

# Coupled-Column RP-HPLC in Combination with Chemometrics for the Characterization and Classification of Wheat Varieties

Róbert Berky<sup>1</sup> · Enikő Sipkó<sup>1</sup> · Gábor Balázs<sup>2</sup> · Anna H. Harasztos<sup>2</sup> · Sándor Kemény<sup>3</sup> · Jenő Fekete<sup>1</sup>

Received: 2 December 2015 / Revised: 25 March 2016 / Accepted: 27 April 2016 / Published online: 18 May 2016  
© Springer-Verlag Berlin Heidelberg 2016

**Abstract** The application of a coupled-column reversed phase high-performance liquid chromatographic method is presented for assessing the quality of industrially relevant wheat flours. Three chromatographic columns filled with core-shell particles were connected in series to generate high resolving power. Peak heights of all 64 peaks in the resulting gliadin protein profiles were used as input data for the statistical comparison of ten wheat varieties. Univariate analysis of variances with nested design indicated that there were significant differences not only between the wheat varieties but also between the plots of lands they were cultivated on. The effect of the repeated sample preparations was found negligible. Principal component analysis and hierarchical clustering revealed that Bezostaja-1, Glenlea, MV-Karizma, MV-Magdaléna, MV-Mazurka and TF-Rétság had unique, while Bánkúti-1201, Fleischmann-481, Székács-1242 and TF-Komádi similar gliadin protein profiles. Nearest mean classification on the scores of the first three principal components could classify six wheat varieties out of ten.

**Keywords** RP-HPLC · PCA · Hierarchical clustering · Varietal classification · Gliadin

✉ Róbert Berky  
robert.berky@gmail.com

<sup>1</sup> Department of Inorganic and Analytical Chemistry, Budapest University of Technology and Economics, 1111, Szt. Gellért tér 4., Budapest, Hungary

<sup>2</sup> Department of Applied Biotechnology and Food Science, Budapest University of Technology and Economics, 1111, Szt. Gellért tér 4., Budapest, Hungary

<sup>3</sup> Department of Chemical and Environmental Process Engineering, Budapest University of Technology and Economics, 1111, Budafoki út 8., Budapest, Hungary

## Introduction

Cereal grains, which contain wheat gluten, are not only sources of proteins for humans, but also they are basic components of animal feeds too. Reliable characterization of different wheat varieties is very important from the industrial point of view for testing and maintaining wheat quality. The protein composition of the gluten content of flour strongly influences the rheological properties of dough during processing. Because of this, flours made from different wheat varieties can be used for different industrial purposes. The technological properties of wheat flour are strongly dependent on genetic background, environmental effects and their interactions [1].

Gluten, which is also named wheat prolamins due to its high amino acid (proline, glutamine) content, is a rubbery mass that remains after wheat dough is washed to remove starch granules and water-soluble constituents [2]. These proteins are the major components of wheat storage proteins. Prolamin proteins have been divided into two fractions according to their solubility in alcohol-water solutions (e.g., 70 % ethanol), the soluble gliadins and the insoluble glutenins. Gliadins are monomeric proteins that form only intramolecular disulphide bonds. They can be divided into four groups on the basis of mobility at low pH in gel electrophoresis:  $\alpha$ ,  $\beta$ ,  $\gamma$ , and  $\omega$  gliadins (in the molecular mass range of 28–55 kDa). Glutenins are polymeric proteins, whose subunits are held together by intermolecular disulphide bonds. These polymeric proteins can be divided into two groups, based on the molecular mass: the high-molecular-weight glutenin subunits (HMW-GS, in the molecular mass range of 70–90 kDa) and the low-molecular-weight glutenin subunits (LMW-GS, in the molecular mass range of 20–45 kDa) [3]. In our experiments, only the gliadin fraction was studied.

Several reports investigated the relationship between the rheological properties of dough and gluten protein composition [4, 5]. Functional tests (for example, extensograph, farinograph, alveograph and baking tests) provide sufficient information about rheological properties of dough. These experiments determined that gliadins are responsible for dough viscosity and extensibility, and glutenins for dough-strength and elasticity [6]. Therefore, flour, which quality depends on the balance of gliadin and glutenin, has to have appropriate protein composition to reach proper dough viscosity and elasticity for the bread making industry [7].

Liquid chromatography is one of the major methods used to isolate and characterize wheat proteins. In the field of cereal analysis, high-performance liquid chromatography has proven its usefulness mostly in varietal identification. Reversed phase high-performance liquid chromatography (RP-HPLC) was first used for the analysis of wheat proteins about 30 years ago by Bietz. The experiments were carried out on a SynChropak RP-P C18 column. The method was able to differentiate the various gliadin protein fractions with good resolution and had the potential of complementing electrophoresis methods or serving as an alternative to them. The obtained results demonstrate that RP-HPLC of gliadins was capable of distinguishing two groups of varieties [8].

The method was improved in 1985 by Bietz et al. [9] by the modification of solvent composition, gradient time, flow rate and temperature. A rapid analysis was developed, but only with poor resolution. Nevertheless, major differences observed in long runs between wheat varieties remained still apparent. Later, Wieser et al. utilized a method with 30 min cycle time in 1998 [10], but once again, only poor resolution was obtained.

The next option for increasing resolution was brought by the improvement of the stationary phase technology. Yu et al. separated water-soluble proteins of wheat grains on a Waters 300SB C18 column filled with 1.7  $\mu\text{m}$  fully porous particles [11]. Fastest mass transfer kinetics of proteins was observed by Gritti et al. for columns filled with core-shell particles having large core-to-particle ratio, or a second external thin layer with large mesopores (200–300 Å) and high porosity (0.5–0.7). For proteins, the eddy diffusion term, which is significantly lower for this type of particles than for conventional ones, together with the solid-liquid mass transfer resistance, governs the kinetic performance of columns [12]. Naeem and Sapirstein applied first a core-shell-type stationary phase for the separation of reduced glutenin subunits in 2007 [13].

