by using Euclidean distance or covariance matrices (plots not shown). When the bidimensional plots of **Figure 5** and the score plots of **Figure 3** were compared, it was evident that the two multivariate statistical techniques provided equivalent results. Relationships between peptide profiles in the MDS space were identical to those in the PCA space, but the two spaces were oriented differently. The proportion of variance explained by MDS was higher (98.4, 97.1, 97.6, and 97.4%, respectively, for VM, HI, MED, and LOW subsets of data) than that explained by PCA.

CA. Nonhierarchical cluster analysis was performed on each data set to assess the effect of the fuzzy approach on clustering data and to evaluate if the VM and the FA approaches provide a similar grouping. The number of clusters used for *K*-means was obtained by two different criteria: (1) a hierarchical cluster analysis was carried out and the number of well-separated clusters that were evident in the dendrogram was used in *K*-means; 2) *K*-means was performed with different numbers of clusters (2, 3, 4, 5, 6), and a scree plot of a number of groups (*x*-axis) and within-group sum of squares (*y*-axis) was produced, so the number of groups used in the final analysis was that corresponding to the elbow of the graph. Samples within data

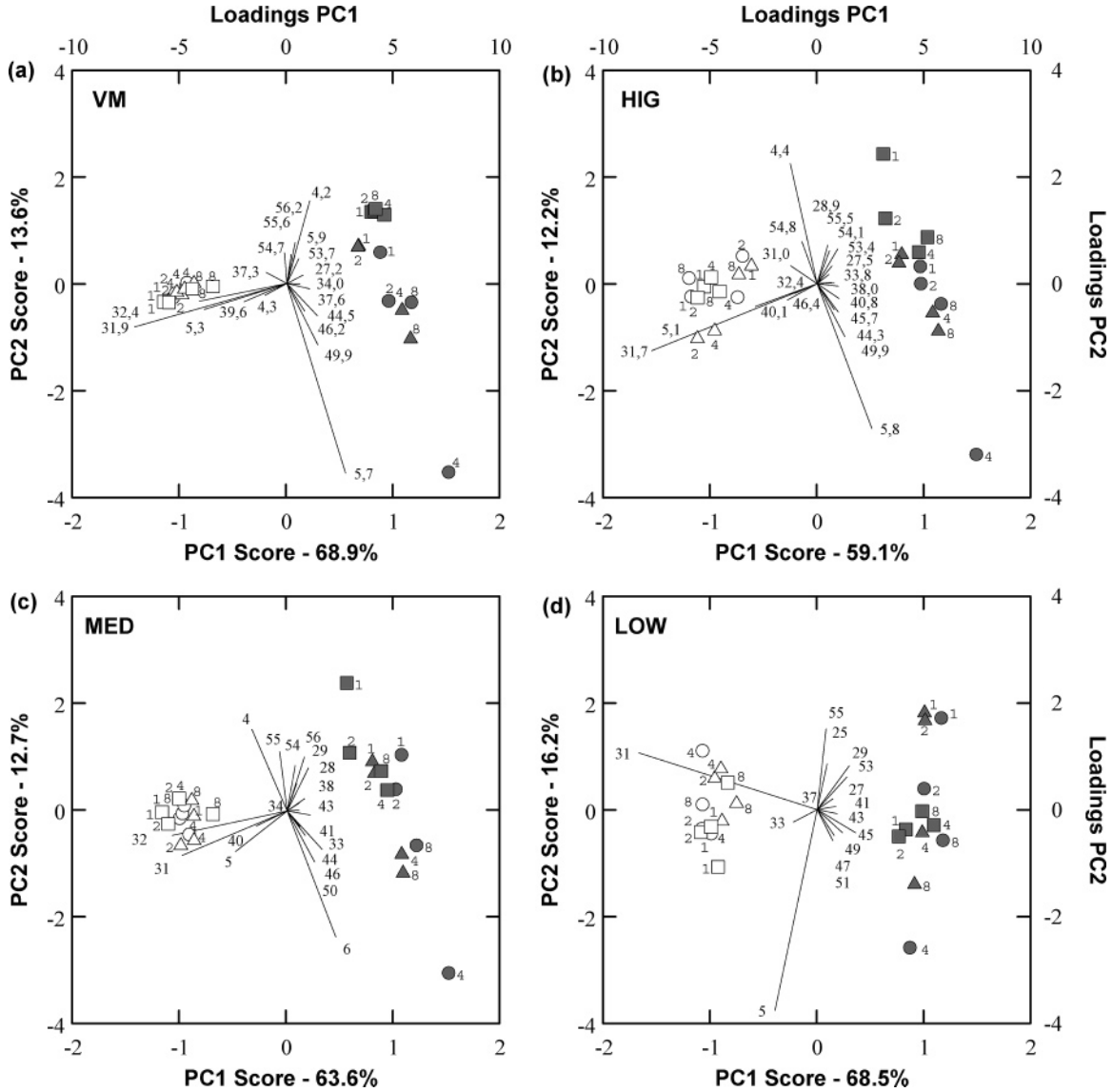

Figure 3. PC1–PC2 score and loading plot vectors of variables obtained by PCA of peptide profiles processed by visual matching (VM, **a**) and by fuzzy approach (HI, **b**; MED, **c**; and LOW, **d**). Variables with loadings outside the range ± 0.5 are shown and labeled by the average retention time of peaks (**a**) or by retention time of class center (**b**–**d**). Score symbols refer to reference smear R (■) and to defined-strain smear mix D (▲) or mix C (●); open and solid symbols are for core and surface samples, respectively. Score numbers refer to ripening time (weeks).

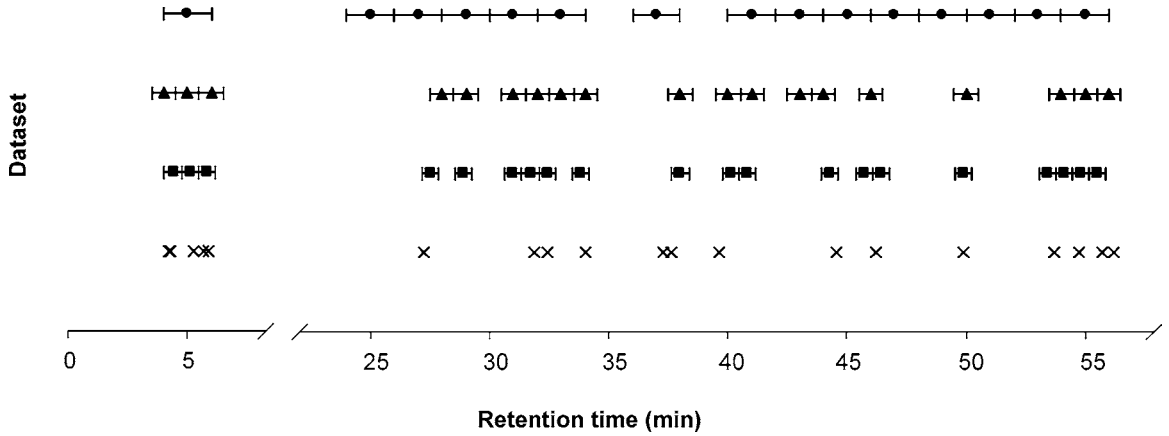

Figure 4. Retention time plot of variables with loadings outside the range ± 0.5 that explained the variance in PCA of peptide profiles processed by visual matching (×) or by fuzzy approach for subsets HI (■), MED (▲), and LOW (●). Bars for variables obtained by fuzzy approach represent class width for each subset of data.

sets were spilt into five groups (clusters 1−5). Groups were represented as convex hulls in the MDS graphs of **Figure 5**. Grouping was similar between data processed by visual matching and data processed by fuzzy approach both at high and

