Principal Component Analysis of Near Infrared Spectra as a Tool of Endosperm Mutant Characterisation and in Barley Breeding for Quality

Lars MUNCK and Birthe MØLLER

Department of Food Science, The Royal Veterinary and Agricultural University, Frederiksberg, Denmark

Abstract: Near infrared technology, now widespread in quality control, makes it possible to obtain a total multivariate physical chemical fingerprint of the barley endosperm with high precision. Whole spectroscopic fingerprints of the physics and chemistry of barley seeds can be interpreted by multivariate analysis (chemometrics), by Principal Component Analysis (PCA) for classification and Partial Least Squares Regression (PLSR) for correlation. PCA classification of Near Infrared Reflectance (NIR) spectra can differentiate between mutants and alleles in the *lys3* and *lys5* loci. PCA on NIR can also be used as a routine in barley breeding to select for a multi-gene quality complex in barley as a whole e.g. increasing starch and reducing fibre content. This is done directly from the PCA classification plot by "data breeding" selecting the recombinants which are approaching the position of the normal high starch controls on the plot. Based on classification of NIR spectra, two alleles in the *lys5* locus were characterised as a new class of (1→3,1→4)-β-glucan compensating starch mutants indicating a metabolic connection between starch and β-glucan.

Keywords: multivariate data analysis; NIR spectroscopy; barley mutants; physical-chemical fingerprint

Near Infrared Spectroscopy (NIR) combined with multivariate data analysis is an established economical, non-destructive technique (Osborne *et al*. 1993) for prediction of chemical composition widely used by plant breeders, seed producers and malt and feed industries. The method is fast, and often more than one component can be determined at the same time. In barley commercial calibrations sold by the instrument manufacturers have been made for moisture, protein, starch and hot water extract. The calibrations are based on multivariate (chemometric) models such as Partial Least Squares Regression (PLSR) and neural nets. While NIR is widely used by plant breeders as a "black box" for prediction of chemical composition, much less attention has been devoted to the use of NIR spectra to be used directly in the breeding work for classification of the whole seed phenotype by Principal Component Analysis (PCA). We know from the calibration work with PLSR that a NIR spectrum of a barley seed sample can be considered as a whole chemical-physical fingerprint. Thus barley mutants and lines bred for improved quality can be explored and compared to controls in a PCA without knowledge of their chemical composition. Spectra from different barley seed sample positions in the PCA score plot represent each a physical-chemical approximation of the "phenome" (Munck *et al*. 2004). From the spectral differences specific changes in composition can be anticipated based on prior spectroscopic knowledge from literature and checked by classical chemical analyses (Jacobsen *et al*. 2005). In the present paper we will demonstrate how NIR and multivariate

analysis can be used in barley genetics and plant breeding on the phenome level to characterise endosperm mutants and gene combinations as specific spectral signatures.