Nowadays, combining mass spectrometry (MS) with the different separation techniques enables the detailed characterization of wheat protein subunits [14–18]. Qian et al. demonstrated that offline coupling of RP-HPLC with matrix-assisted laser desorption ionization mass

spectrometry (MALDI-MS) is a powerful tool in the characterization and identification of wheat proteins. As a part of the analysis, enzymatic digests of the protein fractions collected during the RP-HPLC separation (peptide map) were also analysed by MS, and the results were compared with the predictions made based on wheat DNA databases [14]. Wheat proteins can also be identified by first separating them using 2D gel electrophoresis to reduce the complexity of the sample, and then analysing the spots on the gel using nano-liquid chromatography tandem mass spectrometry (nano-LC-MS/MS) of tryptic peptides [15]. Rombouts et al. used cystein labeling of wheat prolamins prior to analysing them by electrospray ionization-tandem mass spectrometry (ESI-MS/MS) on peptide level after enzymatic digestion [16]. Nevertheless, MS techniques are still too expensive for routine analysis of wheat samples during industrial processes, and because of that, the cost-effective and reliable varietal identification of wheat samples remains a problem to be solved.

Statistical methods are generally used for analysing raw analytical data. The analysis of variance (ANOVA) is often carried out in the beginning of exploratory data analysis. This methodology includes a group of statistical models to test for significant differences between means. In the case of nested designs, the levels of a factor are nested within the levels of other factors. Andersson et al. used nested ANOVA to evaluate variance components of different factors in a study of automatic and manual sampling for ochratoxin A in barley grain [19].

Hierarchical clustering (HC) is also a widely used exploratory data analysis method. This unsupervised pattern recognition algorithm goes through the data joining the most similar clusters together until all observations have been joined together into one cluster. At the end, a cluster tree or dendrogram, which represents a multilevel hierarchy, is created. Lima et al. applied HC for testing the homogeneity of a wheat flour laboratory reference material for inorganic analysis [20].

Principal component analysis (PCA) may also be one of the first steps of a data analysis procedure [20]. It transforms the original, possibly correlated variables into a set of new uncorrelated ones called principal components (PCs). PCs form orthogonal axes representing directions of the greatest variance in the data. The first principal component (PC1) has the largest eigenvalue, and is responsible for the most of the total variance. PC2 accounts for the most of the remaining variance. Only the PCs responsible for a large percentage of the total variance are worth to be retained. The coordinates of the original observations in the subspace of the PCs are the scores. The correlation coefficients of the original variables and the PCs are the loadings. PCA can also be used as a dimension reduction method [21].

Classification or discriminant analysis algorithms are used to assign observations of a test set to different classes on the basis of a training set, which contains objects with known category membership. Quadratic discriminant analysis (QDA) is a method where covariance matrices are assumed to be different for the classes and the decision boundaries are ranging from paraboloids to spheres. In the case of linear discriminant analysis (LDA) and nearest mean classification (NMC), the classes are separated by hyperplanes. A pooled covariance matrix is calculated for LDA and only the mean vectors for NMC. The latter method can be effective in small-sample scenarios [21].

When the number of variables exceeds the number of observations, the covariance matrix becomes singular. In this situation, QDA and LDA cannot be applied directly on the corresponding data set. The applications of either regularization, for example, in the case of regularized discriminant analysis (RDA), or dimension reduction by means of PCA prior to classification are possible strategies to deal with this situation. Classifications are often elaborated on PCA scores [21, 22].

The objective of our study was to develop a high-resolution RP-HPLC method for analysing the gliadin protein pattern of ten different wheat varieties (Bánkúti-1201; Bezostaja-1; Fleischmann-481; Glenlea; MV-Karizma; MV-Magdaléna; MV-Mazurka; Székács-1242; TF-Komádi; TF-Rétság). Three columns, filled with superficially porous particles, were connected in series to enhance the separation efficiency. The gliadin protein pattern obtained by means of this RP-HPLC method can be applied as a fingerprint for the identification of wheat varieties. It is shown that the high-quality chromatographic results form good basis for assessing similarities and differences of wheat varieties using multivariate exploratory statistical tools (PCA, HC). Finally, the classification of wheat varieties is carried out, which is also of great importance for the baking industry, because the protein composition of flour strongly influences the rheological properties of dough during processing.

## Experimental

### Chemicals, Columns

Acetonitrile (gradient grade) was purchased from Merck (Darmstadt, Germany). Water was purified using a Milli-Q Purification System from Millipore (Bedford, MA, USA). Trifluoroacetic acid (TFA) was purchased from Sigma-Aldrich (St. Louis, MO, USA). For preparing samples, absolute ethanol was obtained from Molar Chemicals Kft.

(Budapest, Hungary). Two Phenomenex Aeris Widepore XB-C18 ( $4.6 \times 150$  mm,  $3.6 \mu\text{m}$ ) and one Phenomenex Aeris Widepore C4 ( $4.6 \times 150$  mm,  $3.6 \mu\text{m}$ ) columns (Phenomenex, Torrance, CA, USA) were used for RP-HPLC separations.

### Equipment, Software

In the sample preparation process, an IKA VXR basic VIB-RAX shaker (IKA<sup>®</sup>-Werke GmbH & Co. KG, Staufen, Germany) and a Heraeus Biofuge Pico centrifuge (DJB Labcare Ltd., Newport Pagnell, UK) were used. Polyvinylidene fluoride (PVDF) membrane filters (Millipore, Mississauga, ON, USA) were used for filtering the samples prior to the chromatographic analysis.

All measurements were performed using a Perkin Elmer Flexar FX-15 UHPLC system equipped with a binary solvent delivery pump, an autosampler with a  $50 \mu\text{L}$  sample loop and a UV detector with a  $1.2 \mu\text{L}$  flow-cell. The three columns, connected using  $100$  mm long,  $0.127$  mm internal diameter tubing, were applied in serial arrangement for the measurements. The overall extra-column variance of such a system is about  $30 \mu\text{L}^2$  (system and additional tubing), and the dwell volume about  $330 \mu\text{L}$ , respectively, which are both appropriate for the experimental setup used in this paper [23, 24].

PerkinElmer Chromera Chromatography Data System for LC Software was used for data acquisition, data processing and instrument control. The ANOVA calculations were made using STATISTICA, version 10 (StatSoft, Inc., Tulsa, USA). PCA was carried out by SIMCA v14 (MKS Data Analytics Solutions, Umeå, Sweden). HC and NMC were performed by SCAN, Software for Chemometric Analysis, Release 1.1 (Minitab Inc., State College, USA).