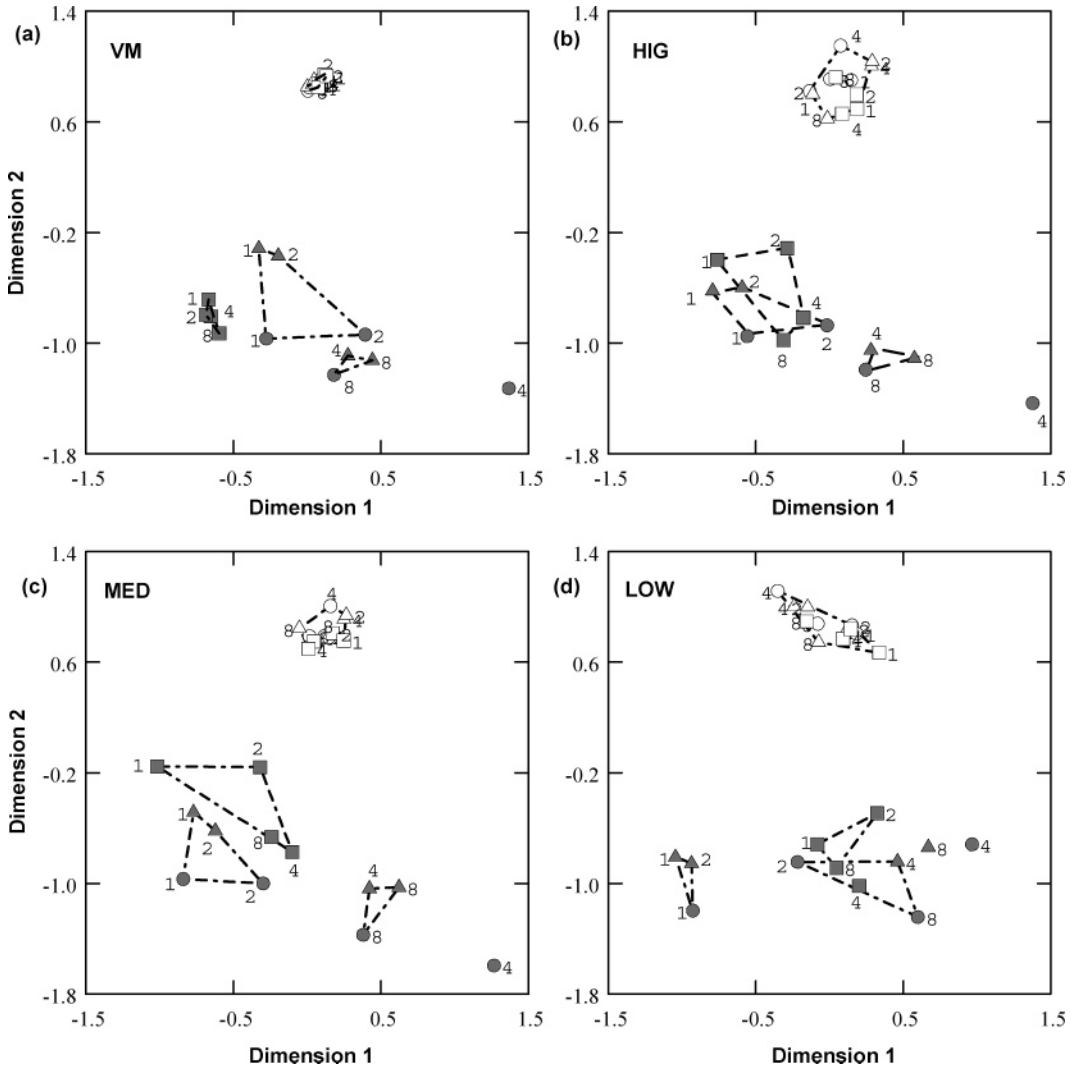

Figure 5. MDS configuration of the matrix of Pearson product-moment correlation between peptide profiles processed by visual matching (VM, **a**) or by fuzzy approach (HI, **b**; MED, **c**; and LOW, **d**). Symbols refer to reference smear R (■) and to defined-strain smear mix D (▲) or mix C (●); open and solid symbols are for core and surface samples, respectively. Convex hulls around samples show the five groups obtained by *K*-means clustering.

medium numbers of classes (subsets HI and MED). Using a lower number of classes (subset LOW), some shifts occurred.

