Available online at www.sciencedirect.com
SCIENCE DIRECT®
PHYTOCHEMISTRY
Phytochemistry 65 (2004) 1531–1548
www.elsevier.com/locate/phytochem

Review

Multivariate approaches in plant science

David M. Gottlieb [a], Jakob Schultz [b], Susanne W. Bruun [b], Susanne Jacobsen [b], Ib Søndergaard [b,*]

[a] *Plasma Product Division, Statens Serum Institut, Artillerivej 5, DK-2300 Copenhagen S, Denmark*
[b] *Biochemistry and Nutrition Group, BioCentrum-DTU, Technical University of Denmark, Søltofts Plads, building 224, DK-2800 Kgs. Lyngby, Denmark*

Received 18 December 2003; received in revised form 1 April 2004
Available online 28 May 2004

Abstract

The objective of proteomics is to get an overview of the proteins expressed at a given point in time in a given tissue and to identify the connection to the biochemical status of that tissue. Therefore sample throughput and analysis time are important issues in proteomics. The concept of proteomics is to encircle the identity of proteins of interest. However, the overall relation between proteins must also be explained. Classical proteomics consist of separation and characterization, based on two-dimensional electrophoresis, trypsin digestion, mass spectrometry and database searching. Characterization includes labor intensive work in order to manage, handle and analyze data. The field of classical proteomics should therefore be extended to also include handling of large datasets in an objective way. The separation obtained by two-dimensional electrophoresis and mass spectrometry gives rise to huge amount of data. We present a multivariate approach to the handling of data in proteomics with the advantage that protein patterns can be spotted at an early stage and consequently the proteins selected for sequencing can be selected intelligently. These methods can also be applied to other data generating protein analysis methods like mass spectrometry and near infrared spectroscopy and examples of application to these techniques are also presented. Multivariate data analysis can unravel complicated data structures and may thereby relieve the characterization phase in classical proteomics. Traditionally statistical methods are not suitable for analysis of the huge amounts of data, where the number of variables exceed the number of objects. Multivariate data analysis, on the other hand, may uncover the hidden structures present in these data. This study takes its starting point in the field of classical proteomics and shows how multivariate data analysis can lead to faster ways of finding interesting proteins. Multivariate analysis has shown interesting results as a supplement to classical proteomics and added a new dimension to the field of proteomics.

Keywords: Multivariate data analysis; Chemometrics; Proteomics; Wheat; Gliadins; Gluten; Quality; Mass spectrometry; 2D-gel electrophoresis; NIR

Contents

* Corresponding author. Tel.: +45-45252733; fax: +45-45886307.
E-mail address: ibs@biocentrum.dtu.dk (I. Søndergaard).

0031-9422/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.phytochem.2004.04.008

1. Introduction

Two-dimensional gel electrophoresis (2DGE) and mass spectrometry (MS) used in combination constitute a strong analytical tool used in "classical" proteomics (Fig. 1), in which MS is used for identification of proteins. By using the two analytical techniques independently of each other, but coupled with multivariate analysis, we have added a new dimension to the field of proteomics. Multivariate analysis improves the data handling in proteomics, and thereby narrowing down proteins of interest much faster (Fig. 2). Our method should therefore be considered as a strong supplement to the "classical" proteomics.

1.1. Multivariate data analysis

One of the troublesome issues in proteomics is the handling of data with respect to characterization. The field of chemometrics mainly concerns multivariate analysis applied to data from chemistry (Martens and Martens, 2001). Chemometric studies deal with the

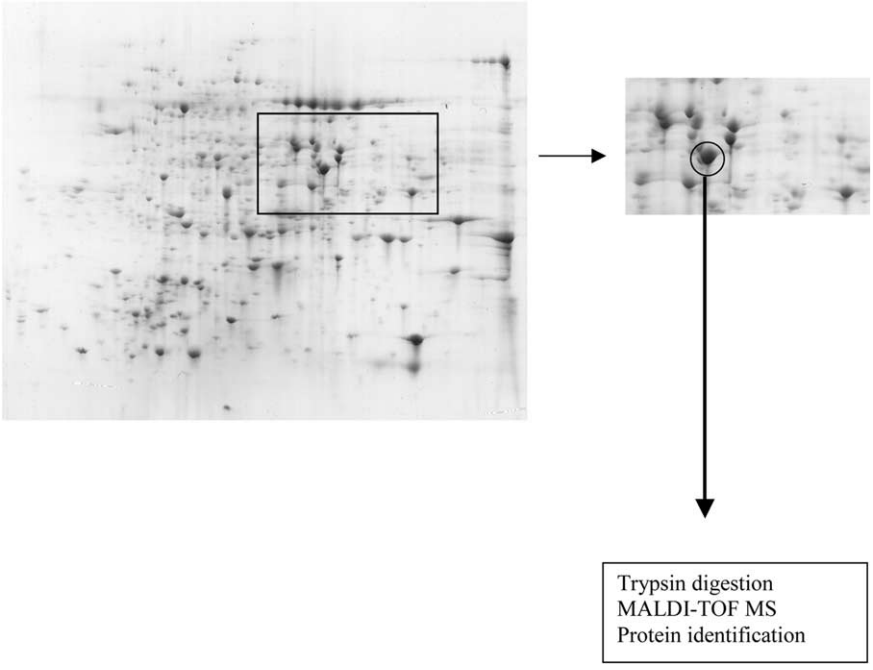


Fig. 1. The water-soluble fraction of a barley variety separated by 2DGE. The proteins are identified after trypsin digestion by mass spectrometry. Labour-intensive work is needed both when many gels are to be compared and when all the protein spots on one gel have to be sequenced and identified.

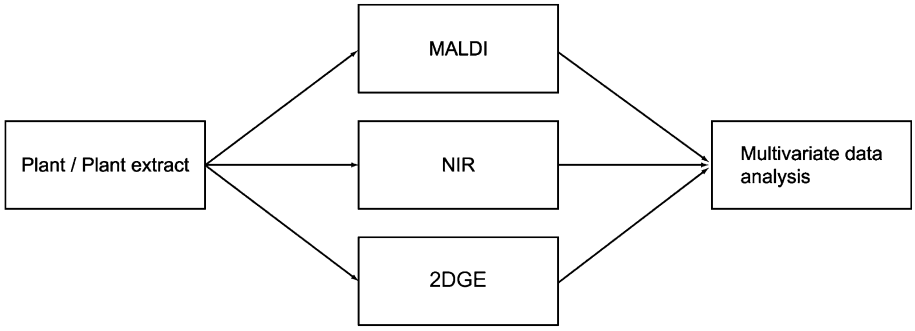


Fig. 2. Multivariate data analysis in plant science.

overall managing, handling and analysis of data collected from, e.g., 2DGE, MS or near infrared (NIR) spectroscopy.

Multivariate analysis builds on the application of statistical and mathematical methods, and includes the analysis of data with many observed variables, as well as the study of systems with many important types of variation (Martens and Martens, 2001).

The performance of multivariate analysis builds on two fundamental principles: (1) formulation of an experiment *before* data analysis (planning) and (2) problem reduction *during* and *after* data analysis (modeling). However, it must be emphasized that prior conventions, theories and expectations may have a restrictive influence on analysis, if used blindly. Multivariate analysis is therefore a balance between prior knowledge and new input gained during analysis.

The explorative data analysis is an important aspect during multivariate analysis. Before any hypotheses are arranged, explorative data analysis can give an insight in the multivariate chaos by means of scores (latent variables) and loading weights. An essential aspect in explorative data analysis is the outlier control. It can always be expected that data include errors as a consequence of typing errors, instrument errors, sampling errors etc. Hypothesis-generating analysis is a natural consequence of the entire concept behind multivariate analysis. In traditional statistical terms, a hypothesis is set up first and then experiments are carried out in order to demonstrate this hypothesis. This is known as deductive analysis. In contrast to traditional statistical methods, multivariate analysis is an inductive analysis, where hypotheses can be set up after having carried out the computational experiments.

Principal Component Analysis (PCA) is an unsupervised multivariate analysis technique used for transforming a set of observed variables into a new set of variables, which then are uncorrelated to one another (Everitt and Dunn, 1991). The basic idea is to find hidden structures in a dataset in order to describe these structures. The strength of PCA is the provision of low-dimensional plots of the data, e.g., to project many dimensions onto a few dimensions. On this basis it is possible to identify outlying observations, clusters of similar observations and other data structures. As the name indicates, the technique is based on principal components, a mathematical technique for an orthogonal orientation to principal axes. A principal component is also referred to as a latent variable. This variable cannot be measured directly but must be expressed as a linear combination of a set of input variables (Martens and Martens, 2001). The PCs form a rearranged multidimensional space based on a bilinear model of the data matrix $\mathbf{X}$, meaning that $\mathbf{X}$ is decomposed into a structural part and an error part. The structural part consists of a scorematrix, $\mathbf{T}$, and a transposed loadingmatrix, $\mathbf{P}^{\mathbf{T}}$, while the error part is termed $\mathbf{E}$ (Esbensen et al., 2000). Equation (1) is the mathematical skeleton of principal component model

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P}^{\mathbf{T}} + \mathbf{E}. \quad (1)$$

PCA is capable of transforming a large number of possible correlated variables to a smaller number of uncorrelated variables, PCs. The original axes are being replaced by PC-axes, where each PC-axis is a linear combination of the original variables.

The relationship of the PCs to the samples (the data rows, t_i) is called scores, and to the variables (data columns, p_i) called loadings. The new uncorrelated variables are represented in decreasing order of importance, which means that the first PC covers as much as possible of the variation in the dataset, and each subsequent component covers as much as possible of the remaining variation. The second PC is calculated orthogonal to the first PC, in that way to ensure complete non-correlation between the first and second PC-axis. The third PC-axis goes through the maximal variation described in the remaining dataset, i.e., not described in the first and second PC. This decomposition continues until all systematic variation is explained. When all variation is explained, the original data matrix has been reduced. It is thus possible to concentrate on only two or three dimensions at a time.

Although most variation is described in the first PC, it does not necessarily make the first PC the most interesting PC.

The **Y**-data structure is used to guide the decomposition of the **X**-matrix in order to reduce the original **X**-data to a small number of latent variables, termed Partial Least Squares (PLS) components (Esbensen et al., 2000). Calibration involves relating the two sets of data by regression modeling:

$$\mathbf{Y} = \mathbf{X} \cdot \mathbf{B}, \tag{2}$$

where **B** is a matrix containing b-regression vectors expressing the link between variation in the predictors and variation in the response. **X** and **Y** are defined in equations (3) and (4), respectively:

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P}^{\mathbf{T}} + \mathbf{E}, \tag{3}$$

$$\mathbf{Y} = \mathbf{T} \cdot \mathbf{Q}^{\mathbf{T}} + \mathbf{F}, \tag{4}$$

where **T** are the scores, **P** and **Q** are the loadings and **E** and **F** are the residuals. **Y**-variables are predicted on the basis of a well-chosen set of relevant **X**-variables with explanatory or predictive purpose (Esbensen et al., 2000). The **Y**-matrix is therefore usually the property to be calibrated for (the response data), while the **X**-matrix (the descriptor data) is defined as the output of the instrument (Martens and Martens, 2001).