### Apparatus and Methodology

#### Sample Preparation

Samples for RP-HPLC analysis were prepared according to Bietz et al. [8] with some modifications. Flour samples were milled in the Agricultural Institute, Centre for Agricultural Research, Hungarian Academy of Sciences. Absolute ethanol was diluted with distilled water to 70 % prior to extraction. As the first step,  $50$  mg of flour sample was weighed into a  $1.5$  mL Eppendorf tube. Gliadins were extracted with  $1$  mL of 70 % ethanol. The sample was vibrated for  $10$  min with  $2000$  rpm, then centrifuged for  $30$  min at  $13,000$  rpm. The clear supernatant was filtered through a  $0.45 \mu\text{m}$  PVDF membrane filter. The filtrate was injected directly into the UHPLC system.

### Chromatographic Separation

For the chromatographic separation, using the two Phenomenex Aeris Widepore XB-C18 and one Phenomenex Aeris Widepore C4 connected in series, the mobile phase was prepared by mixing appropriate amount of HPLC gradient grade acetonitrile, Milli-Q water and TFA. The mixtures were degassed by sonication for 10 min. The weaker mobile phase (A) consisted of water/TFA 100/0.1 V/V, while the stronger mobile phase (B) was a mixture of acetonitrile/TFA 100/0.1 V/V. The separations were carried out with the following gradient program (Table 1): the initial mobile phase composition (10 % B) was held for 1 min followed by a linear gradient program to 80 % B in 60 min. After that, the initial gradient conditions were reset in 1 min and run for 3 min to equilibrate the system. The total run time was 65 min. The mobile phase flow rate was 0.5 mL min<sup>-1</sup>. The oven temperature was set to 40 °C, and 5 µl of the sample solution was injected. The chromatograms were recorded at 210 nm.

### Experimental Design, Statistical Evaluation

An experimental design was carried out to gain data, which could be used to perform comprehensive statistical evaluation. The wheat variety was treated as fixed factor, but the plot of land used for cultivation, and the repeated sample preparations were treated as random factors. The ten different wheat varieties used in this experiment were each grown on three separated plots of land simultaneously. They were assigned randomly to the plots available for this experiment, so that differences in expression pattern of proteins caused by environmental effects could also be taken into consideration. Every sample was prepared in duplicates. The 60 samples of this experimental design (10 wheat variety, 3 plots of land, and 2 repeated sample preparations) were prepared at the same time and measured in the same sequence. They could be handled in one block during the statistical evaluation.

Despite using extremely efficient columns with core-shell-type stationary phases in combination with UHPLC system, both capable of very high resolving power,

several unresolved peak pairs still remained in these chromatograms. In such a situation, peak heights are more robust estimates of the amounts of the different compounds in the samples [25]. The uncertain position of the start and end points of the peaks or even minor changes in that of the baseline in the case of multiple coeluting peaks makes the calculation of the peak area less reproducible. After the integration, peak heights were then extracted and organized in a proper form for the statistical analysis using the STATISTICA, SCAN and SIMCA software.

ANOVA was used to gain preliminary information about the effect of the factors (wheat variety, plot of land, and repeated sample preparations) on the gliadin protein pattern. The calculations were carried out using the STATISTICA software. The results for every variable were evaluated using nested designs. The repeated sample preparations were nested in the plot of land and wheat variety factors and the plot of land in the wheat variety factor. Due to the characteristics of the factors (random and fixed factors), univariate statistical tests were performed using the heights of every 64 peak.

PCA was elaborated on the 60 × 64 data set for evaluating the variability of the gliadin protein profile of the different wheat flours. The calculations were made by the SIMCA software. The results were visualized in the subspace of the PCs and discussed in terms of PCA scores. Unique characteristics of the different classes were assessed by means of contribution plots. The goodness of fit and the predictive power of the model are given by the R<sup>2</sup>X (explained variance) and Q<sup>2</sup>X (predicted variation) parameters, respectively. The latter is estimated using cross-validation. R<sup>2</sup>X varies between 0 (no fit) and 1 (perfect fit) and approaches unity as model complexity increases. Q<sup>2</sup>X will not automatically come close to 1 with increasing number of PCs incorporated in the model. In SIMCA, PCA model is evaluated by simultaneously considering R<sup>2</sup>X and Q<sup>2</sup>X [26].

HC was also utilized to recognize similarities between samples. The two parameters of this clustering method, the linkage (complete, average, centroid, McQuitty, median, Ward) and distance (Euclidean distance, Manhattan distance, Pearson correlation) rules were varied systematically. The linkage rule is used to determine when two clusters are similar enough to be linked together. The distance rule defines the way of measuring similarity. The variables were autoscaled prior to analysis in the case of Euclidean and Manhattan distance, because the peak heights of the different peaks had very different ranges. Pearson correlation was used with absolute values. The algorithm was set up to find ten clusters. The results were compared with those of PCA.

After evaluating the characteristics of the data set using PCA and HC, NMC was chosen for the classification of the RP-HPLC results. It is well suited for classification tasks

**Table 1** RP-HPLC gradient program used for the separation of gliadin proteins

Time (min)	A (%)	B (%)
0	90	10
1	90	10
61	20	80
62	90	10
65	90	10

with small number of observations, since only the coordinates of the cluster centroids have to be calculated using the observations of the training data set. The algorithm was elaborated on the scores of the PCA, which was applied for dimension reduction. The number of components in the model was increased up to 8, and error rates of NMC were evaluated. To assess the predictive accuracy of the algorithm, a threefold cross-validation was carried out. The data set was split into training and validation sets. The plots, on which the different wheat varieties were cultivated, were used as splitting criteria. Two-third of the original data set was used as training set, and the remaining results were then classified. At the end of this procedure, every observation was used for validation only once.