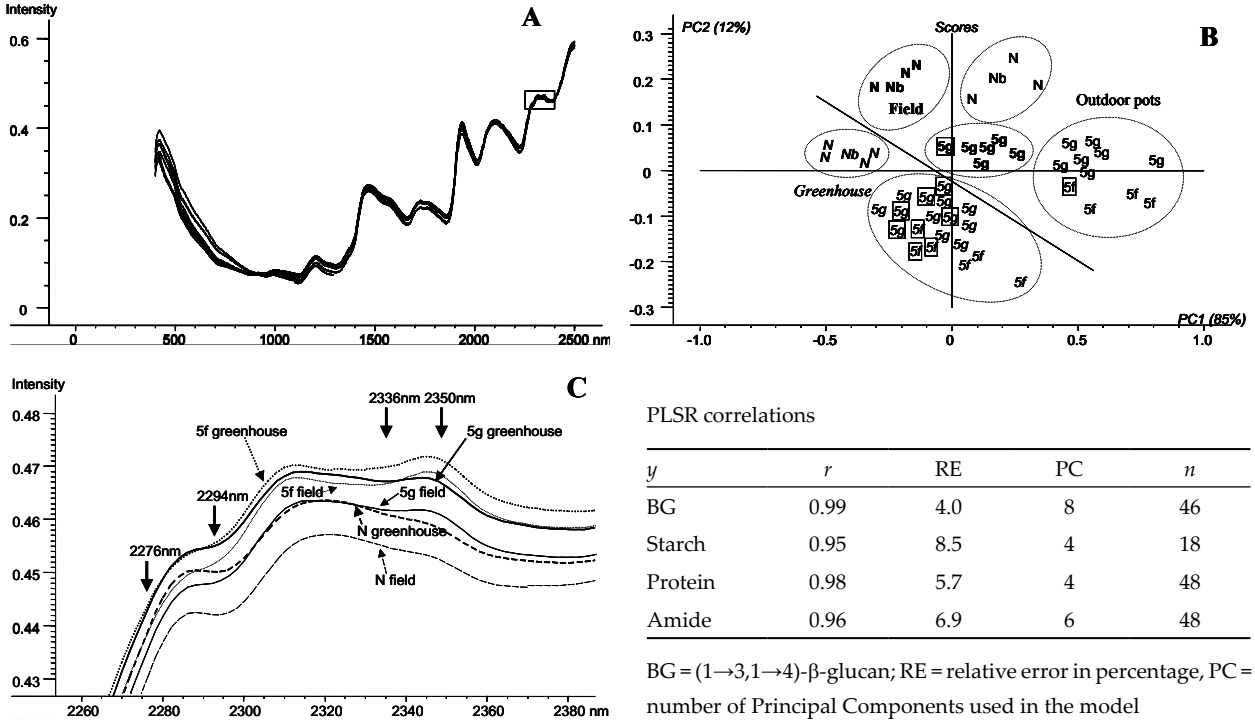
MATERIAL AND METHODS

Two barley materials are used: Material I – 49 samples of *lys5f* and *lys5g* low starch mutants (Doll 1983; Jacobsen *et al*. 2005) with moderate increase in lysine (appr. 10%) including homozygous lines (doubled monoploids) from crosses between these mutants and wild type barley varieties as well as other wild type barley control lines were grown in greenhouse (n = 13), outdoor pots (n = 7) and field (n = 6) in 2000. Material II – 14 samples of the very high mutants *lys3a* and *lys3m* and seven segregants (0502, 0505, 0531, 0538, 0556, Lysimax, Lysiba) as well as six barley lines (two samples of each of Minerva, Bomi and Triumph) were grown in the field in 1991. Multiplicative signal corrected (MSC) NIR spectra (400–2500 nm) were obtained on flour from all samples (Munck *et al*. 2004). Every second wavelength was used in the classification (1050 data points). Protein, amide, starch and (1→3,1→4)-ß-glucan were determined according to Munck *et al*. (2001). A Principal Component Analysis (PCA) software data program for classification and Partial Least Squares Regression (PLSR) program for correlation ("Unscrambler", Camo A/S, Trondheim, Norway) were used to analyse the NIR spectra.

RESULTS AND DISCUSSION

Differentiating between gene specific and environmental effects on NIR spectroscopy patterns from barley endosperms

Figure 1A shows the MSC-corrected NIR spectra from 49 samples of barley grown in greenhouse, outdoor pots and in the field. The barley material consists of normal barley varieties (N) and two low starch mutants (*lys5f* and *lys5g*) selected at Risø (Munck *et al*. 2004) as moderate high-lysine lines by the dye-binding method. The *lys5f* mutant is more drastic than *lys5g* because it has a lower



PLSR correlations

y	*r*	RE	PC	*n*
BG	0.99	4.0	8	46
Starch	0.95	8.5	4	18
Protein	0.98	5.7	4	48
Amide	0.96	6.9	6	48

BG = (1→3,1→4)-β-glucan; RE = relative error in percentage, PC = number of Principal Components used in the model