In PCA-calibration one set of loadings from the **X**-matrix (**P**-loadings, cf. equation (3)) is obtained, whereas PLS-calibration also includes the usage of loading weights (termed **W**-loadings). The **P**-loadings express the relationship between the raw data matrix **X** and its scores (**T**), whereas the **W**-loadings represent the effective loadings between **X** and **Y**. The differences between **P**- and **W**-loadings describe the influence of the **Y**-guidance on the decomposition of **X**. The loadings from the **Y**-matrix (**Q**-loadings, cf. equation (4)) are the regression coefficients from the **Y**-variables onto the scores (**U**). Together with the **Q**-loadings, the **W**-loadings are used to interpret the relationships between **X**- and **Y**-variables as well as interpreting the scores related to these loadings (Esbensen et al., 2000).

As an extension to PLS, Nørgaard et al. (2000) have developed iPLS. The purpose of iPLS is to divide the variables into subintervals of equal width in order to carry out local PLS on the subintervals, and thereby detect specific variables of interest. In this way one can get as large score vectors as possible in order to obtain more stable predictions (Höskuldsson, 2001). PLS is then carried out on each of the subintervals (local models) and the full-spectrum (global model). The procedure is a stepwise calculation with the aim of zooming into local models of interest, based on modeling performance between local models and the global model. The selection of intervals by iPLS is furthermore of importance in PCA in order to remove noise from the **X** data matrix.

1.2. Proteomics: how can data handling be improved?

The purpose of proteomics is to find ways of focusing on those proteins that are involved in a particular biological function of interest (Godovac-Zimmermann and Brown, 2001). The field of proteomics mainly consists of the following two stages (Rabilloud and Humphery-Smith, 2000): (1) separation of the proteins to be analyzed and (2) characterization of the separated proteins.

Examinations of proteomic maps have revealed more additional proteins than was expected when compared with the corresponding genomic maps (Corthals et al., 2000). It has therefore been suggested that the additional proteins found via proteome analysis are modified proteins, which could not be accounted for by genome analysis.

For every gene expressed in a cell at a given time, three times as many cellular proteins must be expected as a result of mRNA splicing and posttranslational modifications (Naaby-Hansen et al., 2001). Posttranslational modifications, which include simple proteolytic cleavage as well as covalent modification of specific amino acid residues, like, e.g., glycosylation, phosphorylation and acylation, are not detectable by analysis of RNA (Hille et al., 2001).

Although time-consuming, 2DGE is the favored separation technique in proteomics by virtue of the extremely high resolution obtained (Küster et al., 2001). While separation constitutes the first half of proteomics, characterization constitutes the other half. Proteins of interest, separated by 2DGE and electroblotted, may be submitted to N-terminal sequencing and succeeding database cataloguing in order to determine their identity. The advantage of N-terminal sequencing is the ability to directly sequence the N-terminus of electroblotted proteins without any need of specific preparation procedures (Kinter and Sherman, 2000). Characterization of proteins may also be carried out by application of MS. Matrix assisted laser desorption/ionisation time of flight (MALDI-TOF) MS is the most common type of MS combined with 2DGE (Hille et al., 2001). After 2DGE proteins are concentrated in individual spots as SDS–protein complexes within the polyacrylamide gel matrix. The protein spot must be pre-treated prior to analysis by mass spectrometry (Eckerskorn and Strupat, 2000). If the protein spot is excised from a dry gel, it must first be washed out and then cleaved by site-specific proteolysis with, e.g., trypsin (Naaby-Hansen et al., 2001). The resulting peptide fragments of a given protein spot can thereby be viewed as the third dimension separation, being independent of the two separation modes from 2DGE (Hanash, 2000). The peptide masses obtained from the mass spectrum are subjected to database cataloguing, where theoretically digestions of proteins are available, either in protein databases or in translated genomic databases (Naaby-Hansen et al., 2001).

A general problem in proteomics when starting with 2DGE is the characterization procedure. Approximately one day per gel is needed for analysis. When also including complete structural analysis by MS, about one month is required per gel (Hille et al., 2001). Identification of proteins by combination of MALDI TOF-MS and 2DGE, including the processes of image analysis, spot detection and enzymatic digestion prior to MS, is not possible to carry out automatically (Hille et al., 2001). A compromise between sample throughput and analysis time is therefore an important issue in proteomics of today. Since complete automation is not possible, the field of proteomics must be extended to include methods that can ease the evaluation of results obtained from 2DGE. Focus is turned to optimization of 2DGE and MS, and particularly on how to optimize the combination between these techniques.

Although the concept of proteomics is to encircle the identity of certain proteins of interest, the overall relation between proteins must also be explained. Since large amounts of data are collected, the overview may easily be lost. If the overview is lost, how can the conditions for proteomics then be fully obtained? The field of proteomics should therefore be extended to also include proper handling of large datasets. Image analysis of 2D gels is the basis for characterization of proteins. If the steps in image analysis could be speeded up, and at the same time be improved, it would ease the proteomic procedure essentially. Chemometrics may be the answer to a faster and more reliable analysis in 2DGE.

1.2.1. 2DGE

2DGE separates proteins according to two independent physical and chemical properties. Thousands of different proteins can thus be separated, and information such as the protein pI, the apparent molecular weight, and the amount of each protein is obtained. As the need for high throughput methods in proteomics increases, focus has shifted towards automation (Lopez, 2000; Patterson, 2000), by this new focus the bottleneck has moved from the protocol itself to the gel analysis (Lopez, 2000; Smilansky, 2001). The 2D protocol is still time consuming, however, it is important to notice that the subsequent gel analysis is just as time consuming. To automate gel analysis, several software programs have been developed (Appel et al., 1997; Lopez, 2000; Mahon and Dupree, 2001; Raman et al., 2002; Smilansky, 2001). Gel analysis involves three steps: (1) spot detection, (2) spot/gel alignment and (3) identification of interesting spots (Lopez, 2000). Development of efficient and reliable algorithms to perform the two first steps has been subject to much work (Gustafsson et al., 2002; Kaczmarek et al., 2002; Kriegel et al., 2000; Pleissner et al., 1999; Veeser et al., 2001), which can be seen in the new generation of 2DGE software (Raman et al., 2002). With the continued progress in development of 2DGE analytical software, the full potential of the 2DGE is anticipated in the near future.

The use of multivariate methods in the analysis of 2DGE is an emerging application (Appel et al., 1988; Jessen et al., 2002; Pun et al., 1988; Rabilloud et al., 1985; Tarroux, 1983; Tarroux et al., 1987; Vohradsky, 1997). By mathematical modeling of the data contained in 2D gels, it is possible to make fast extraction of data from gels. Traditionally the use of spot volume data has been applied as this makes direct use of the spot lists generated by most 2D analysis software packages. Alternatively it is possible to use the presence of a spot as indicator, so that the dataset is a binary matrix, where 1 shows that a spot is present in a gel and 0 that it is not (Radzikowski et al., 2002). This makes the classification of gel images based on expression patterns for protein spots possible. Moreover it is possible to deduce biological information from the loading plots, i.e., which spots contribute to the differentiation of the gels. Using an image-analytical approach, it is possible to do much of the same work in an automated and fast process that does not involve the subjective assessments of an operator.

However, although the developments in the 2DGE protocol and instrumentation have greatly improved the reliability and reproducibility of 2DGE, much focus is still on whether the method will turn into what everybody hopes; a fast, reliable method for high throughput proteomic research.

1.2.2. Speed/Automation

The speed and degree of automation are two areas where the 2DGE protocol can be improved. It has been estimated that it can take as much as one month to fully analyze one gel (Hille et al., 2001) with the current degree of automation. Improvements in the degree of automation, however it is estimated, can bring this down to 3–7 days (Hille et al., 2001). In addition, the gel analysis involves subjective assessments by the analyzer, which can make the analysis operator dependent.

Automated units to perform the first and second dimension runs as well as visualization combined with the development of IPG have brought down the number of process steps and operator dependent variables (Görg et al., 2000). Prototypes of fully automated robots for spot identification, excision and analysis with MALDI TOF-MS have likewise been described (Harry et al., 2000; Nordhoff et al., 2001). It is therefore in the gel analysis that we find the bottleneck in large-scale proteomics today.

1.2.3. Problematic proteins

Generally, two groups of proteins have been a problem in 2DGE separations – very basic and/or insoluble membrane proteins. The problem with the basic proteins has been the lack of commercial products to

create a pH gradient above 10 and that basic proteins have been difficult to focus in the first dimension because of reverse electro-endosmotic flow. With IPG it is possible to make pH gradients up to 12 and at the same time use a standard protocol (Görg et al., 2000). Narrow pH gradients up to pH 12 require changes to the protocol to minimize the transportation of water from the cathode to the anode. Gradients with pH 10–12 and pH 9–12 have successfully been applied (Görg et al., 2000).

Hydrophobic membrane proteins are very challenging to handle and it is estimated that only about 1% of the membrane proteins are separated in 2DGE with a standard protocol (Fey and Larsen, 2001). The development of sample solubilization has improved the possibility of solubilizing these proteins by use of zwitterionic- and organic-detergents. But it is still an area that needs to be developed (Anderson et al., 2000).

1.2.4. Visualization of proteins

Traditionally proteins are stained with Coomassie brilliant blue (CBB) or silver. CBB staining has a relatively low sensitivity but is compatible with Western blotting and subsequent protein sequencing. Using CBB the spot intensities correlate linearly with protein amount. Silver staining can detect as little as 0.1 ng protein and is thereby much more sensitive than CBB staining but has disadvantages: (1) Silver staining does have a lower reproducibility between replicates; (2) It does only stain quantitatively in a narrow range, which means that silver staining is not useful to study differences in protein expression between different stages; (3) Some proteins are only stained weakly or not at all (Görg et al., 2000).

As an alternative to silver staining, fluorescence staining can be used. Fluorescence staining is less labour intensive and has a detection limit of 1–4 ng protein (Steinberg et al., 2000), which is better than CBB and at the same level as some silver staining procedures. The staining is linearly in a wide interval. The protocol is simple and can therefore more easily be used in an automated system.

Several alternatives to the classic methods of visualizing have been reported. In (Bienvenut et al., 1999; Binz et al., 1999) a molecular scanner is described. In this system all proteins in the gel are digested followed by transfer to a polyvinylidene difluoride (PVDF) membrane with a matrix solution compatible with MALDI TOF-MS. The PVDF membrane is then scanned directly with the MALDI TOF-MS instrument with a resolution of 0.4 mm. The spectrum obtained in each point is automatically submitted to a protein sequence database for identification. In this way a complete map of identified proteins is created. However, the method is not quantitative and the scanning of the gels with the spectrometer is very time consuming. Thus it takes about 36 days to scan one gel measuring 16×16 cm^2, generating approximately 40 GB of data. However, the authors estimate that it will be possible to bring this down to a matter of hours.