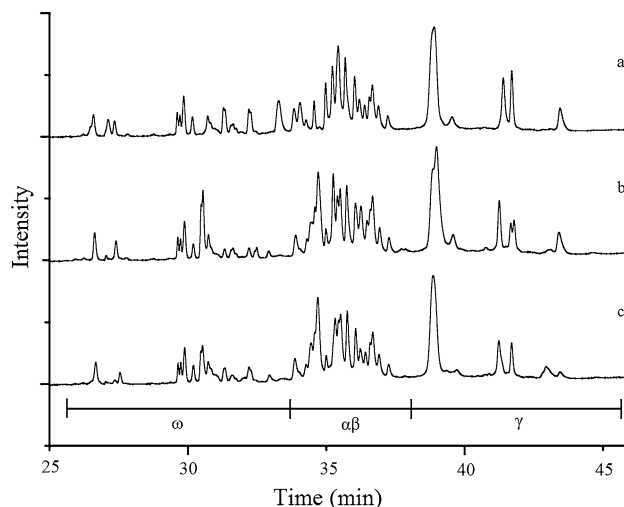
## Results and Discussion

### Chromatographic Method Development

The gliadin fraction of the wheat protein consists of high number of components which have to be separated to reveal the fine structure of the protein distribution prior to integration. It is very important to reach very high resolution to be able to recognize even the smallest differences in the protein pattern of the corresponding wheat samples.

The two Aeris Widepore XB-C18 and the single Aeris Widepore C4 columns were connected in series to gain as much separation power as possible for the separations. Nevertheless, it was more time efficient to perform the early experiments with a single Aeris Widepore XB-C18 column. Before using a single column as a downscale model for the latter experiments, the selectivity of the Aeris Widepore XB-C18 and the Aeris Widepore C4 columns had to be compared. Under the same conditions, no significant differences in the separation of gliadin proteins were observed using these columns with different surface chemistry (data not shown). During the initial experiments, gradient time was varied in the range of 10–30 min and separation temperature in the range of 30–50 °C. Higher oven temperatures were not applied during the separations, because the maximum allowed temperature for the Aeris Widepore C4 material was 60 °C. Lower oven temperatures also support the preservation of the stationary phase.

Higher gradient times improved slightly the resolution, but no significant effect of the temperature on the overall separation was observed in the range applied. In every case, some peaks remained unresolved in the chromatograms. Because of these results for the setup, in which all the three columns mentioned previously were connected in series, 60 min gradient time—3 times 20 min for a single column—and 40 °C oven temperature were applied to ensure high resolution in a reasonable analysis time. Figure 1



**Fig. 1** Example of the separation of gliadin proteins extracted from three different wheat samples using the coupled-column chromatographic method. *a* Sample 12 (Bezostaja-1), *b* sample 18 (Fleischmann-481), *c* sample 24 (Glenlea). The retention time windows of the  $\omega$ ,  $\alpha\beta$  and  $\gamma$  gliadin fractions are highlighted below the chromatograms

shows chromatograms in overlay representation obtained using this novel three-column chromatographic method of the gliadin proteins of three different wheat varieties. The location of the  $\omega$ ,  $\alpha\beta$  and  $\gamma$  gliadin fractions on the chromatograms is highlighted according to Mejías et al. [27].

The integration of such chromatograms, which contains several poorly resolved peaks, is very difficult using preset integration parameters. Most of the software support manual integration, which offers the easiest way to deal with such situations. Due to these facts, in the case of the 60 chromatograms obtained by analysing the samples from the experiment described in the previous section, manual integration was applied. The peak heights of 64 different peaks, which were identified based on retention time, were extracted and used for further statistical analysis. As discussed before in the “[Experimental Design, Statistical Evaluation](#)” section, peak heights are better candidates for representing the amounts of the compounds remained unresolved in the chromatograms.

### Statistical Evaluation

#### Anova

As a result of the chromatographic measurements of the 60 samples (10 wheat varieties cultivated on 3 different plots of land each, and prepared in duplicates prior to analysis), a data matrix was obtained, which contained 3840 data points (60 samples  $\times$  64 peaks). For studying the effect of the different factors (plot of land; wheat variety; repeated sample



**Table 2** Univariate test of significance (nested design) for peak 28 obtained with STATISTICA 10 data analysis software. Significant effects are highlighted with bold text

Effect	Effect (F/R)	SS	Degr. of freedom	MS	Den. Syn. error df	Den. Syn. error MS	<i>F</i>	<i>p</i>
Intercept	F	$3.7 \times 10^9$	1	$3.7 \times 10^9$	20	$3.0 \times 10^6$	1212.9	0.000
1-Wheat variety	F	$5.5 \times 10^9$	9	$6.1 \times 10^8$	20	$3.0 \times 10^6$	199.9	0.000
2-Plot of land(2)	R	$6.1 \times 10^7$	20	$3.1 \times 10^6$	30	$3.5 \times 10^5$	8.7	0.000
3-Sample prep. (1-2)	R	$1.1 \times 10^7$	30	$3.5 \times 10^5$	0	0		
Error			0					

*F* for fixed and *R* for random in effect column, *SS* for sum of squares, *MS* for mean squares, *Den.Syn.* for Denominator synthesis, which refers to the Satterthwaite's method of denominator synthesis used by STATISTICA 10 for constructing error terms in mixed or random models. *F* indicates the calculated value for *F* test, *p* the calculated value for probability. Wheat variety and *plot* of land had significant effect. Sample preparation was found negligible

preparations) on the protein pattern, ANOVA with a nested design was applied. Due to the characteristics of the factors (random and fixed factors), univariate statistical tests were carried out using the heights of every 64 peak. The results of this statistical analysis showed that in most of the cases (58 from 64 peaks), the wheat variety and the plot of land had significant effect on protein pattern. Table 2 shows an example for the ANOVA. The effect of the repeated sample preparations was found negligible. This pattern was also confirmed later by the hierarchical clustering methods and PCA, where the results of the repeated sample preparations were found to be highly similar in several cases.

### Principal Component Analysis

The comparison of similarities and differences of the RP-HPLC results was done first in a multivariate form after the dimension reduction by means of PCA. The first four PCs were responsible for 76.8 % of the total variance (PC1 27.1 %, PC2 23.7 %, PC3 14.9 %, PC4 11.1 %). Further components had less contributions to the model (PC5 7.2 %, PC6 4.5 %, PC7 3.4 %, PC8 2.6 %). Despite the fact that the value of  $Q^2X$ , which corresponds to the predictive ability of the PCA model, had maximum at PC12, only eight PCs with the largest eigenvalues were kept in the model for further evaluation with NMC.