Figure 1. **A.** All 49 NIR spectra (MSC) from material 1 in Material and Methods **B.** PCA (PC1:2) score plot of NIR spectra (MSC) for 49 samples of normal barley (N, Nb = Bomi) and mutants and mutant recombinants crossings of *lys5f* (*5f*) and *lys5g* (*5g*) grown in three environments: *greenhouse*, **field** and in pots outdoors. Squared samples are original mutants. **C.** Average spectra of normal barley (N) and mutants of *lys5f* (*5f*) and *lys5g* (*5g*) grown in greenhouse and field. **Table.** PLSR predictions of chemical properties BG from NIR spectra (MSC)

starch content. The mutants are in the same locus and are identical to Risø mutant 13 (background Bomi) and mutant 29 (background Carlsberg II) respectively. By using the unsupervised classification algorithm PCA in evaluating the whole spectra in Figure 1A, six clusters according to genetic parameters as well as for environment can be interpreted by consulting the field book (Figure 1B). Normal samples are located in the top left corner, whereas *lys5* mutants and recombinants are spread diagonally in the bottom right corner. There is a tendency that the more drastic mutant *lys5f* is placed in the periphery of the *lys5* clusters (Figure 1B). Furthermore, it is seen that samples in the two clusters grown in greenhouse are located diagonally in the bottom left corner below the division line, samples from the field (bold) are located in the middle and samples grown in outdoor pots are placed diagonally in the top right corner.

Because samples located near each other in a PCA plot indicate similar pattern of variables, here NIR wavelengths, mean spectra from the six different clusters in Figure 1B can be characterised by visual inspection. In Figure 1C an interesting small area from 2260 to 2380 nm marked with a square in Figure 1A is enlarged for the average spectra of the six classes in the PCA in Figure 1B. It is easily recognised that the average curve form in Figure 1C of the normal barley (N, stippled lines) is quite different from those of the mutants (punctuated line *lys5f*, whole line *lys5g*). While the normal barleys have a decided plateau at 2285–2295 nm the slope is here much steeper with the two *lys5* mutants. Likewise the mutants have a characteristic bulb at approximately 2350 nm which is almost non existent in normal barley. This bulb is larger in the spectra of the *lys5f* mutant than for that of *lys5g*. When comparing growing locations in Figure 1C it is seen that the curve form for each mutant genotype is approximately the same while an offset base line indicates the growing location where the greenhouse has a higher intensity than the field. As indicated from the PCA in Figure 1B, the original *lys5f* and *lys5g* mutations (marked with a square) are obtaining full penetration in the recombinants. Thus the recombinants are classified together with the original mutants in the PCA. As demonstrated by us (Munck *et al*. 2004; Jacobsen *et al*. 2005) it is also possible to differentiate between several mutants and alleles from the *lys3*, *lys5* loci and other loci in the same data set with PCA.

Verifying the chemical nature of the *lys5* PCA cluster as (1→3,1→4)-β-glucan compensated starch mutants

In an unsupervised PCA, where one does not know the origin of the samples, the spectra may be inspected using prior knowledge regarding critical wavelengths indicative of certain chemical bonds in order to define analyses for chemical validation. In comparing the wavelength areas indicating different chemical compounds and bonds (Figure 1C), it is seen that the mean spectra from the mutants

Table 1. Chemical properties of the six groups in Figure 1 as well as for *lys5f* and *lys5g* compared to Bomi

		n	BG	Starch (S)	BG + S	Protein	Amide
N	greenhouse	5	7.4 ± 1.7	50.1 ± 3.9[a]	57.5 ± 2.5[a]	15.0 ± 1.4	0.41 ± 0.05
5		19	14.6 ± 3.0	33.5 ± 7.5[a]	51.6 ± 4.5[a]	17.5 ± 1.3	0.44 ± 0.04
N	field	4	5.0 ± 0.9	55.9 ± 3.5	60.9 ± 2.8	10.5 ± 0.8	0.26 ± 0.03
5		6	8.5 ± 0.9	43.2 ± 4.0[a]	55.1 ± 4.4[a]	12.4 ± 0.7	0.28 ± 0.02
N	outdoor pots	4	7.4 ± 0.6[b]	44.6 ± 1.3[c]	52.0 ± 0.6[c]	16.7 ± 0.4[b]	0.44 ± 0.02[b]
5		11	13.3 ± 2.6	–	–	18.1 ± 1.1	0.44 ± 0.04
Bomi compared to *lys5f* and *lys5g* mutants							
Bomi		1	6.8	48.8	55.6	14.6	0.4
lys5f	greenhouse	3	19.6 ± 0.4	29.8 ± 0.6	49.4 ± 1.0	17.0 ± 1.4	0.42 ± 0.06
lys5g		6	13.1 ± 0.8	44.7[d]	58.2[d]	17.4 ± 1.0	0.43 ± 0.04