The method eliminates some of the classic flaws in 2DGE such as matching of protein spots, sensitivity, identification and to some extent co-migration. As all proteins in principle are identified, the need for matching of protein spots is eliminated. The method is therefore an interesting alternative to the classic 2DGE analysis, in that several steps are combined to one automated process.

In (Walker et al., 2001) a method is described where the second dimension is replaced by MALDI TOF-MS, thus creating a so-called virtual gel. The first dimension is run in the traditional manner, but the IPG strip is prepared for MALDI TOF-MS. By scanning the IPG with the MALDI TOF-MS spectrometer the virtual gel is created where the second dimension is constructed by MS spectra. Thus the method also eliminates some of the classic 2DGE problems.

1.3. Near infrared spectroscopy

Near infrared (NIR) spectroscopy provides a method for rapid, non-destructive and accurate analysis of the composition of a sample. It allows discrimination of various organic compounds and can be used both to acquire qualitative and quantitative information. It not only supplies chemical information, but also information of whether the physical properties of a sample can be obtained.

NIR has been widely used in the field of agriculture, and one of its first applications was the determination of moisture in agricultural products (Pasquini, 2003). Now it is also used in various other fields such as food and medicine, and it is an increasingly accepted tool for academic research and industrial quality control in many areas ranging from chemistry to agriculture and from life science to environmental analysis (Foley et al., 1998; Siesler, 2002). A merit of NIR is the simultaneous determination of multiple constituents in a sample, which also allows for estimation of complex attributes such as the susceptibility of plants to insect attack. NIR is not used for very sensitive analysis since the detection limit in general is only about 0.1% (w/w) for most constituents (Iwamoto and Kawano, 1992).

Infrared (IR) is the part of the electromagnetic spectrum that covers the wavelength region from 0.7 to 200 μm. The region of IR, which is nearest to the region of visible light, is called the near infrared (NIR) region, and it includes the wavelength range from about 780 to 2500 nm. The mid infrared (MIR) spans the higher wavelength range from 2500 to 15,000 nm (Davies, 1993).

The electromagnetic radiation can interact with matter to give rise to an absorption spectrum. In

vibrational spectroscopy, which employs the MIR and NIR regions, the absorption bands originate predominantly from radiation energy transferred to mechanical energy associated with the vibration of atoms.

In a molecule, atoms or groups of atoms participating in chemical bonds are displacing one in relation to the other in a frequency that is defined by the type of bond of vibration (Davies, 1993). Absorption of infrared radiation induces the transition between vibrational energy levels, and the frequency and amount of the absorbed radiation gives information about the types and number of bonds between atoms or functional groups in the molecules. Consequently, the absorption spectrum reflects the chemical composition of the material being analyzed, and gives information on the amount of protein, fat, starch or any other organic molecule in a sample. However, NIR is a secondary method requiring calibration against a reference method for the constituent, because of influence also from physical properties (Osborne et al., 1993).

While the MIR-region possesses the energy that is necessary to promote molecules from their lowest excited vibrational states, the NIR region is of higher energy, and the absorptions originate from overtones or combinations of the fundamental absorptions seen in the MIR region.

IR spectroscopy that uses the MIR-region has been a well-established tool for elucidation of structure, because the peaks are relatively distinct and can be attributed to the presence of certain functional groups (Siesler, 2002). In the NIR region, however, direct interpretation of the spectral absorbances is very difficult for complex mixtures because of broad overlapping absorption bands. NIR thus relies on multivariate methods to quantify the properties or constituents of interest.

One of the advantages of NIR over IR is that NIR requires a minimum of sample preparation and provides the possibility for analysis on, e.g., intact fruit and also opaque samples. When a beam of IR radiation containing different frequencies is directed on to a molecule, an absorption spectrum (plot of energy versus wavelength) is produced, because only the radiation of frequencies capable of supplying exactly the energies between allowed transitions is absorbed. Each kind of molecule has a characteristic spectrum depending on the number and types of bonds, since the transition energies are defined by the vibrational frequencies of the different bonds.

Spectra of polyatomic molecules show absorptions from the distinct chemical groups, which vibrate at their characteristic group vibrations. The characteristic vibrations are relatively constant in their frequencies from molecule to molecule, but some adjustment takes place due to influence from different molecular environments and molecular interactions (e.g., degree of hydration) which influence the force constant (Bokobza, 2002). It is therefore possible for example to differentiate C–H stretching stemming from, e.g., alkanes, methanol and ethanoic acid (Osborne et al., 1993).

For polyatomic molecules interbond coupling can occur between stretching and bending vibrations of the same functional group, meaning that their vibrational energies are dependent on each other. This complicates the spectrum, but also causes some distinct vibrations for complex molecules. Proteins, for example, show characteristic absorption bands in the IR and NIR due to the vibrational modes: C=O stretching coupled to N–H bending and C–N stretching (amide I), and N–H bending coupled to C–N stretching (amide II) (Osborne et al., 1993). In the NIR some combination bands involving these modes (and, e.g., N–H stretching) appear, and such a band has been found very useful for estimation of protein concentration. Many bands in the NIR spectrum of protein are sensitive to changes in secondary structure and degree of hydration, and therefore can be used, e.g., for monitoring the denaturation of a protein (Wu et al., 2000). This is the consequence of NH-bands being displaced by hydrogen bonding like any other X–H-band. Hydrogen bonding changes the force constant of the covalent X–H bond thus causing a small shift in the wavelength at which the absorption band appears. This sensitivity of NIR to hydrogen bonding is the reason why NIR also can be used for studying the state of water in foods. The O–H absorption band, however, becomes very broad due to the hydrogen bonding.

2. Results and discussion

2.1. Proteomics 'classic'

The traditional way of doing proteomics is outlined in Fig. 1. Gels are evaluated using image-processing software; interesting spots are pointed out and identified. The visual image of protein spots is invaluable in proteome analysis as far as characterization of single proteins is concerned, which is why 2DGE is the favoured separation technique in proteomics. However, the characterization of proteins from 2D gels often requires many 2D images being compared to each other. However, when adding just a few more gels to the analysis it is almost impossible to maintain an overall view of the data. Image Master® and other software programmes like, e.g., CAROL (Kriegel et al., 2000), Z3 (Smilansky, 2001), PDQuest, Melanie and Progenesis have been developed, making attempts on easing the 2D image analysis. However, the real breakthrough will only appear when a full-automated analysis of 2D images is possible. An unquestionable obstacle towards the full-automated analysis of 2D gels is the problem of gel alignment.

2.2. Explorative data analysis

In order to increase the effectiveness of proper spot selection, the data from 2DGE can be subjected to multivariate analysis in order to point out which combination of spots could be valuable to sequence. This way a lot of time and effort can be saved when only the proper spots are identified. Although the images from 2DGE are obvious subjects to multivariate data analysis by virtue of the many variables they create, there still are some obstacles to pass before it is practically possible. The 2D gel patterns are exposed to geometrical distortions, locally as well as globally, with decisive impact on the grade of reproducibility. In order to analyze 2D images properly by multivariate analysis, they must first of all be aligned.

There are two ways to go. Multivariate data analysis can be used on either the spot list produced by the image processing software after alignment or directly on the aligned images. The first procedure has been used in some studies (Jessen et al., 2002, Radzikowski et al., 2002). In the latter study concerning rye proteins it was shown that the results from the different analysis could then be combined and analyzed by PCA to give an improved characterization of the varieties. The PCA of the 2D spot data was able to group the spots according to the varieties in which they were present and this improved the evaluation of the 2D gels. The resulting data from PCA can also be used to create a dendrogram of the investigated varieties. The PCA of the 2D spot data in combination with the functional properties data showed a similar grouping of the varieties and that there was one spot that was close to the properties, bread volume and bread height. The PCA of the 2D spot data can be useful in any 2D electrophoretic analysis where the aim is to find protein spots that are characteristic for a given sample or find protein spots that are present in a selected group of the investigated samples. The PCA of 2D spot data reduces the time spent on analysis of the results obtained from image analysis of 2D gels, and it also makes it easier to analyze a large number of gels. Another advantage is that it is possible to combine results from many different experiments and analyze them together.

2.3. Spot detection of 2DGE gels

Here, we present an analysis of 2DGE patterns of the storage proteins from ten different wheat varieties by PCA and PLSR. An analysis of the volume spot lists showed that the selected wheat varieties were represented in two groups. To avoid the generation of spot volume lists, i.e., to avoid spot detection, we used a method in which the gels were analyzed as images to test if the gels could be differentiated. The latter approach gave the same classification of the ten varieties as the use of spot volume lists, although without the prior work of spot detection and spot matching which is both time consuming and subjective. For further screening purposes the use of this approach in the initial screening of a large number of gels is therefore a promising alternative to the usual spot detection and matching.

Multivariate analysis is implemented in recent versions of popular 2DGE analysis software packages. The implementation, however, is solely based on the subsequent analysis of spot list data. The present method is based on sampling of real-spot data as basis for the detection.

The algorithm: Based on a data matrix of unfolded spot images we have used a singular value decomposition (SVD) to build a PC Model and used this model to create virtual gels of probability to indicate where the spots are located. This approach is described in Fig. 3. A more detailed description of the algorithm is found in Appendix A.

The algorithm has been used to identify spots on 2DGE gels of wheat storage proteins (Schultz et al., 2004) for 2DGE procedure. In Fig. 4, the spots used to construct the Peak matrix are marked.

Spot identification was done on three different gels of the wheat varieties: Pentium, Hussar and Trintella. All gels were sub-images extracted from whole gels and the background has been subtracted and the intensities adjusted (Fig. 5).

The results from the identifications are shown in Fig. 6. It is clear that almost all spots have been identified. However, there is a tendency towards missing identification of the weakest spots as well as some symmetric noise around each spots. Moreover, it is seen that the algorithm identifies the gravity point in the spot. This is in accordance with what is expected as the spots were sampled from the centre of gravity. Moreover, it should be noted that the intensities of the identified spots reflect the degree to which a specified pixel fits the reference model and not the original spot intensity. On the gel of the variety Hussar an area is seen where the number of spots are difficult to identify on the original gel. In this area three spots have been identified, which are in accordance with the actual gel. The algorithm has also been tested with other spots as reference data, which differ in size and form (data not shown). These tests show that the shape of the spot has less influence on the performance of the algorithm than the size (Fig. 7).

We have here demonstrated an alternative approach to 2DGE spot detection as well as shown how a multivariate approach can be used for other purposes than analysis of spectroscopic spectres. We believe that the algorithm as presented here can contribute to further development of powerful 2DGE analytic software packages, further fuelling the widespread use of the 2DGE technology in modern proteomics research.