MV-Magdaléna and TF-Rétság are very different from the other wheat varieties in terms of PC2 and PC3 scores, respectively. The average PC2 score of the observations of MV-Magdaléna was 11.26 and those of the other wheat varieties ranged from -3.26 to 0.45 (Fig. 2a). The average PC3 score of the observations of TF-Rétság was 7.19 and those of the other wheat varieties ranged from -4.42 to 2.77 (Fig. 2b). These unique characteristics emphasize the fact that PC2 and PC3 contribute 48.6 % to the total variance. The observations are best separated in the plane of PC1 and PC4 (Fig. 2c).

It can be seen in Fig. 2 that Bezostaja-1, Glenlea, MV-Karizma, MV-Magdaléna, MV-Mazurka, and TF-Rétság

form completely separated groups in the subspace of PC1 and PC4. Clusters of Bánkúti-1201 and Fleischmann-481 are overlapped, and observations of Székács-1242 and TF-Komádi are located very close to each other. Based on the contribution plots, the similar behavior of the different wheat varieties in terms of the original features (peak heights) was evaluated. Bánkúti-1201 and Fleischmann-481 expressed similar levels of  $\alpha\beta$  and  $\gamma$ , but very different levels of  $\omega$  gliadin proteins, while Székács-1242 and TF-Komádi showed quite similar behavior in terms of all gliadin protein fractions (Fig. 3). Only minor differences can be seen on the contribution plots of the latter pair of wheat varieties. MV-Mazurka and TF-Rétság had balanced gliadin pattern compared to the average (data not shown). Bezostaja-1, Glenlea and MV-Karizma expressed less and MV-Magdaléna on the contrary more gliadin proteins overall compared to the average expression levels of the tested wheat varieties (data not shown).