DISCUSSION

Why the Fuzzy Set Theory Should Be Used. Fuzzy set theory was proposed by Zadeh (*17*, *18*) to provide a mathematical tool for dealing with linguistic variables (i.e., concepts described in natural languages). More than a single theory, the author proposed the "fuzzification" process as a methodology to generalize any specific theory from a crisp (discrete) to a continuous (fuzzy) form, so handling the concept of partial truth—truth values between "completely true" and "completely false". In this way, propositions can be represented with degrees of truthfulness and falsehood, and formalized tools for dealing with the imprecision intrinsic to chemical analysis or data interpretation can be provided. Visual matching, the common approach used to obtain variables from proteolytic profile, is a procedure in which the researcher decides, on the basis of retention time, if peaks present in different chromatograms are equivalent or not. In practice, visual matching is performed by answering the questions "is the peak with retention time A in chromatogram X equivalent to the peak with retention time B in chromatogram Y?" and "are differences in retention time due to the instrumental error only?" This is a discrete situation of partial truth—truth values with concepts described in natural languages. Using a fuzzy approach it is possible to quantify linguistic inputs and to give quickly a working approximation of complex and often unknown system input−output rules. We used a fuzzy approach to transform proteolytic profiles in a continuous (fuzzy) subset of data, so that the new subset of data was suitable for statistical analysis. The transformation was made by a rule-based membership function (**Figure 1**) that established degrees of truthfulness in grouping peaks into classes. For more general information on fuzzy sets and systems see Dubois and Prade (*19*) and Kosko (*20*).

Fuzzy Approach. Retention time of peaks is the chemical information of RP-HPLC profiles of cheese extracts, which can express the hydrophobicity and/or molecular weight of the peptides associated with each peak (depending also on the column properties and gradient characteristics). With the fuzzy approach a predefined number of intervals in the retention time axis are fixed, and each interval maintains its chemical information. As shown in **Figure 2**, a processed profile appears to be very similar to the corresponding raw chromatogram at the high number of classes, so that the chemical information is not removed by the "fuzzification" process, but it is combined or summarized. However, peaks that are not well separated in the chromatogram become a single variable in the fuzzy profile,

and this might represent a disadvantage of the fuzzy approach in some cases. Results of **Figures 3** and **4** show that almost the same information was extracted from the data set using VM or FA: cheese samples were classified in a similar way (internal layer versus external layer, by ripening time, etc.), and the variables that explained the variance of the data set were equivalent in the two approaches.

Three main features distinguished the new approach (FA) from the standard method (VM). FA was much faster compared to VM (a set of chromatograms can be processed in a few minutes compared with several hours for VM). FA was more objective than VM (data processing by FA is automated and achieved by a data sheet, whereas the results of VM may be affected by the ability of an operator to match peaks, as indicated by the statistics for peaks position shown in **Table 1**). FA was a flexible tool (the number of classes can be chosen depending on the purpose of the analysis) with the potential to be used in data reduction contexts (e.g., to make inferential analysis easier by reducing greatly the number of dependent variables). In general, FA facilitated the extraction of information from RP-HPLC peptide profiles compared to VM.

Settings and Data Subsets. Even if the number of classes can be freely chosen by using the fuzzy approach, in practice, the number of classes ranges from a minimum dependent on the purpose of the analysis to a maximum dependent on the experimental error (or precision of the chromatographic analysis). The minimum number of classes depends on the purpose of the analysis because the relevant information might be lost if the data set is reduced to too few variables. On the other hand, the number of variables should be balanced with the number of samples, especially when the purpose of the study is to develop automated identification methods as in discriminant analysis or artificial neural networks (*21*). The maximum number of classes depends on the experimental error. In fact, in the absence of peak synchronization (retention time adjustment for peak position), class width must keep into account the variance associated with peak position. Therefore, because the number of classes is inversely proportional to class width, the maximum number of classes is given by the minimum width for a class enclosing the error of peak position.