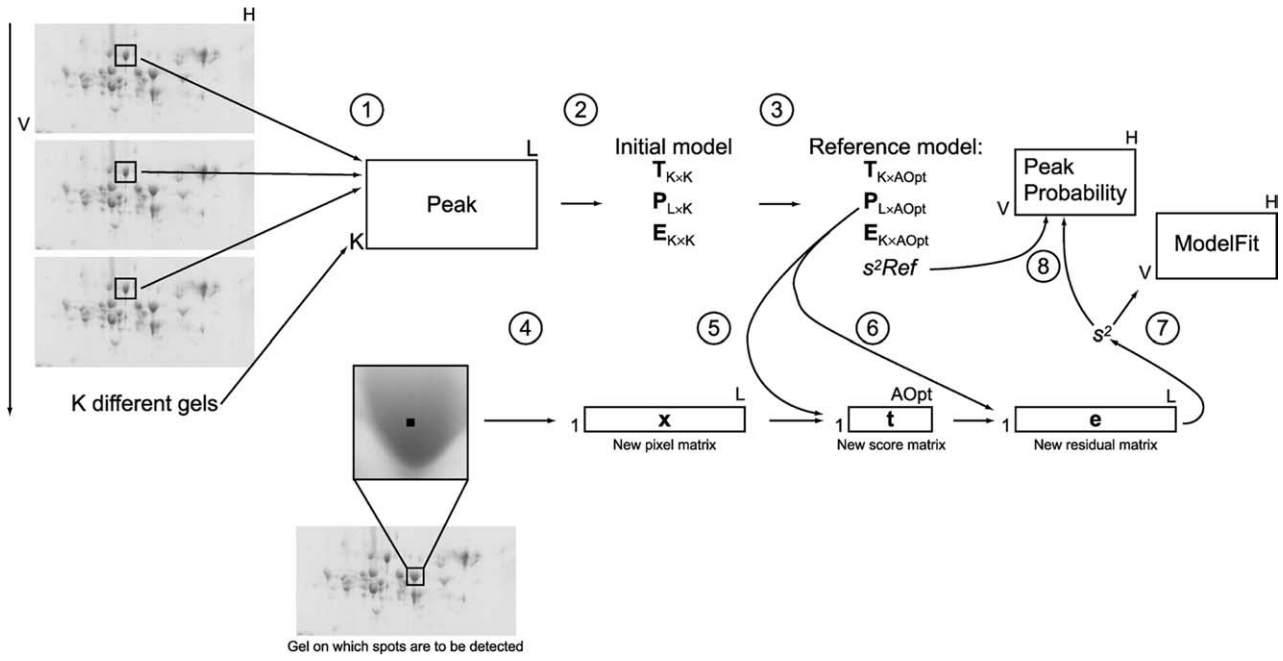


Fig. 3. Diagram of the FindPeak algorithm. (1) The spots are sampled, and the Peak matrix is constructed of the unfolded spot images. Spots are sampled from K different gels. One or more spots from each gel can be sampled. (2) The initial PC model is calculated, consisting of the scores matrix (**T**), the loadings matrix (**P**) and the residual matrix (**E**). (3) The optimal number of principal components is determined (AOpt) and the reference PC model is constructed based on this number of principal components. From the residual matrix the reference variance (s^2Ref) can be calculated. (4) In each pixel in the gel image to be analyzed, a sub-image with centre in the pixel is extracted and is unfolded to **x**. (5) The new score matrix, **t,** for the pixel is calculated from **x** and **P**. (6) The new residual matrix, **e**, for the pixel is calculated from the new score matrix **t** and **P**. (7) From the new residual matrix, the pixel variance, s^2 is calculated. The numeric value of this describes how well the pixel sub-image fits the model and forms the ModelFit matrix. (8) The relationship between s^2 and s^2Ref makes up the PeakProbability matrix.

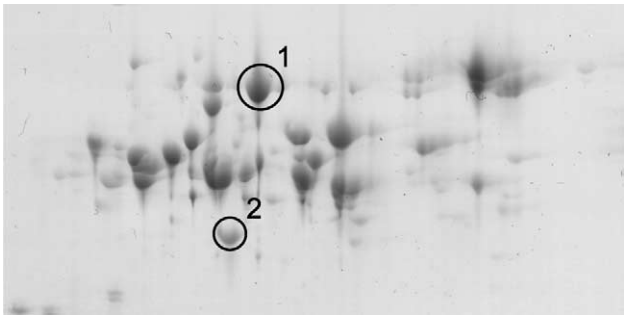


Fig. 4. DGE gel marked with the spots used to construct the Peak matrix. Spots 1 and 2 were sampled from 39 different gels of different wheat varieties.

2.4. Mass spectrometry and multivariate data analysis

MS of whole protein extracts together with chemometrics can be used to classify complex mixtures of proteins. We have used this chemometric approach as a supplement to the proteome analysis of the alcohol-soluble proteins (gliadins) from the wheat gluten complex (Gottlieb et al., 2002). Based on classic proteome analysis with 2DGE, a specific gliadin was found to only be present in wheat varieties unqualified for bread-making. By means of N-terminal sequencing, the identity of the protein was then encircled. Gliadin-data obtained from MALDI-TOF MS ranging 31 kDa were subject to multivariate analysis. By means of multivariate analysis on the MS data narrow molecular weight intervals of interest, with sizes of only few hundreds of Da, were repeatedly detected (Fig. 7). The study revealed that application of multivariate analysis could detect the molecular weight area in which the gliadin of interest was found by the classic labour-intensive proteome analysis. From the study it was concluded that the use of multivariate analysis on data output from separation of gliadins is a strong tool that can contribute substantially to the field of proteomics.

2.5. NIR spectroscopy and multivariate analysis

Quantitative information of a specific analyte is reflected in the intensity at the wavelength at which it absorbs according to Beers law, stating that absorbance presents a linear behavior with the concentration of the analyte for a fixed path length (Heise and Winzen, 2002). The NIR spectra are however often rather featureless, which prevents identification of bands for the analyte of interest. Use of a single wavelength will seldom provide a good model because of the occurrence of overlapping absorption bands and deviations from Beers law. Deviations from Beers law occur, e.g., at high

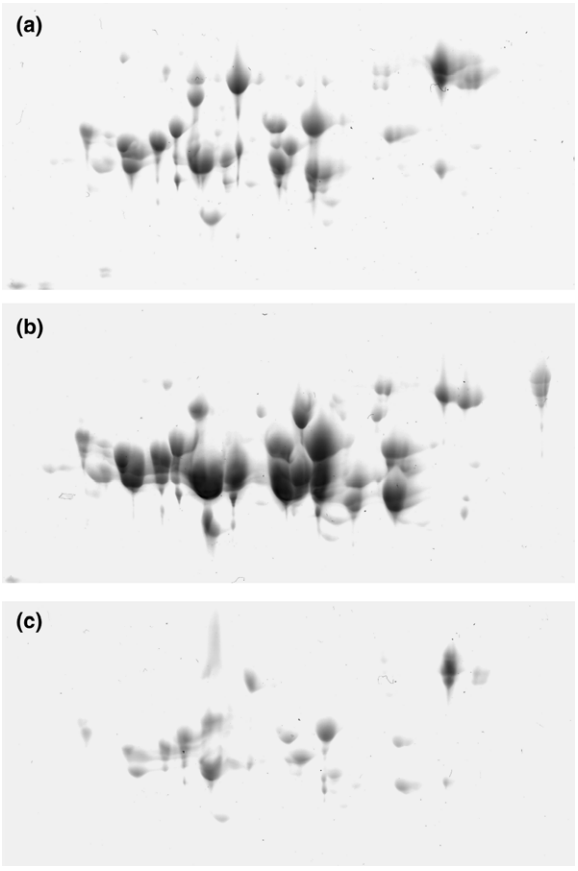


Fig. 5. Gels on which spots have been identified. (a) Pentium; (b) Hussar and (c) Trintella.

analyte concentrations, because of light scattering phenomena in solid samples, and when there are changes in the hydrogen bonding pattern such as temperature or relative concentration changes (Pasquini, 2003). In particulate samples light scattering phenomena further introduce nonlinearities (Martens et al., 2003).

Extraction of quantitative information therefore relies on multivariate models. Multiple linear regression (MLR) that uses only a few wavelengths is a usual method for regression of the reference data on the spectral data. However, PCR and PLSR that can separate out the relevant and reliable covariation patterns from the background noise in the full wavelength range are increasingly being used (Heise and Winzen, 2002). PCA provides a quick overview of the spectral data and reveals clusters and trends which could otherwise be hard to see (Fig. 8).

2.5.1. Barley mutants

NIR spectroscopy has shown promising results for application in plant biotechnology for gaining insight into the phenotypes that result from perturbation of the gene expression by genetic and environmental changes (Jacobsen et al., 2004; Munck et al., 2001). Processing of the NIR spectra by classification techniques yields a metabolic fingerprint of the organisms without differentiation of the individual metabolites. In this way NIR has potential as a fast screening method revealing organisms with altered phenome, but it is also possible to obtain quantitative information of specific metabolites

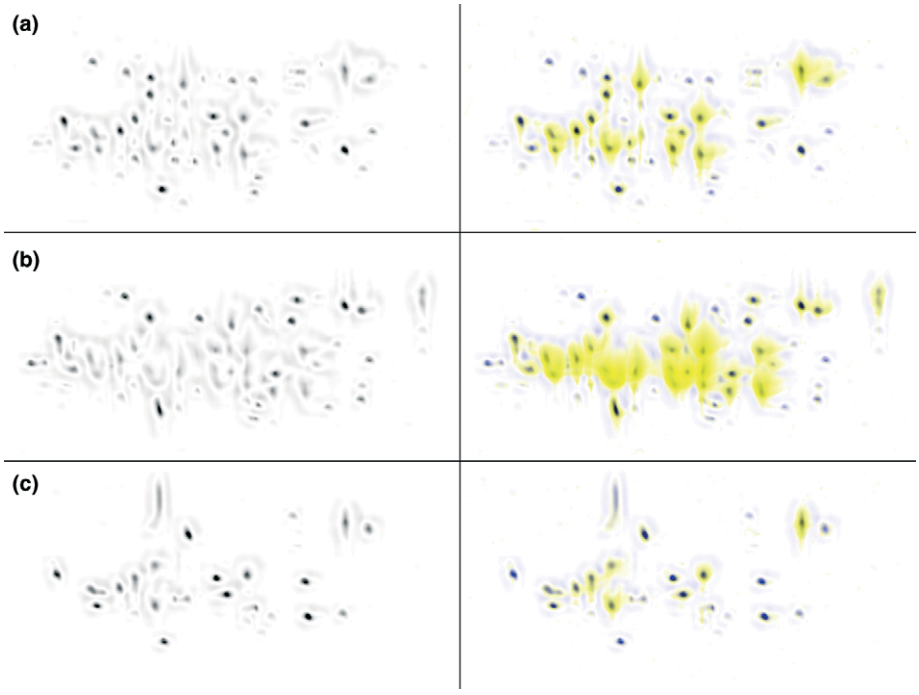


Fig. 6. Left column shows the PeakProbability matrix. Right column shows a composite image of the PeakProbability matrix and the original gel. (a) Pentium; (b) Hussar; (c) Trintella.