[a]*n* = 4, [b]*n* = 3, [c]*n* = 2, [d]*n* = 1; BG = (1→3,1→4)-β-glucan

and normal barleys are different from each other at the wavelength about 2276 nm correlating to starch, at 2294 nm describing amino acids, at 2336 and 2352 nm characteristic of cellulose content and unsaturated fat at 2347 nm. From here it is possible to make hypotheses about differences in chemical composition and to evaluate these by choosing proper reference analyses for final validation. Several authors like Doll (1983) and Greber *et al*. (2000) suggested that the high lysine mutants of barley could relate to starch synthesis which was more or less reduced in these mutants. It was therefore surprising when we found (Munck *et al*. 2004; Jacobsen *et al*. 2005) that *lys5f* and *lys5g*, which all were classified by NIR spectroscopy and PCA in the same cluster, had a strongly increased content of (1→3,1→4)-β-glucan which compensated the low starch content to a large extent. In Table 1 the chemical composition of the six clusters and the individual genotypes classified in the experiment in Figure 1 are shown. The drastic increase in (1→3,1→4)-β-glucan especially for the *lys5f* genotype – up to 20% in dry matter – is clearly seen. The *lys5g* genotype seems to fully compensate the loss in starch with (1→3,1→4)-β-glucan as seen from the total figure of (1→3,1→4)-β-glucan plus starch (BG + S, Table 1) which is comparable to the normal controls.

Greenhouse conditions increase both protein and (1→3,1→4)-β-glucan while keeping the relative differences between the different genotypes. The function of the highly reproducible NIR spectrograph as a "multimeter" is clearly pointed out in Figure 1 (Table) displaying fully cross validated PLSR prediction values for whole spectra in correlations to the different chemical analyses including (1→3,1→4)-β-glucan. This explains also the high discriminating power of the PCA in Figure 1C using the same NIR spectral data.

The question is now if NIR spectroscopy interpreted by chemometrics is restricted to detecting mutants with major effects on chemical composition? Is it possible to use this technology in regular plant breeding e.g. to improve starch and reduce fibre content in the very high lysine genotype *lys3a* – Risø mutant 1508 by crossbreeding and selection?

The feasibility of selecting for an improved gene background for high starch, low fibre to the *lys3* gene for starch by NIR spectroscopy and chemometrics

As demonstrated by Munck (1992) it is possible to breed for an improved kernel quality and yield with the *lys3a* mutant to reach the level of normal