Statistical Analysis. PCA is the most common statistical tool used in chemometric analysis of proteolytic profiles of cheese extracts (*11*−*14*, *22*) or peptide profiles of milk protein hydrolysates in general (*23*, *24*). As shown in **Figure 3**, score and loading plots obtained by PCA of data processed according to the new approach were almost the same as that obtained from data processed according to the standard method. Because PCA has some disadvantages [for example, PCA of covariance matrix is not robust and is affected by outliers, and the choice and exploration of principal components may be complicated or questionable (*21*)], multivariate statistical techniques that can be used as an alternative to PCA, such as MDS, were tested in this study to confirm the suitability of the new approach in obtaining variables for statistical analysis of RP-HPLC profiles. MDS is not commonly used in chemometric analysis of proteolytic profiles. MDS can often fit an appropriate model in fewer dimensions than can PCA, and if it is implausible to assume a linear relationship between dissimilarities, multidimensional scaling provides a simple dimensional model. On the other hand, although MDS maps are easy to interpret, the relationship between original variables and MDS coordinates is not always clear, especially when the number of variables is high, so it is not possible to give information on the variables responsible for the observed differences. However, as observed with PCA, MDS represented almost in an equivalent way data processed by FA and data processed by VM (**Figure 5**), thus confirming that the new approach could be used as an alternative to VM in obtaining variables. Results from cluster analysis showed that the fuzzy approach has no effect on clustering data (if a high number of classes are used), so that the groupings were similar between the VM and FA approaches. Clusters were represented as convex hulls on the MDS plot (**Figure 5**) only to summarize visually results from *K*-means, but the grouping obtained by *K*-means was not necessarily evident from the MDS graph (overlapping of the convex hulls of the clusters) because the two procedures evaluate distance relationships in different ways. In any case, the association of MDS with cluster analysis could be an alternative in representing and grouping cheese samples based on their proteolytic profiles, especially when data are processed using the fuzzy approach.

ABBREVIATIONS USED

VM, visual matching technique; FA, novel fuzzy approach; PCA, principal component analysis; MDS, multidimensional scaling; CA, cluster analysis; FR, flat range; MFR, membership in the flat range; HI, subset of data with a number of 101 classes; MED, subset of data with a number of 71 classes; LOW, subset of data with a number of 36 classes; HPLC, high-performance liquid chromatography; SDS-PAGE or urea-PAGE, polyacrylamide gel electrophoresis with sodium dodecyl sulfate or urea, respectively.

LITERATURE CITED

(1) Cordella, C.; Moussa, I.; Martel, A. C.; Sbirrazzuoli, N.; Lizzani-Cuvelier, L. Recent developments and adulteration detection: technique-oriented perspectives. *J. Agric. Food Chem.* **2002**, *50*, 1751−1764.

(2) McSweeney, P. L. H.; Fox, P. F. Chemical methods for the characterization of proteolysis in cheese during ripening. *Lait* **1997**, *77*, 41−76.

(3) Ardö, Y.; Polichroniadou, A. Analysis of peptides. In *Laboratory Manual for Chemical Analysis of Cheese*; Office for the Official Publications of the European Communities: Luxembourg, 1999; pp 31−64.

(4) Sousa, M. J.; Ardö, Y.; McSweeney, P. L. H. Advances in the study of proteolysis during cheese ripening. *Int. Dairy J.* **2001**, *11*, 327−345.

(5) Pillonel, L.; Tabacchi, R.; Bosset, J.-O. Analytical methods for the determination of the geographic origin of Emmental cheese. Summary of a screening study. *Mitt. Lebensm. Hyg.* **2003**, *94*, 60−69.

(6) Wold, S.; Sjöström, M. Chemometrics, present and future success. C*hemom. Intell. Lab.* **1998**, *44*, 3−14.

(7) Massart, D. L.; Vandeginste, B. G. M.; Deming, S. N.; Michotte, Y.; Kaufman, L. *Chemometrics: a Textbook*; Elsevier Science Publisher: Amsterdam, The Netherlands, 1988.