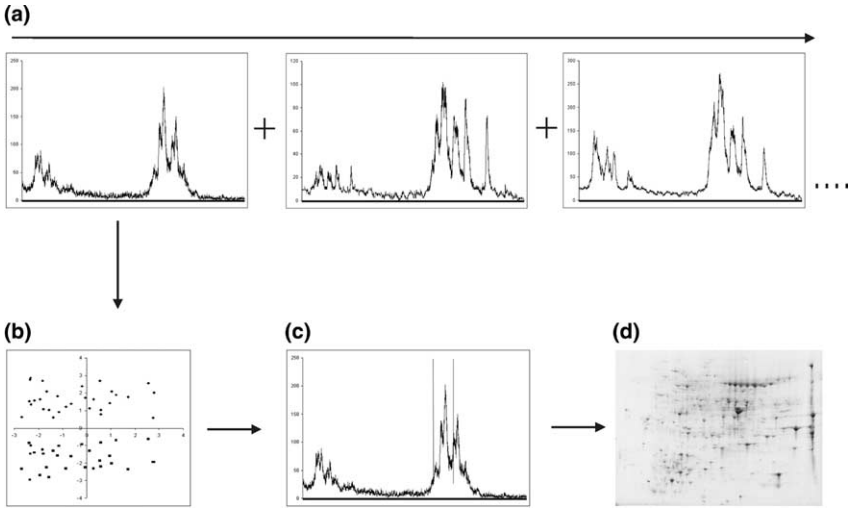


Fig. 7. Multivariate workflow combined with proteomics. (a) Mass spectra are collected, in this case of the alcohol-soluble fraction from wheat varieties. (b) By using PCA to analyse the collected spectra it is possible to compare the objects to each other in a score-plot (each spectrum is represented by a spot). (c) Variables of interest can be detected by iPLS. (d) The high-resolution obtained by 2DGE is then used to isolate the few proteins in the detected interval of interest. Further analysis is done by enzymatic digestion of the selected proteins and final identification by database searching.

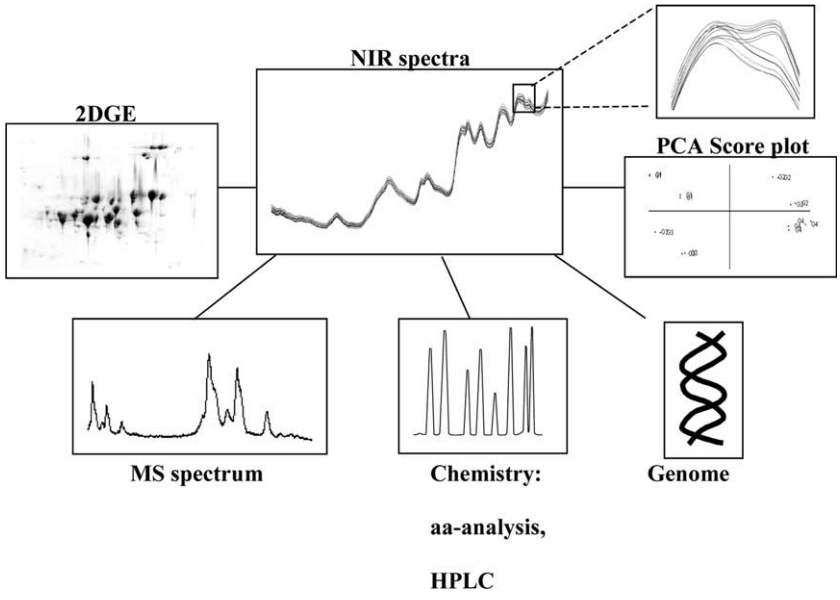


Fig. 8. NIR as an exploratory tool. NIR spectra contain 'hidden' information about the sample compositions, but groupings and trends in samples can easily be surveyed in a PCA based on the spectral data. The basis for the groupings is examined by use of various analytical methods, and gene sequencing reveals the underlying genome.

for use in metabolomics. Accurate and reproducible quantitative methods are necessary to differentiate samples where the result of the changed gene expressions is only quantitative changes in the metabolite concentrations (Sumner et al., 2003).

It has been shown by Munck et al., 2001 that NIR of barley flour provides a spectral fingerprint of the barley endosperm phenome, which can be used for discrimination of normal barley and high-lysine mutants. In the near isogenic background of the advanced barley lines, the effects of the high-lysine genes and also different growth environments were easily detected by NIR. A clear discrimination was seen in a PCA using the wavelength region 400–2500 nm. Also the ability of NIR to discriminate different high-lysine mutant genotypes has been demonstrated. In a work by Jacobsen et al., 2004, even the phenotypic effects of different alleles in the same locus were differentiated in a PCA, where a more extreme mutant was shown to form a distinct cluster.

Comparison of the mean spectra from the PCA clusters lead to identification of the spectroscopic signatures that discriminated the mutant genotypes. A

small region (2280–2360 nm) in the spectra was identified as basis for visual discrimination of mutants and also their differentiation from normal barley. Observed absorption bands in this range were in the literature assigned to protein side chains (amino acid-determinant), cellulose and unsaturated fat. The effect of the different high-lysine mutant genotypes on the amino acid composition were thus reflected in the spectral shape, but also the effects of the altered proteome on other constituents such as starch, fat and fibre were evident in the spectra. These pleiotropic effects, which are often forgotten in analysis, are revealed by NIR, and the perspective is that mutants and transformants can be identified from their pleiotropic effects.

The ability of NIR to measure on the intact plant thus provides a holistic fingerprint of the metabolic status in contrast to other chemical methods applying to plant extracts and thus being biased towards specific chemicals. The advantage of using spectral information about the total endosperm composition for classification of unknown barley lines was demonstrated. A barley line that was formerly considered a waxy line due to its low amylose content was recognized as another mutant, since its spectra grouped together with *lys5* mutants in the PCA model. The *lys5* mutants were low in starch but compensated for that by high β-glucan content and thus differed from the classical waxy lines.

The use of the combination of NIR and PCA is a totally exploratory approach. After identification of clusters or outliers in a PCA, the proteome and metabolome can be further investigated by more selective methods like 2DGE, MS, amino acid analysis and other chemical analyses. In this respect, knowledge of the wavelengths at which the different constituents absorb can be of great help for targeting the chemical analysis. Genome analysis reveals the functional relationship between the genomes and the metabolomes. The fingerprinting approach allows for generating new hypotheses about the gene functions and is more objective than the traditional procedure in functional genomics, where only test of the logical response to a perturbation is made (Gidman et al., 2003).

2.5.2. Wheat quality

NIR has long been a recognized method for accurate prediction of the protein content of wheat for assessment of its breadmaking potential (Morris and Rose, 1996). The baking quality of flour, however, relates to both the amount and quality of the gluten proteins and is also determined by the complex interactions of all the biochemical constituents in flour (Veraverbeke and Delcour, 2002). Providing a measure of all the primary constituents simultaneously, NIR should have potential for determination of this quality. Various biochemical and physical properties of dough, relating to the baking quality, have been reasonably estimated by NIR, but a strong correlation between the measured property and the total protein content can lead to wrong conclusions. NIR is, however, sensitive not only to protein content but also to protein quality to some degree (Wesley et al., 2001). The quality of gluten protein is partially determined by the glutenin to gliadin ratio and the weight distribution of glutenins (Wesley et al., 2001). NIR is generally not very sensitive to individual levels of different proteins, but is has been found anyway that the individual contents of gliadin and glutenin can be estimated to some degree from NIR spectra not only because of their correlation to total protein content (Wesley et al., 2001).

2.5.3. Characterization of gluten

Dried gluten is used in the baking industry for improving the bread-making performance of wheat flour. To assure satisfactory performance of the gluten, assesment of both composition and functional end-use properties is required (Czuchajowska and Pomeranz, 1991). An important quality parameter of gluten is the moisture content, since high moisture content (above 10%) promotes deterioration of gluten quality. Other quality parameters include protein, free lipid and ash content as well as particle size and various rheological properties (Czuchajowska and Pomeranz, 1991). NIR has been found useful for determination of all these parameters, though composition was much better predicted than the physical and rheological properties (Czuchajowska and Pomeranz, 1991).

It is more desirable to know how well gluten performs in bread-making than just knowing its individual quality parameters. An accepted test for gluten functionality is measurement of the increase in volume of bread baked from flour fortified by gluten, and NIR has been tested for its ability to predict this end-use property. In the work by Czuchajowska and Pomeranz, 1991, it was found that a calibration model based on three wavelengths had limited power for predicting increase in loaf volume, but it was, however, found that some rheological properties of hydrated gluten (which correlates to its end use properties) could be well predicted by three-wavelengths MLR models.

2.5.4. Experiment: moisture in gluten

In an experiment FT-NIR reflectance spectra were measured on samples of freeze-dried gluten powder, which contained different amounts of moisture. In the beginning of the experiment spectra were measured on the dried gluten powders, and then samples were left to absorb moisture in a moist chamber at room temperature (26 °C). Spectra were again recorded after 2, 4 and 24 h on the same samples (Fig. 9). The final water content after 24 h was around 18% on a wet basis.

Broad water bands appear at around 1450 and 1930 nm for the moist samles and the intensity of the bands increases as the water content becomes higher. At the

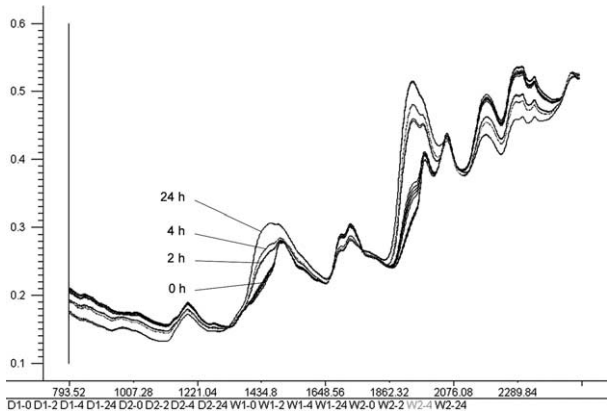


Fig. 9. NIR spectra measured on samples of gluten powder with different water contents.