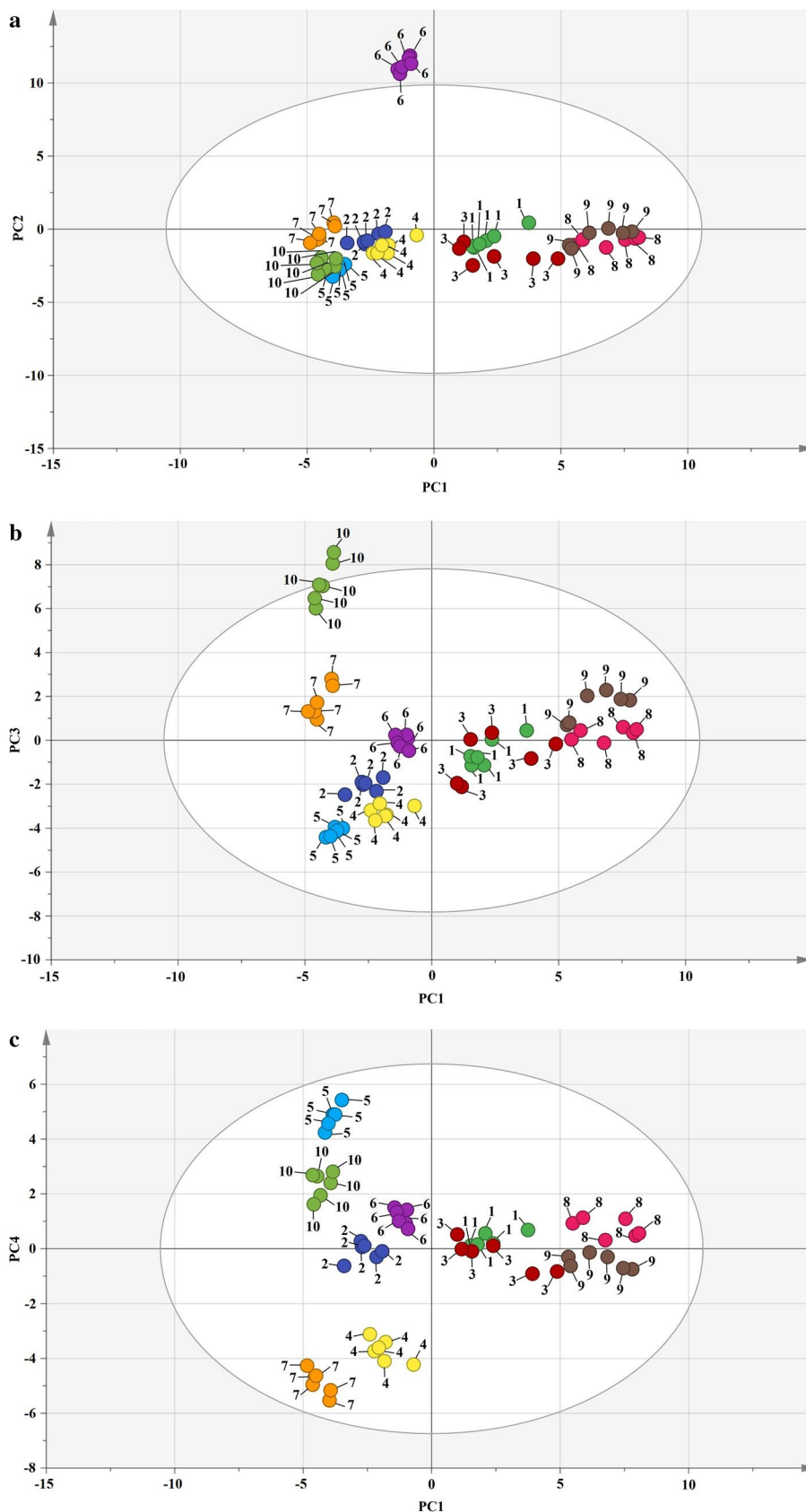
### Hierarchical Clustering

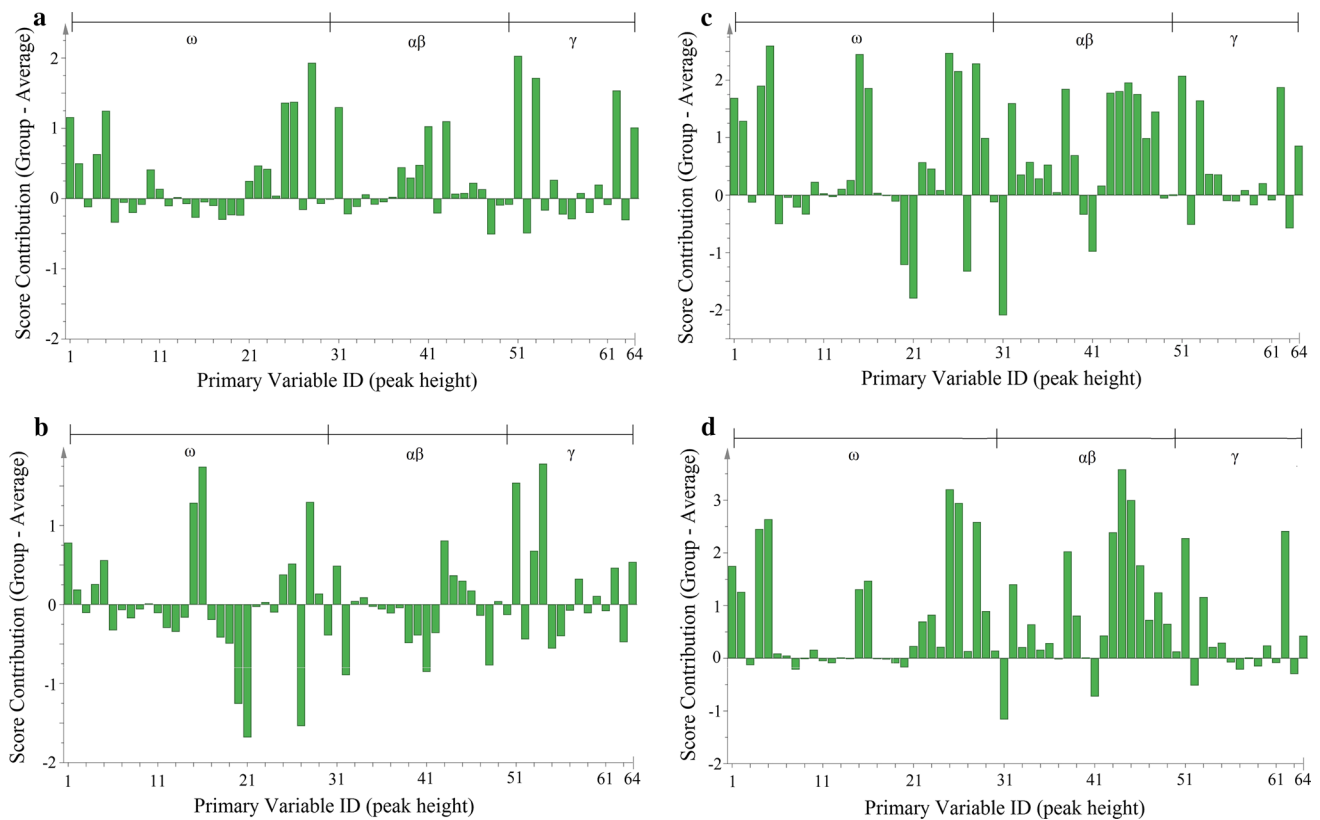
With the chromatographic results in hand (peak heights in the present study), the similarities and differences of the protein distributions of the samples were also evaluated using this pattern recognition methodology. HC was carried out in several ways as described in the “[Experimental Design, Statistical Evaluation](#)” section.

In most of the cases, the observations of the same wheat variety were found to be more similar to each other than to those of other cultivars. Only two parallel samples (same plot of land) of Fleischmann-481 were assigned occasionally to other clusters. The clusters of Székács-1242 and TF-Komádi were every time joined at the next hierarchical level, and all observations of Bánkúti-1201 and four of Fleischmann-481 were also grouped together in all HC experiments. These findings are in accordance with the PCA results.

The structure of the experimental setup described in the “[Experimental Design, Statistical Evaluation](#)” section is

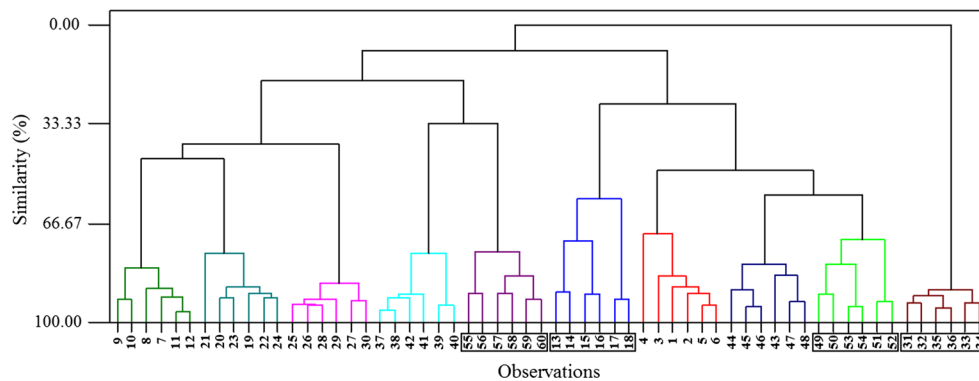
**Fig. 2** PCA score plot in the plane of PC2 and PC1 (a), PC3 and PC1 (b), PC4 and PC1 (c). 1 Bánkuti-1201 (observations 1–6); 2 Bezostaja-1 (observations 7–12); 3 Fleischmann-481 (observations 13–18); 4 Glenlea (observations 19–24); 5 MV-Karizma (observations 25–30); 6 MV-Magdaléna (observations 31–36); 7 MV-Mazurka (observations 37–42); 8 Székács-1242 (observations 43–48); 9 TF-Komádi (observations 49–54); 10 TF-Rétság (observations 55–60)





**Fig. 3** PCA contribution plots of Bánkúti-1201 (a), Fleischmann-481 (b), Székács-1242 (c) and TF-Komádi (d). The protein expression levels of the wheat varieties are compared to that of the average of

all observations. “ $\omega$ ”, “ $\alpha\beta$ ” and “ $\gamma$ ” refers to the  $\omega$ ,  $\alpha\beta$  and  $\gamma$  gliadin fractions, respectively