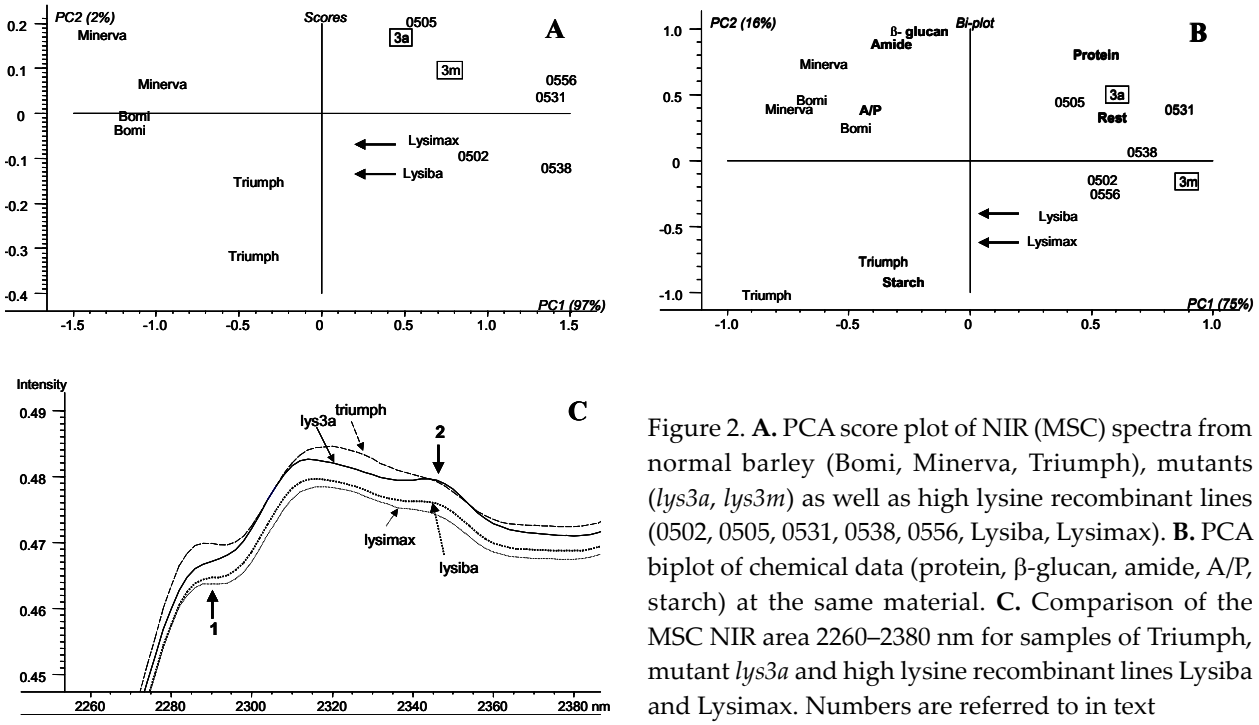


Figure 2. **A.** PCA score plot of NIR (MSC) spectra from normal barley (Bomi, Minerva, Triumph), mutants (*lys3a*, *lys3m*) as well as high lysine recombinant lines (0502, 0505, 0531, 0538, 0556, Lysiba, Lysimax). **B.** PCA biplot of chemical data (protein, β-glucan, amide, A/P, starch) at the same material. **C.** Comparison of the MSC NIR area 2260–2380 nm for samples of Triumph, mutant *lys3a* and high lysine recombinant lines Lysiba and Lysimax. Numbers are referred to in text

barleys. This results in improved starch content from about 48 to 53% which is equal to normal feed barleys. A collection of the original mutants, normal barley lines and yield-improved crosses was evaluated by NIR. A section of this data bank from material grown in the field in 1991 is demonstrated in a PCA (Figure 2A) which shows the pattern of the 14 NIR spectra 400–2500 nm characteristic of six normal barleys, two mutants and seven crosses (generation 5–10).

It is clearly seen that the samples divide into two clusters: one with normal barleys (Bomi, Minerva and Triumph) to the left, and the original *lys3* mutants (squared) *lys3a* and *lys3m* and their high-lysine segregates to the right. Lysiba and Lysimax are semi-commercial (registered but not widely grown) high-lysine varieties with verified improved yield (about 110–115% of the original Risø 1508 mutant). The lines Lysiba and Lysimax are thus *a priori* defined as positive selections. They are situated in the NIR PCA closer to the normal samples than for example lines 502, 531 and 538, which are negative selections confirmed by their locations above and to the far right in the PCA (Figure 2A). If an unsupervised PCA is performed on chemical data (protein, amide, A/P, starch, β-glucan and rest = (100 – (protein + starch + β-glucan)), almost the same pattern (Figure 2B) as with NIR spectra (Figure 2A) is seen, indicating the physical-chemical basis of the NIR measurements. The PCA bi-plot of chemical data in Figure 2B shows that samples of Triumph, which are known to have a very high starch content, are located close to the starch variable, indicating that these samples are highly influenced by this variable. Opposite, still to the left, are Minerva and Bomi samples highly influenced by amide, A/P and β-glucan. To the very right the mutants *lys3a* and *lys3m* are placed with the high-lysine segregates and together with the variables protein and rest. The samples of Lysiba and Lysimax (Group 1) are located closer to the normal varieties than the others, indicating a positive selection (with regard to starch) as in the PCA with the NIR spectra. Samples 502 and 556 (Group 2) are located close together, whereas the negatively selected samples 505, 531 and 538 (Group 3) are placed in between the mutants.