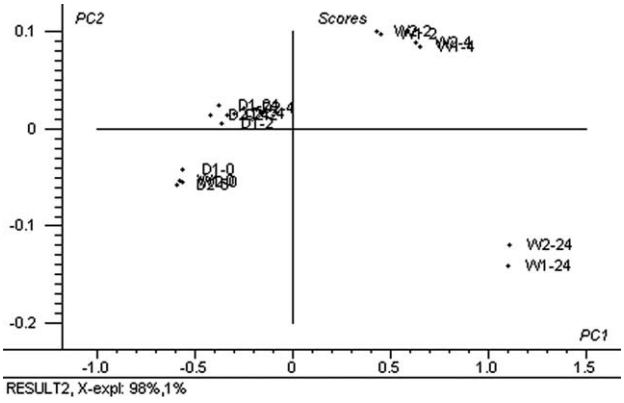


Fig. 10. Scoreplot from a PCA on the NIR spectra seen in Fig. 9. D, dry gluten samples. W, moist gluten samples. Last number refers to hours of hydration.

same time protein bands decrease in intensity. The spectra have been pretreated by multiple scatter correction (MSC) in order to remove undesired spectral variation caused by light scattering. Light scattering results from physical phenomena in the samples like particle shape, size, size distribution and sample packing and it introduces both multiplicative and additive effects in the spectra (Martens et al., 2003). To demonstrate how PCA can provide an overview of the variance and groupings in data a score plot from PCA on the MSC corrected full spectrum is shown in Fig. 10.

PC1 which describes 99% of the spectral variance reflects the increase in water content. PC2 shows another smaller phenomenon taking place.

To get more insight into the spectral changes that take place when water increases and interacts with the gluten proteins, the second derivative was taken. Second derivative spectra are shown in Fig. 11. Taking the second derivative of the spectra facilitates the visual inspection of the spectra since peaks in the original spectrum appear as more clearly separated downward peaks in the second derivative spectrum and at the same time multiplicative and additive effects are removed. The signal to noise ratio is, however, decreased.

The spectral changes upon hydration can originate from the changed concentrations, the changed hydration of the protein, changes in protein secondary structure upon hydration or from changes in protein side chains. Also changes in other minor components cannot be excluded.

A zoom picture of a small wavelength region with interesting spectral changes is seen in Fig. 12. It seems that two peaks that are close to each other change shape

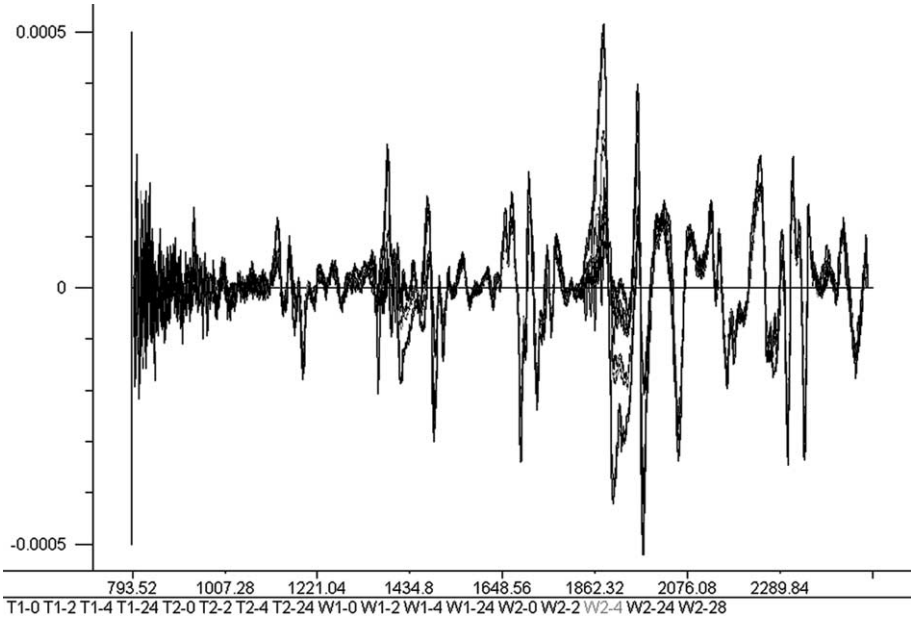


Fig. 11. Second derivative spectra. Original NIR spectra are seen in Fig. 9 and are measured on samples of gluten powder with different water contents.

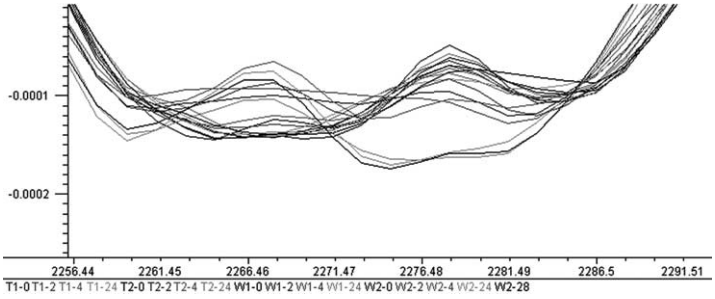


Fig. 12. Zoom picture of second derivative spectra in Fig. 11. Dry and moist gluten samples show distinct spectral features.

and shift wavelength positions to lower wavelengths when the samples become moist.

Changes in protein secondary structure are possibly reflected in the amid combination bands, which are sensitive to the degree of hydration, but also protein side chain vibrations can change upon secondary structure changes (Sefara et al., 1997). The interesting region encloses the wavelength range 2255–2290 nm, and in this area combinations of CH2 stretch and H–C–H bending vibrations from protein side chains and carbohydrates are found (starch content is low). It has been found that increased hydration of some gluten proteins leads to increased β-sheet content and decreased unordered structure, and at a certain point the β-turn content is increased (Belton et al., 1995). The analysis methods that have been used for elucidation of the changes in secondary structure that arise upon hydration, are yet needed for relating these structural events to the changes seen in the NIR spectra.

2.5.5. Applications to fruits and vegetables

The lower absorption intensities of water in the NIR compared to the MIR means that NIR is useful for measurements on high moisture samples such as fruit. A mode for measurements that is especially useful for measurements on intact fruits is called interaction, and it uses a fiber-optic probe, where the probability of the incident beam to interact with the sample is increased (Pasquini, 2003).

NIR is thus used for the assessment of the quality of fruits and vegetables. It has been successfully applied for determination of some of the most important quality aspects of fruits such as soluble solid, sugar and acidity content. Soluble solid content (SSC) or total solid content (TSC) (or dry matter content) has been determined for various fruits such as apples, melons, peaches, tomatoes, kiwis and dates and for vegetables such as onions, potatoes and corn by use of NIR (Kawano, 2000; Lammertyn et al., 2000; Schmilovitch et al., 1999; Slaughter et al., 2003).

NIR spectra have been found useful for determination of sucrose, glucose, fructose, citric acid, malic acid and ascorbic acid content in strawberries (Jin and Cui, 1994). In an experiment using reflectance measurements on potatoe slices, a calibration of sugar content, however, did not perform well, and NIR was also found insensitive to the fructose content (Scanlon et al., 1999). On the other hand, much better calibrations of sugar content have been reported when using transmission measurements on thin potato slices (Mehrubeoglu and Cote, 1997). NIR is in addition able to detect secondary metabolites in plants since, e.g., phenols, alkaloids, tannins and glucosinolate have distinct absorptions in the spectrum (Foley et al., 1998).

Also textural properties of fruits and vegetables can correlate to NIR spectra. Calibration models of NIR spectra of pears could predict fruit hardness, juiciness and mealiness. NIR can also predict the firmness, waxiness and mealiness of boiled potatoes as well as hardness and crispiness of boiled carrots (De Belie et al., 2003).

3. Conclusions

The field of classical proteomics should be extended to also include handling of large datasets by appropriate data analysis. The analysis performed by 2DGE, MS and NIR give rise to many data and multivariate data analysis can unravel the complicated data structures, which can relieve the characterization phase in classical proteomics. Based on analysis of proteins from the wheat gluten complex, we have used this technique to focus on the interesting spots or the interesting part of spectra before the actual identification phase. Multivariate analysis has shown interesting results as a supplement to classical proteomics and added a new dimension to the field of proteomics.

4. Experimental

For preparation of gluten powder, dough was made by mixing commercial wheat flour with distilled water (2:1) by hand. Gluten was washed out manually with distilled water from the dough and freeze-dried. The freeze-dried gluten was ground to powder and sieved through a 500-μm screen.

Moist samples were obtained by placing 2×700 mg gluten powder in a sealed container with water in the

bottom, and FT-NIR spectra were collected at 2, 4, and 24 h after leaving the samples for moisture absorption. FT-NIR spectra were also obtained from two gluten samples, which were kept dry.

For the FT-NIR measurements, powders were compressed in a sample cup and spectra were recorded using a Spectrum One NTS, Perkin–Elmer spectrometer in reflectance mode. Spectral data were recorded from 793 to 2495 nm at 1.67 nm intervals with co-addition of 50 scans and use of a spectral resolution of 8 cm^{-1}. A Spectralon® diffuse reflectance standard was used as reference. NIR spectra were analyzed by using The Unscrambler Software version 8.0. Spectral preprocessing included MSC.

Acknowledgements

Harald Martens is thanked for very inspiring multivariate data analysis discussions. Ljiljana Nesic and Marianne K. Petersen are greatly appreciated for their crucial experimental assistance. Center for Advanced Food research is gratefully acknowledged. This work is part of the Danish Cereal Network supported by The Ministry of Food, Agriculture and Fisheries.

Appendix A

Here we describe the spot detection algorithm. Notation and syntax are that of Matlab.

The initial model is calculated using Singular Value Decomposition (SVD) as:

$[\mathbf{U},\mathbf{S},\mathbf{V}] = \mathrm{svd}(\mathbf{X})$.

From this we get **T** and **P** as:

$\mathbf{T} = \mathbf{VS}$,

$\mathbf{P} = \mathbf{U}$.

The optimal number of principal components to use in the model is obtained by studying the **S** matrix, which is a diagonal matrix of the eigenvalues. The eigenvalues describe how much of the variance is described by each principal component, and is commonly expressed in percents as:

$$\frac{\mathrm{diag}(\mathbf{S})\cdot 100}{\mathrm{sum}(\mathrm{diag}(\mathbf{S}))}.$$

From this the optimal number of principal components (AOpt) is determined and the **T**, **P** and **E** matrixes are constructed and the reference variance is calculated (Fig. 3, step 2):

$\mathbf{T} = \mathbf{T}(:,1:A\mathrm{Opt})$,

$\mathbf{P} = \mathbf{P}(:,1:A\mathrm{Opt})$,

$\mathbf{E} = \mathbf{Peak}' - \mathbf{T}\cdot\mathbf{P}'$,

$s2\mathrm{Ref} = \mathrm{mean}(\mathbf{E}(:).\hat{}2)$.