**Fig. 4** Dendrogram resulting from HC using complete linkage, Manhattan distance and autoscaled variables. The samples from the different wheat varieties were classified into separate classes (10 clusters, each having 6 samples of the same wheat variety). Classes having the repeated sample preparations in separate clusters are bracketed. Bánkúti-1201 (observations 1–6); Bezostaja-1 (observations 7–12); Fleis-

chmann-481 (observations 13–18); Glenlea (observations 19–24); MV-Karizma (observations 25–30); MV-Magdaléna (observations 31–36); MV-Mazurka (observations 37–42); Székács-1242 (observations 43–48); TF-Komádi (observations 49–54); TF-Rétság (observations 55–60)

clearly recognizable on the dendrogram in Fig. 4 (HC with Manhattan distance and complete linkage with autoscaled variables) in the case of wheat variety Fleischmann-481, TF-Rétság, TF-Komádi and MV-Magdaléna. The repeated

sample preparations were more similar to each other than to those of other plots of lands. Similar dendrogram was obtained in the case of HC with Manhattan distance and Ward linkage (autoscaled variables), Euclidean distance and



**Table 3** Confusion matrix of the threefold cross-validation representing the classification accuracy of NMC algorithm on PCA scores

Number of principal components in the PCA model: 3											
Assigned class	True class										Total
	1	2	3	4	5	6	7	8	9	10	
1	<b>4</b>	0	5	0	0	0	0	0	0	0	9
2	0	<b>6</b>	0	0	0	0	0	0	0	0	6
3	1	0	<b>0</b>	0	0	0	0	0	0	0	1
4	0	0	0	<b>6</b>	0	0	0	0	0	0	6
5	0	0	0	0	<b>6</b>	0	0	0	0	0	6
6	0	0	0	0	0	<b>6</b>	0	0	0	0	6
7	0	0	0	0	0	0	<b>6</b>	0	0	0	6
8	0	0	1	0	0	0	0	<b>6</b>	2	0	9
9	1	0	0	0	0	0	0	0	<b>4</b>	0	5
10	0	0	0	0	0	0	0	0	0	<b>6</b>	6
Total	6	6	6	6	6	6	6	6	6	6	Error rate: 16.7 %

The first three PCs were used in the PCA model. Class 1: observations 1–6 (Bánkuti-1201); Class 2: observations 7–12 (Bezostaja-1); Class 3: observations 13–18 (Fleischmann-481); Class 4: observations 19–24 (Glenlea); Class 5: observations 25–30 (MV-Karizma); Class 6: observations 31–36 (MV-Magdaléna); Class 7: observations 37–42 (MV-Mazurka); Class 8: observations 43–48 (Székács-1242); Class 9: observations 49–54 (TF-Komádi); Class 10: observations 55–60 (TF-Rétság)

bold in matrix as a mathematical modelling

Ward linkage (autoscaled variables) and Pearson distance and Ward linkage (without autoscaling) regarding this phenomenon. The observations originating from the different plots in the case of the other wheat varieties showed different patterns. This is probably because the cultivation on different plots has not the same effects on the protein expression in the case of the different wheat varieties. These findings are confirmed by the results of the ANOVA, where in most of the cases (58 from 64 peaks), only the wheat variety and the plot of land proved to have a significant effect on protein pattern.

According to our results, the linkage rule had much stronger impact on the quality of the classification than the distance rule (data not shown).

### Classification

As described in the “[Experimental Design, Statistical Evaluation](#)” section, after a dimension reduction using PCA, NMC was applied because of its simplicity for the classification of the different wheat varieties. This supervised pattern recognition algorithm, which uses Euclidean distance, assigns the observation to clusters with the closest centroid.

The prediction accuracy of NMC strongly depends on the number of PCs used in the PCA model. Keeping only the first, then the first two PCs resulted in 50 and 25 % error rates, respectively. Even observations of different wheat varieties with unique gliadin protein pattern were assigned not to their true classes. Using only the scores of first PC, two observations of TF-Rétság were assigned to the class of MV-Mazurka, which were considered to be very different based

on both PCA and HC. However, using the first 8 PCs, responsible for the 94.5 % of the total variance, the classification ended without any misclassification. Further error rates after the threefold cross-validation procedure were 16.7 % (3PCs), 13.3 % (4 PCs), 5 % (5 PCs) and 3 % (6 PCs). No misclassification occurred first when scores of the PC1–7 were applied.

When the first three PCs remained in the PCA model, the observations of Bánkuti-1201, Fleischmann-481, Székács-1242 and TF-Komádi were mixed. Table 3 shows the confusion matrix of the threefold cross-validation. These wheat varieties were found also to be similar in the case of PCA and HC. The error rate using PC1–4 decreased slightly, but again, the previously mentioned four varieties were involved in misclassification.

### Conclusion

HPLC–MS techniques can produce more specific information than 1D chromatographic methods. However, the investments and operating costs, which are of great importance in the baking industry, are significantly higher than those of a single HPLC instrument used in this study. The presented RP-HPLC method provided detailed information regarding the gliadin fraction of the wheat proteins. This can be very useful in the process of wheat varietal identification. According to our best knowledge, such separation of gliadin proteins has never been published before.

Similar separations of the gliadin proteins were obtained in terms of selectivity using the Aeris Widepore XB-C18

and the Aeris Widepore C4 columns. Varying the temperature (30–50 °C) and the gradient time (10–30 min) for a single XB-C18 phase (downscale model) has not resulted in an overall better separation in the initial experiments. Based on these results, for the column-coupling RP-HPLC measurements using the two XB-C18 and one C4 columns, 60 min gradient time and 40 °C oven temperature were applied.

Nested ANOVA revealed that there were significant differences not only between the wheat varieties but also between plots of lands they were cultivated on. The repeated sample preparations factor proved to be not significant in most of the cases. Using PCA and HC as multivariate exploratory data analysis methods, Bezostaja-1, Glenlea, MV-Karizma, MV-Magdaléna, MV-Mazurka and TF-Rétság were found to have quite unique gliadin protein patterns, based on which they could be easily differentiated from the other wheat varieties. The four remaining cultivars, Bánkúti-1201, Fleischmann-481, Székács-1242 and TF-Komádi proved to be more similar to each other.

The cross-validated error rate of NMC decreased with every principal component added to the PCA model. Six wheat varieties out of ten could be classified on the scores of the first three PCs. The observations of the four wheat varieties found to be similar also with HC were involved in misclassification.

This column-coupling RP-HPLC method was used only for the separation of gliadin proteins, which are responsible for dough viscosity and extensibility [6]. In the future, this study could be extended also for the whole protein pattern of wheat flours. Two-dimensional chromatographic methods should also be taken into consideration in the future.