When comparing the chemical data of these four groups one can easily (see Table 2) that Group 1 with Lysiba and Lysimax has protein and starch content closer to normal varieties. The amide content and A/P index are, however, more closely related to the mutants. This means that they still are likely to have a content of essential amino acids as high as the original high-lysine mutants combined with an increase in starch (mean) from 48.7% to 52.6%. Group 2 is intermediate in starch content with reduced protein content, while Group 3 has as high protein content as the original mutants with only a slight increase in starch. The wavelength area used in earlier investigations (2260–2380 nm) is compared for spectra from samples of the normal barley Triumph, *lys3a* and from the positive selected lines Lysiba and Lysimax. From Figure 2C it is seen that the curve forms for Triumph and *lys3a* are very different, and that those of Lysiba and Lysimax are intermediate between *lys3a* and normal barley. The improved high-lysine lines have adopted some characteristics from the normal curve form. In the wavelength area 2285 nm to 2295 nm (arrow marked 1) the Triumph curve form is flat parallel to the *x*-axis while the original *lys3a* mutant curve form is steeper. As also seen in Figure 2C the improved crossbred lines carrying the *lys3a* gene, Lysiba and especially Lysimax are approaching the curve form of the reference barley variety Triumph.

It is also seen that the bulb in *lys3a* at the arrow marked 2 (Figure 2C) is much reduced in Lysiba and especially Lysimax indicating a more normal barley state. This area correlates to unsaturated fat (2347 nm) and cellulose (2352 nm). The content of these analytes should be tested further to detect the reasons why the samples separate from each other. A lower content of unsaturated fat and cellulose (fibre) in Lysiba and Lysimax should be expected.

Data breeding – NIR spectroscopy and chemometrics in breeding for improved physical-chemical composition as whole integrated genetic complex

We first introduced the concept of "data breeding" at the IBGS – VIII meeting in Adelaide in 2000 (Munck *et al*. 2000). As seen in the PCA of the whole NIR spectra in Figure 2A it is possible by crossbreeding and selection to move the *lys3a* segregants Lysimax and Lysiba in the direction of the arrow towards normal barley (Triumph), signifying an improved starch content verified by the PCA for the chemical analyses in Figure 2B and Table 2. Near Infrared Transmission (NIT) spectroscopy is today used as a "blackbox" tech-

Table 2. Average and standard deviations of chemical data for the five groups from the *lys3* breeding experiment

	Normal (n = 6)	Group 1	Group 2	Group 3	Group 4
Protein (P)	11.3 ± 0.4	11.7 ± 0.1	11.7 ± 0.1	12.6 ± 0.2	12.5 ± 0.2
Amide (A)	0.28 ± 0.03	0.21 ± 0.007	0.21 ± 0.007	0.22 ± 0.02	0.23
A/P	15.5 ± 0.9	11.0 ± 0.3	10.9 ± 0.4	10.7 ± 0.8	11.4
Starch	54.6 ± 2.5	52.6 ± 0.5	50.0 ± 0.1	49.4 ± 1.5	48.7 ± 0.2
ß-glucan	4.7 ± 1.1	3.1 ± 0.1	3.1 ± 0.2	3.1 ± 0.3	2.8 ± 0.5
Rest (100 – P + S + BG)	29.5 ± 1.8	32.7 ± 0.5	35.3 ± 0.3	34.9 ± 1.8	36.1 ± 0.5

Group 1 = Lysiba, Lysimax; Group 2 = 502, 556; Group 3 = 505, 531, 538; Group 4 = *lys3a*, *lys3m*
BG = (1→3,1→4)-β-glucan

nology in plant breeding by analysing for specific chemical parameters such as protein and water by a closed chemometric prediction model given by the instrument manufacturer. By introducing chemometrics to be used directly by the plant breeders the closed box can be opened so that they themselves can make their own NIT barley seed classifications and correlation models tailored to their own materials and needs. This is demonstrated for malting barley in these proceedings by Møller and Munck (2004).