In every pixel of the gel image to be analyzed is a sub-image the same size as the spot sampling sub-image is extracted and unfolded to row vector **x**, which is used as data in a new model that is calculated from the loadings matrix from the reference model. Hereby it is possible to estimate a new score matrix, by projecting **x** on **P** and hence the residual matrix and variance for the sub-image:

$\mathbf{x} = \mathbf{t}\cdot\mathbf{P}' + \mathbf{e}$,

$\mathbf{t} = \mathbf{x}\cdot\mathbf{P}\cdot(\mathbf{PP}')^{-1}$,

$\mathbf{t} = \mathbf{x}\cdot\mathbf{P}$,

$\mathbf{e} = \mathbf{x} - \mathbf{t}\cdot\mathbf{P}'$,

$s^2 = \mathrm{mean}(\mathbf{e}(:).\hat{}2)$.

The variance, $s2$ is used to calculate a peak-probability from the $s2$Ref (Fig. 3, step 7):

$s2\mathrm{Ratio} = s2/s2\mathrm{Ref}$,

$\mathrm{Prob} = \mathrm{ScalingFactor} * 1/s2\mathrm{Ratio}$,

$\mathrm{Prob} = \max(\mathrm{Prob},0)$,

$\mathrm{Prob} = \min(\mathrm{Prob},1)$,

$\mathrm{PeakProbability} = \mathrm{Prob} + f(\mathrm{leverage})$.

This is shown below:

- Make reference PCA model and calculate reference scores and loadings.
- Study the **S** matrix and determine the optimal number of principal components (AOpt) and estimate the reference variance ($s2$Ref) with this number of principal components.
  - for $h = 1:nh$ (number of horisontal pixels in the gel-image to be analyzed)
    - for v = 1:nv (number of vertical pixels in the gel-image to be analyzed)
      - Extraxt a sub-image with centre in the pixel $(v-dv:v+dv, h-dh:h+dh)$
      - Reshape to a row vector, **x**
      - Estimate the score matrix for the new **x** data from the reference loadings.
      - Estimate the residual for the new **x** data: $\mathbf{e} = \mathbf{x} - \mathbf{t} * \mathbf{P}'$
      - Calculate the variance: $s2 = \mathrm{mean}(\mathbf{e}(:)).\hat{}2)$
      - Calculate the probability factor for the pixel (v,h)
    - end
  - end

Before the reference model is made the data is normalised. This is done by Multiplicative Scatter Correction (MSC) (Martens et al., 2003).

References

Anderson, N.L., Matheson, A.D., Steiner, S., 2000. Proteomics: applications in basic and applied biology. Current Opinion in Biotechnology 11, 408–412.

Appel, R., Hochstrasser, D., Rock, C., Funk, M., Muller, A.F., Pellegrini, C., 1988. Automatic classification of two-dimensional gel-electrophoresis pictures by heuristic clustering analysis – a step toward machine learning. Electrophoresis 9, 136–142.

Appel, R.D., Palagi, P.M., Walther, D., Vargas, J.R., Sanchez, J.-C., Ravier, F., Pasquali, C., Hochstrasser, D.F., 1997. Melanie II – a third-generation software package for analysis of two-dimensional electrophoresis images: I. Features and user interface. Electrophoresis 18, 2724–2734.

Belton, P.S., Colquhoun, I.J., Grant, A., Wellner, N., Field, J.M., Shewry, P.R., Tatham, A.S., 1995. FTIR and NMR-studies on the hydration of a high-M(R) subunit of glutenin. International Journal of Biological Macromolecules 17, 74–80.

Bienvenut, W.V., Sanchez, J.-C., Karmine, A., Rouge, V., Rose, K., Binz, P.-A., Hochstrasser, D.F., 1999. Toward a clinical molecular scanner for proteome research: parallel protein chemical processing before and during western blot. Analytical Chemistry 71, 4800–4807.

Binz, P.A., Muller, M., Walther, D., Bienvenut, W.V., Gras, R., Hoogland, C., Bouchet, G., Gasteiger, E., Fabbretti, R., Gay, S., Palagi, P., Wilkins, M.R., Rouge, V., Tonella, L., Paesano, S., Rossellat, G., Karmine, A., Bairoch, A., Sanchez, J.-C., Appel, R.D., Hochstrasser, D.F., 1999. A molecular scanner to automate proteomic research and to display proteome images. Analytical Chemistry 71, 4981–4988.

Bokobza, L., 2002. Origin of near-infrared absorption bands. In: Siesler, H.W., Ozaki, Y., Kawata, S., Heise, H.M. (Eds.), Near-infrared Spectroscopy. Principles, Instruments, Applications. Wiley, Weinheim, pp. 11–41.

Corthals, G.L., Wasinger, V.C., Hochstrasser, D.F., Sanchez, J.-C., 2000. The dynamic range of protein expression: a challenge for proteomic research. Electrophoresis 21, 1104–1115.

Czuchajowska, Z., Pomeranz, Y., 1991. Evaluation of vital dry gluten composition and functionality in breakmaking by near-infrared reflectance spectoscopy. Cereal Foods World 36, 439–464.

Davies, A.M.C., 1993. Introduction to NIR spectroscopy. The Second European Symposium on Near Infrared (NIR) Spectroscopy, 1–16.

De Belie, N., Pedersen, D.K., Martens, M., Bro, R., Munck, L., De Baerdemaeker, J., 2003. The use of visible and near-infrared reflectance measurements to assess sensory changes in carrot texture and sweetness during heat treatment. Biosystems Engineering 85, 213–225.

Eckerskorn, C., Strupat, K., 2000. Mass spectrometry of intact proteins from two-dimensional PAGE. In: Rabilloud, T. (Ed.), Proteome Research: Two-dimensional Gel Electrophoresis and Identification Methods. Springer Verlag, Berlin, pp. 233–244.

Esbensen, K.H., Guyot, D., Westad, F., 2000. Multivariate Data Analysis – In Practice. Camo ASA, Oslo.

Everitt, B.S., Dunn, G., 1991. Applied Multivariate Data Analysis. Edward Arnold, London.

Fey, S.J., Larsen, P.M., 2001. 2D or not 2D. Current Opinion in Chemical Biology 5, 26–33.

Foley, W.J., McIlwee, A., Lawler, I., Aragones, L., Woolnough, A.P., Berding, N., 1998. Ecological applications of near infrared reflectance spectroscopy a tool for rapid, cost-effective prediction of the composition of plant and animal tissues and aspects of animal performance. Oecologia 116, 293–305.

Gidman, E., Goodacre, R., Emmett, B., Smith, A.R., Gwynn-Jones, D., 2003. Investigating plant–plant interference by metabolic fingerprinting. Phytochemistry 63, 705–710.

Godovac-Zimmermann, J., Brown, L.R., 2001. Perspectives for mass spectrometry and functional proteomics. Mass Spectrometry Reviews 20, 1–57.

Gottlieb, D.M., Schultz, J., Petersen, M., Nesic, L., Jacobsen, S., Søndergaard, I., 2002. Determination of wheat quality by mass spectrometry and multivariate data analysis. Rapid Communications In Mass Spectrometry 16, 2034–2039.

Gustafsson, J.S., Blomberg, A., Rudemo, M., 2002. Warping two-dimensional electrophoresis gel images to correct for geometric distortions of the spot pattern. Electrophoresis 23, 1731–1744.

Görg, A., Obermaler, C., Boguth, G., Harder, A., Scheibe, B., Wildgruber, R., Weiss, W., 2000. The current state of two-dimensional electrophoresis with immobilized pH gradients. Electrophoresis 21, 1037–1053.

Hanash, S.M., 2000. Biomedical applications of two-dimensional electrophoresis using immobilized pH gradients: current status. Electrophoresis 21, 1202–1209.

Harry, J.L., Wilkins, M.R., Herbert, B.R., Packer, N.H., Gooley, A.A., Williams, K.L., 2000. Proteomics: capacity versus utility. Electrophoresis 21, 1071–1081.

Heise, H.M., Winzen, R., 2002. Chemometrics in near-infrared spectroscopy. Principles, instruments, applications. In: Siesler, H.W., Ozaki, Y., Kawata, S., Heise, H.M. (Eds.), Near-infrared Spectroscopy. Wiley, Weinheim, pp. 125–162.

Hille, J.M., Freed, A.L., Wätzig, H., 2001. Possibilities to improve automation, speed and precision of proteome analysis: a comparison of two-dimensional electrophoresis and alternatives. Electrophoresis 22, 4035–4052.

Humphery-Smith, I., 2000. Introduction: the virtue of proteomics. In: Rabilloud, T. (Ed.), Proteome Research: Two-dimensional Gel Electrophoresis and Identification Methods. Springer Verlag, Berlin, pp. 1–8.

Höskuldsson, A., 2001. Variable and subset selection in PLS regression. Chemometrics and Intelligent Laboratory Systems 55, 23–38.

Iwamoto, M., Kawano, S., 1992. Advantages and disadvantages of NIR applications for the food industry. In: Murray, I., Cowe, I.A. (Eds.), Making Light Work: Advances in Near Infrared Spectroscopy. Wiley, Cambridge, UK, pp. 367–375.

Jacobsen, S., Søndergaard, I., Møller, B., Desler, T., Munck, L., 2004. The barley endosperm as a data interface for the expression of genes and gene combinations at different levels of biological organization explored through pattern-recognition data evaluation. J. Cereal Science (submitted).

Jessen, F., Lametsch, R., Bendixen, E., Kjærsgård, I.V.H., Jørgensen, B.M., 2002. Extracting information from two-dimensional electrophoresis gels by partial least squares regression. Proteomics 2, 32–35.

Jin, T.M., Cui, H.C., 1994. A new method for determination of nutrient contents of intact strawberries – near infrared spectrometry. Acta Agriculture Boreal Singapore 9, 120–123.

Kaczmarek, K., Walczak, B., de Jong, S., Vandeginste, B.G.M., 2002. Feature based fuzzy matching of 2D gel electrophoresis images. Journal of Chemical Information and Computer Science 42, 1431–1442.

Kawano, S., 2000. Application to agricultural products and foodstuffs. In: Siesler, H.W., Ozaki, Y., Kawata, S., Heise, H.M. (Eds.), Near-infrared Spectroscopy. Principles, Instruments, Applications. Wiley, Weinheim, pp. 269–287.

Kinter, M., Sherman, N.E., 2000. Protein Sequencing and Identification Using Tandem Mass Spectrometry. Wiley, New York.

Kriegel, K., Seefeldt, I., Hoffmann, F., Schultz, C., Wenk, C., Regitz-Zagrosek, V., Oswald, H., Fleck, E., 2000. An alternative approach to deal with geometric uncertainties in computer analysis of two-dimensional electrophoresis gels. Electrophoresis 21, 2637–2640.

Küster, B., Krogh, T.N., Mørtz, E., Harvey, D.J., 2001. Glycosylation analysis of gel-separated proteins. Proteomics 1, 350–361.

Lammertyn, J., Peirs, A., De Baerdemaeker, J., Nicolai, B., 2000. Light penetration properties of NIR radiation in fruit with respect to non-destructive quality assessment. Postharvest Biology and Technology 18, 121–132.