**Acknowledgments** The authors are very grateful for the valuable inputs and discussions in combination with chemometric methods provided by Dr. Pál Szepesváry from Eötvös Lóránd University.

#### Compliance with Ethical Standards

**Conflict of Interest** All authors declare that they have no conflict of interest.

**Ethical approval** This article does not contain any studies with human participants or animals performed by any of the authors.

## References

- Malik AH, Kuktaite R, Johansson E (2013) Combined effect of genetic and environmental factors on the accumulation of proteins in the wheat grain and their relationship to bread-making quality. *J Cereal Sci* 57:170–174
- Herbert W (2007) Chemistry of gluten proteins. *Food Microbiol* 24:115–119
- D'Ovidio R, Masci S (2004) The low-molecular-weight glutenin subunits of wheat gluten. *J Cereal Sci* 39:321–339
- Khatkar BS, Fido RJ, Tatham AS, Schofield JD (2002) Functional properties of wheat gliadins. II. Effects on dynamic rheological properties of wheat gluten. *J Cereal Sci* 35:307–313
- Khatkar BS, Fido RJ, Tatham AS, Schofield JD (2002) Functional properties of wheat gliadins. I. Effects on mixing characteristics and bread making quality. *J Cereal Sci* 35:299–306
- Barak S, Mudgil D, Khatkar BS (2013) Relationship of gliadin and glutenin proteins with dough rheology, flour pasting and bread making performance of wheat varieties. *Food Sci Technol* 51:211–217
- Li Vigni M, Baschieri C, Marchetti A, Cocchi M (2013) RP-HPLC and chemometrics for wheat flour protein characterisation in an industrial bread-making process monitoring context. *Food Chem* 139:553–562
- Bietz JA (1983) Separation of cereal proteins by reversed-phase high performance liquid chromatography. *J Chromatogr* 255:219–238
- Bietz JA, Cobb LA (1985) Improved procedures for rapid wheat varietal identification by reversed-phase high performance liquid chromatography of gliadin. *Cereal Chem* 62:332–339
- Wieser H, Antes S, Seilmeier W (1998) Quantitative determination of gluten protein types in wheat flour by reserved-phase high performance liquid chromatography. *Cereal Chem* 75:644–650
- Yu Z, Han C, Wang S, Lv D, Chen G, Li X, Yan Y, Jiang G-L (2013) Fast separation and characterization of water-soluble proteins in wheat grains by reversed-phase ultra performance liquid chromatography (RP-UPLC). *J Cereal Sci* 57:288–294
- Gritti Fabrice, Horvath Krisztian, Guiochon Georges (2012) How changing the particle structure can speed up protein mass transfer kinetics in liquid chromatography. *J Chromatogr A* 1263:84–98
- Naeem HA, Sapirstein HD (2007) Ultra-fast separation of wheat glutenin subunits by reversed-phase HPLC using a superficially porous silica-based column. *J Cereal Sci* 46:157–168
- Qian Y, Preston K, Krokshin O, Mellish J, Ens W (2008) Characterization of wheat gluten proteins by HPLC and MALDI TOF mass spectrometry. *J Am Soc Mass Spectr* 19:1542–1550
- Mamone G, Addeo F, DeCaro S, Ferranti P, Di Luccia A (2009) Proteomic-based analytical approach for the characterization of glutenin subunits in durum wheat. *J Mass Spectrom* 44:1709–1723
- Rombouts I, Lagrain Bunnbauer M, Delcour JA, Koehler P (2013) Improved identification of wheat gluten proteins through alkylation of cysteine residues and peptide-based mass spectrometry. *Sci Rep* 3:2279
- Liu L, Wang A, Appels R, Ma J, Xia X, Lan P, He Z, Bekes F, Yan Y, Ma W (2009) A MALDI-TOF based analysis of high molecular weight glutenin subunits for wheat breeding. *J Cereal Sci* 50:295–301
- Lagrain B, Brunnbauer M, Rombouts I, Koehler P (2013) Identification of intact high molecular weight glutenin subunits from the wheat proteome using combined liquid chromatography-electrospray ionization mass spectrometry. *PLoS One* 8:1–10
- Andersson MG, Reiter EV, Lindqvist P-A, Razzazi-Fazeli E, Häggblom P (2011) Comparison of manual and automatic sampling for monitoring ochratoxin A in barley grain. *Food Addit Contam* 28:1066–1075
- Lima DC, dos Santos AMP, Araujo RGO, Scarmínio IS, Bruns RE, Ferreira SLC (2010) Principal component analysis and hierarchical cluster analysis for homogeneity evaluation during the preparation of a wheat flour laboratory reference material for inorganic analysis. *Microchem J* 95:222–226
- Dougherty ER, Shmulevich I, Chen J, Wang ZJ (2005) Genomic signal processing and statistics. Hindawi Publishing Corporation, New York

22. Park S, Ku YK, Seo MJ, Kim DY, Yeon JE, Lee KM, Jeong S-C, Yoon WK, Harn CH, Kim HM (2006) Principal component analysis and discriminant analysis (PCA–DA) for discriminating profiles of terminal restriction fragment length b polymorphism (T-RFLP) in soil bacterial communities. *Soil Biol Biochem* 38:2344–2349
23. Fekete S, Fekete J (2011) The impact of extra-column band broadening on the chromatographic efficiency of 5 cm long narrow-bore very efficient columns. *J Chromatogr A* 1218:5286–5291
24. Fekete S, Kohler I, Rudaz S, Guillarme D (2014) Importance of instrumentation for fast liquid chromatography in pharmaceutical analysis. *J Pharm Biomed Anal* 87:105–119
25. Berky R, Fekete S, Fekete J (2012) Enhancing the quality of separation in one dimensional peptide mapping by using mathematical transformation. *Chromatographia* 75:305–312
26. Eriksson L, Byrne T, Johansson E, Trygg J, Vikström C (2013) Multi- and megavariate data analysis. Malmö, Sweden
27. Mejías JH, Lu X, Osorio C, Ullman JL, von Wettstein D, Rustgi S (2014) Analysis of wheat prolamins, the causative agents of celiac sprue, using reversed phase high performance liquid chromatography (RP-HPLC) and matrix-assisted laser desorption ionization time of flight mass spectrometry (MALDI-TOF-MS). *Nutrients* 6:1578–1597