We conclude that multivariate statistical methods (PCA) are indispensable for the classification of original barley endosperm mutants and their recombinants with standard varieties. Classic statistics of variance cannot handle large horizontal data sets where the samples have 1050 wavelength variables in this case. Thus different mathematical models should be used for the gamete and zygote levels of biological organisation (Munck 2003) where classic probability statistics effective on the gene recombination level should be complemented by pattern recognition data analysis (chemometrics) for analysing gene expression by NIR at the phenotype level.

References

Doll H. (1983): Barley seed proteins and possibilities for their improvement. In: Gottschalk W., Muller H.P. (eds): Seed Proteins. Martinus Nijhoff/Dr. W. Junk Publishers, The Hague: 205–223.

Greber B., Waite D, Fahy B., Hylton C., Parker M., Laurie D., Smith A.M., Denyer K. (2000): Use of barley mutants to understand starch synthesis. In: Proc. 8th Int. Barley Genetics Symp., 22–27 October 2000, Vol. I: 196–198.

Jacobsen S., Søndergaard I., Møller B., Desler T., Munck L. (2005): The barley endosperm as a data interface for expression of genes and gene combinations at different levels of biological organisation explored through pattern-recognition data evaluation. Journal of Cereal Science (accepted).

Munck L. (1992): The case of high lysine barley. In: Shewry P.E. (ed.): Barley: Genetics, Biochemistry, Molecular Biology and Biotechnology. CAB International, Wallingford, 573–602.

Munck L. (2003): Detecting diversity – a new holistic, exploratory approach bridging the geneotype and phenotype. In: Bothmer R. von, Hintum T.J.L. van, Knupffer H., Sato K. (eds): Diversity in Barley (*Hordeum vulgare*). Elsevier Science B.V., Amsterdam, Chapter 11: 227–245.

Munck L., Nielsen J., Møller B. (2000): From plant breeding to data breeding – How new multivariate screening methods could bridge the information gap between the genotype and the phenotype. In: Proc. 8th Int. Barley Genetic Symp., Adelaide, 22–27 October 2000, Vol. I: 179–182.

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.

Munck L., Møller B., Jacobsen S., Søndergaard I. (2004): Spectral multivariate indicators for mutant genes evaluated by chemometrics reveal a new mechanism for substituting starch with (1→3,1→4)β-glucan in barley. Journal of Cereal Science, **40**: 213–222.

Møller B., Munck L. (2004): A new two dimensional germinative classification for malting barley quality based on separate estimates for vigour and viability. Czech Journal of Genetics and Plant Breeding, **40**: 102–108.

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

Received for publication July 1, 2004
Accepted after corrections July 26, 2005

Abstrakt

Munck L., Møller B. (2005): **Užití analýzy hlavních komponent infračervených spekter pro charakterizaci mutací endospermu a ve šlechtění ječmene na zlepšenou jakost.** Czech J. Genet. Plant Breed., **41**: 89–95.

NIR, nyní široce používaná pro kontrolu jakostních parametrů, umožňuje získat přesný fyzikálně-chemický „fingerprint" endospermu ječmene. Spektroskopické „fingerprinty" semen ječmene mohou být interpretovány pomocí multivariační analýzy metodou hlavních komponent (PCA) využitelné pro klasifikaci a metodou parciální regresní analýzy nejmenších čtverců (PLSR) pro analýzu korelačních vztahů. PCA klasifikace metodou NIR umožnila diferenciaci mezi mutanty a alelami v lokusech *lys3* a *lys5*. Tato metoda může být také běžně použita ve šlechtění ječmene pro selekci na jakost, speciálně např. pro zvýšení obsahu škrobu a snížení obsahu vlákniny. To lze provádět přímo jako selekci rekombinací s přihlédnutím k pozici kontrol s vysokým obsahem škrobu. Na základě klasifikace pomocí NIR spekter byly dvě alely lokusu *lys5* charakterizovány jako nová kategorie indikující metabolický vztah mezi obsahy škrobu a β-glukanu.

Klíčová slova: multivariační analýza dat; NIR spektroskopie; mutanty ječmene; fyzikálně-chemický fingerprint

Corresponding author:

Prof. Lars Munck, The Royal Veterinary and Agricultural University, Department of Food Science, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark
tel: + 45 35 283 507, fax: + 45 35 283 265, e-mail: lmu@kvl.dk