(8) Carpino, S.; Acree, T. E.; Barbano, D. M.; Licitra, G.; Siebert, K. J. Chemometric analysis of Ragusano cheese flavor. *J. Agric. Food Chem.* **2002**, *50*, 1143−1149.

(9) Pripp, A. H.; Stepaniak, L.; Sørhaug, T. Chemometrical analysis of proteolytic profiles during cheese ripening. *Int. Dairy J.* **2000**, *10*, 249−253.

(10) Dewettinck, K.; Dierckx, S.; Eichwalder, P.; Huyghebaert, A. Comparison of SDS-PAGE profiles of four Belgian cheeses by multivariate statistics. *Lait* **1997**, *77*, 77−89.

(11) Saldo, J.; McSweeney, P. L. H.; Sendra, E.; Kelly, A. L.; Guamis, B. Proteolysis in caprine milk cheese treated by high pressure to accelerate cheese ripening. *Int. Dairy J.* **2002**, *12*, 35−44.

(12) Poveda, J. M.; Sousa, M. J.; Cabezas, L.; McSweeney, P. L. H. Preliminary observations on proteolysis in Manchego cheese made with a defined-strain starter culture and adjunct starter (*Lactobacillus plantarum*) or a commercial starter. *Int. Dairy J*. **2003**, *13*, 169−178.
(13) Di Cagno, R.; De Angelis, M.; Upadhyay, V.; McSweeney, P. L. H.; Minervini, F.; Gallo, G.; Gobetti, M. Effect of proteinases of starter bacteria on the growth and proteolytic activity of *Lactobacillus plantarum* DPC2741. *Int. Dairy J*. **2003**, *13*, 145−157.
(14) Sørensen, J.; Benfeldt, C. Comparison of ripening characteristics of Danbo cheeses from two dairies. *Int. Dairy J*. **2001**, *11*, 355−362.
(15) Piraino, P.; Ricciardi, A.; La Norte, M. T.; Malkhazova, I.; Parente, E. A new procedure for data reduction in electrophoretic fingerprints of whole-cell proteins. *Biotechnol. Lett.* **2002**, *24*, 1477−1482.
(16) Hannon, J. A.; Lillevang, S.; Sepulchre, A.; Bockelmann, W.; McSweeney, P. L. H. Defined-strain surface starters in Tilsit cheese. *Aust. J. Dairy Technol.* **2002**, *57*, 73−75.
(17) Zadeh, L. A. Fuzzy sets. *Inform. Control* **1965**, *8*, 338−353.
(18) Zadeh, L. A. Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Trans. Syst. Man Cyb.* **1973**, *3*, 28−44.
(19) Dubois, D.; Prade, H. *Fuzzy Sets and Systems: Theory and Applications*; Academic Press: New York, 1980.
(20) Kosko, B. *Fuzzy Engineering*; Prentice-Hall: Upper Saddle River, NJ, 1997.
(21) Everitt, B. S.; Dunn, G. *Applied Multivariate Data Analysis*; Arnold: London, U.K., 2001.
(22) Pripp, A. H.; Shakeel-Ur-Rehman; McSweeney, P. L. H.; Fox, P. F. Multivariate statistical analysis of peptides profiles and free amino acids to evaluate effects of single-strain starters on proteolysis in miniature Cheddar-type cheeses. *Int. Dairy J.* **1999**, *9*, 473−479.
(23) Ven, C.; Gruppen, H.; Bont, D. B. A.; Voragen, A. G. J. Reversed phase and size exclusion chromatography of milk protein hydrolysates: relation between elution from reversed phase column and apparent molecular weight distribution. *Int. Dairy J.* **2001**, *11*, 83−92.
(24) Pripp, A. H.; Shakeel-Ur-Rehman; McSweeney, P. L. H.; Sørhaug, T.; Fox, P. F. Comparative study by multivariate statistical analysis of proteolysis in sodium caseinate solution under cheese-like conditions caused by strains of *Lactococcus*. *Int. Dairy J*. **2000**, *10*, 25−31.

Received for review March 10, 2004. Revised manuscript received August 5, 2004. Accepted September 10, 2004.

JF049606N