Lopez, M.F., 2000. Better approaches to finding the needle in a haystack: optimizing proteome analysis through automation. Electrophoresis 21, 1082–1093.

Mahon, P., Dupree, P., 2001. Quantitative and reproducible two-dimensional gel analysis using Phoretix 2D Full. Electrophoresis 22, 2075–2085.

Martens, H., Martens, M., 2001. Multivariate Analysis of Quality – An Introduction. John Wiley & Sons Ltd, Chichester.

Martens, H., Nielsen, J.P., Engelsen, S.B., 2003. Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures. Analytical Chemistry 75, 394–404.

Mehrubeoglu, M., Cote, G.L., 1997. Determination of total reducing sugars in potato samples using near-infrared spectroscopy. Cereal Foods World 42, 409–413.

Morris, C.F., Rose, S.P., 1996. Wheat. In: Henry, R.J., Kettlewell, P.S. (Eds.), Cereal Grain Quality. Chapmann & Hall, London, pp. 5–54.

Munck, L., Nielsen, J.P., Møller, B., Jacobsen, S., Søndergaard, I., Engelsen, S.B., Nørgaard, L., Bro, R., 2001. Exploring the phenotypic expression of a regulatory proteome-altering gene by spectroscopy and chemometrics. Analytica Chimica Acta 446, 171–186.

Naaby-Hansen, S., Waterfield, M.D., Cramer, R., 2001. Proteomics – post-genomic cartography to understand gene function. Trends in Pharmacological Sciences 22, 376–384.

Nordhoff, E., Egelhofer, V., Giavalisco, P., Eickhoff, H., Horn, M., Przevieslik, T., Theiss, D., Schneider, U., Lehrach, H., Gobom, J., 2001. Large-gel two-dimensional electrophoresis-matrix assisted laser desorption/ionization-time of flight-mass spectrometry: an analytical challenge for studying complex protein mixtures. Electrophoresis 22, 2844–2855.

Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J.P., Munck, L., Engelsen, S.B., 2000. Interval partial least squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. Applied Spectroscopy 54, 413–419.

Osborne, B.G., Fearn, T., Hindle, P.H., 1993. Practical NIR Spectroscopy with Applications in Food and Beverage Analysis. Longman Scientific & Technical, Harlow.

Pasquini, C., 2003. Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. Journal of the Brazilian Chemical Society 14, 198–219.

Patterson, S.D., 2000. Proteomics: the industrialization of protein chemistry. Current Opinion in Biotechnology 11, 413–418.

Pleissner, K.P., Hoffmann, F., Kriegel, K., Wenk, C., Wegner, S., Sahlstrom, A., Oswald, H., Alt, H., Fleck, E., 1999. Proteome data analysis and management – new algorithmic approaches to protein spot detection and pattern matching in two-dimensional electrophoresis gel databases. Electrophoresis 20, 755–765.

Pun, T., Hochstrasser, D.F., Appel, R.D., Funk, M., Villars-Augsburger, V., Pellegrini, C., 1988. Computerized classification of two-dimensional gel electrophoretograms by correspondence analysis and ascendant hierarchical clustering. Applied Theoretical Electrophoresis 1, 3–9.

Rabilloud, T., Vincens, P., Tarroux, P., 1985. A new tool to study genetic expression using 2-D electrophoresis eata – the functional map concept. FEBS Lett. 189, 171–178.

Radzikowski, L., Nesic, L., Hansen, H.B., Jacobsen, S., Søndergaard, I., 2002. Comparison of ethanol-soluble proteins from different rye (Secale cereale) varieties by two-dimensional electrophoresis. Electrophoresis 23, 4157–4166.

Raman, B., Cheung, A., Marten, M.R., 2002. Quantitative comparison and evaluation of two commercially available, two-dimensional electrophoresis image analysis software packages, Z3 and Melanie. Electrophoresis 23, 2194–2202.

Scanlon, M.G., Pritchard, M.K., Adam, L.R., 1999. Quality evaluation of processing potatoes by near infrared reflectance. Jornal of the Science of Food and Agriculture 79, 763–771.

Schmilovitch, Z., Hoffman, A., Egozi, H., Ben Zvi, R., Bernstein, Z., Alchanatis, V., 1999. Maturity determination of fresh dates by near infrared spectrometry. Journal of the Science of Food and Agriculture 79, 86–90.

Schultz, J., Gottlieb, D.M., Petersen, M., Nesic, L., Jacobsen, S., Søndergaard, I., 2004. Explorative data analysis of 2-D electrophoresis gels. Electrophoresis 25, 502–511.

Sefara, N.L., Magtoto, N.P., Richardson, H.H., 1997. Structural characterization of p-lactoglobulin in solution using two-dimensional FT mid-infrared and FT near-infrared correlation spectroscopy. Applied Spectroscopy 51, 536–540.

Siesler, H.W., 2002. Introduction. In: Siesler, H.W., Ozaki, Y., Kawata, S., Heise, H.M. (Eds.), Near-infrared Spectroscopy. Principles, Instruments, Applications. Wiley, Weinheim, pp. 1–10.

Slaughter, D.C., Thompson, J.F., Tan, E.S., 2003. Nondestructive determination of total and soluble solids in fresh prune using near infrared spectroscopy. Postharvest Biology and Technology 28, 437–444.

Smilansky, Z., 2001. Automatic registration for images of two-dimensional protein gels. Electrophoresis 22, 1616–1626.

Steinberg, T.H., Lauber, W.M., Berggren, K., Kemper, C., Yue, S., Patton, W.F., 2000. Fluorescence detection of proteins in sodium dodecyl sulfate–polyacrylamide gels using environmentally benign, nonfixative, saline solution. Electrophoresis 21, 497–508.

Sumner, L.W., Mendes, P., Dixon, R.A., 2003. Plant metabolomics: large-scale phytochemistry in the functional genomics era. Phytochemistry 62, 817–836.

Tarroux, P., 1983. Analysis of protein-patterns during differentiation using 2-D electrophoresis and computer multidimensional classification. Electrophoresis 4, 63–70.

Tarroux, P., Vincens, P., Rabilloud, T., 1987. Hermes – A 2nd generation approach to the automatic-analysis of two-dimensional electrophoresis gels. 5. Data-analysis. Electrophoresis 8, 187–199.

Veeser, S., Dunn, M.J., Yang, G.-Z., 2001. Multiresolution image registration for two-dimensional gel electrophoresis. Proteomics 1, 856–870.

Veraverbeke, W.S., Delcour, J.A., 2002. Wheat protein composition and properties of wheat glutenin in relation to breadmaking functionality. Critical Reviews in Food Science and Nutrition 42, 179–208.

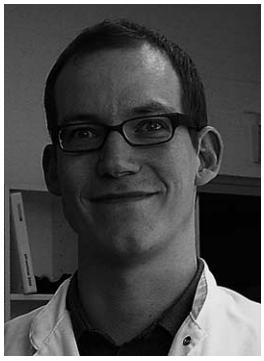
Vohradsky, J., 1997. Adaptive classification of two-dimensional gel electrophoretic spot patterns by neural networks and cluster analysis. Electrophoresis 18, 2749–2754.

Walker, A.K., Rymar, G., Andrews, P.C., 2001. Mass spectrometric imaging of immobilized pH gradient gels and creation of virtual two-dimensional gels. Electrophoresis 22, 933–945.

Wesley, I.J., Larroque, O., Osborne, B.G., Azudin, N., Allen, H., Skerritt, J.H., 2001. Measurement of gliadin and glutenin content of flour by NIR spectroscopy. Jornal of Cereal Science 34, 125–133.

Wu, Y.Q., Czarnik-Matusewicz, B., Murayama, K., Ozaki, Y., 2000. Two-dimensional near infrared spectroscopy study of human serum albumin in aqueous solutions: using overtones and combination modes to monitor temperature-dependent changes in the secondary structure. Journal of Physical Chemistry B 104, 5840–5847.

David M. Gottlieb is Chemist at the Plasma Product Division at Statens Serum Institute in Copenhagen, Denmark. Before entering his current position he worked as a research assistant in the Biochemistry and Nutrition Group at BioCentrum-DTU. The work of Dr. Gottlieb has been centred on the use of multivariate statistics in the analysis of wheat proteins and wheat properties. He graduated from the Technical University of Denmark as chemical engineer under the supervision of Prof. Ib Søndergaard and Prof. Susanne Jacobsen. His studies at the Technical University of Denmark have been supplemented with studies in pharmacology and biochemistry at the University of Liverpool.



Jakob Schultz is working as a research assistant in the Biochemistry and Nutrition Group at BioCentrum-DTU. The work of Dr. Schultz is centred on the development of novel ways to analyse two-dimensional electrophoresis gels, focusing on easy extraction of relevant data. This among other things involves the use of multivariate statistics, multivariate image analysis and proteomics. He graduated from the Technical University of Denmark as chemical engineer under the supervision of Prof. Ib Søndergaard and Prof. Susanne Jacobsen. His studies at the Technical University of Denmark have been supplemented with studies in food technology and brewery technology at the Technical University of Munich.

Susanne Jacobsen is a Ph.D. in biochemistry and associate professor at BioCentrum-DTU, Biochemistry and Nutrition Group, at the Technical University of Denmark in Kgs. Lyngby, Denmark.The work of Prof. Jacobsen has been centred on proteins, proteomics and the integration of biochemical analysis methods – mainly electrophoretic and mass spectrometric methods – with the use of digital image processing, neural networks and multivariate statistical methods. The goal is to gain insight into complex biochemical and biological system using non-destructive and non-invasive methods.

Susanne Wrang Bruun is Ph.D. student in the Biochemistry and Nutrition Group at BioCentrum-DTU. The work of Dr. Bruun is centred on the development of novel ways to analyse cereal protein interactions with proteins and carbohydrates. Near infrared spectrometry is used in combination with multivariate statistics. She graduated from the Biochemistry and Nutrition Group at BioCentrum-DTU, Technical University of Denmark as chemical engineer. As part of the Ph.D. study she is currently studying infrared, near infrared technology and multivariate data analysis at MatForsk, Aas, Norway.

Ib Søndergaard is biochemist and associate professor at BioCentrum-DTU, Biochemistry and Nutrition Group, at the Technical University of Denmark in Kgs. Lyngby, Denmark. Before entering his current position he worked at the Royal Veterinary and Agricultural University, the University Hospital of Copenhagen and the University of Copenhagen. Prof. Søndergaard is currently a member of the Danish Agricultural and Veterinary Research Council and the Danish Nutrition Council. The work of Prof. Søndergaard has been centered on proteins, proteomics and the integration of biochemical analysis methods – mainly electrophoretic and mass spectrometric methods – with the use of digital image processing, neural networks and multivariate statistical methods. The goal is to gain insight into complex biochemical and biological system using non-destructive and non-invasive methods.