Near infrared reflectance spectroscopy and computer graphics visualises unique genotype specific physical-chemical patterns from barley endosperms

Moller B., Norgaard L., Munck L.

in

Molina-Cano J.L. (ed.), Christou P. (ed.), Graner A. (ed.), Hammer K. (ed.), Jouve N. (ed.), Keller B. (ed.), Lasa J.M. (ed.), Powell W. (ed.), Royo C. (ed.), Shewry P. (ed.), Stanca A.M. (ed.).
Cereal science and technology for feeding ten billion people: genomics era and beyond

Zaragoza : CIHEAM / IRTA
Options Méditerranéennes : Série A. Séminaires Méditerranéens; n. 81

2008
pages 253-259

Near infrared reflectance spectroscopy and computer graphics visualises unique genotype specific physical-chemical patterns from barley endosperms

B. Møller, L. Nørgaard and L. Munck
Spectroscopy & Chemometrics Group, Quality & Technology, Department of Food Science,
The Royal Veterinary and Agricultural University, Rolighedsvej 30,
DK-1958 Frederiksberg C, Denmark
lmu@kvl.dk

SUMMARY – Near Infrared instruments today are conveniently used by plant breeders and industry as spectral "multimeters" for instant analysis of a range of chemical components. These are based on multivariate computer calibrations by chemometrics (e.g. Partial Least Squares Regression - PLSR - and neural nets) facilitated by the instrument suppliers. Here we will demonstrate in barley how the near infrared spectra that are routinely measured for specific chemical screening by breeders can now be exploited for chemical-physical characterisation of whole samples or genotypes. From near infrared data in a Principal Component Analysis (PCA) score plot the breeder can now directly select cereal seed genotypes with improved chemical-physical quality by "data breeding". This is achieved by selecting recombinants which are positioned close to high quality control genotypes in the score plot. Differences in local spectra can be deduced by PLSR to physical/chemical components by destructive analyses. Mathematical data compression models are useful. However, they are all destructive to the finely tuned, highly reproducible spectral patterns that originate from the self-organised biological network represented by the developing barley endosperm. The evaluation of Near Infrared Reflectance Spectroscopy (NIRS) data must therefore include a final direct visual inspection of spectra to trace reproducible patterns of biological significance from individual phenotypes. The "LatentiX" graphic data interface that is demonstrated at the symposium is constructed to visualise spectral areas correlated to specific chemical components and genotypes. It is available at www.latentix.com.

Genotype Classification by Near Infrared Spectroscopy and Principal Component Analysis

Breeding for quality implies access to inexpensive and reliable screening methods for chemical and physical variables. The fact that near infrared data in the range of 800-2500nm after calibration to classical methods (Williams and Norris, 2001) can be used as a "multimeter" for a wide range of parameters substitutes at once tens of analyses previously made using 10 different instruments with just one measurement plus calibration checks. The required software, calibrated to the classical analyses, is usually supplied by the instrument manufacturer and based on chemometric pattern recognition data models (Martens and Næs, 1989) such as neural nets and PLSR. Today, plant breeders and the cereal industry routinely utilise Near Infrared Transmission (NIT) spectrometers for estimation of chemical composition in substances such as water, protein, starch and malt extract. The instrument directly measures the intact seed sample in the range 800-1050 nm in seconds. In this investigation we use a Visual - Near Infrared Reflectance (NIR) Instrument 400-2500 nm for more detailed information. It is not yet acknowledged by plant breeders outside the cereal laboratories that a PCA classification on seed NIT and NIR spectra is an efficient tool in breeding for quality and for characterisation of genetic and technological quality complexes. The recent textbook on Near Infrared Analytic Technology edited by Williams and Norris in 2001 just mentions the PCA option on a few pages without reference to plant breeding. However, Campell *et al*., 2000 demonstrated that NIT data evaluated by PCA can classify a range of single and double mutant maize endosperm genotypes. Discriminate PLSR was used by Wang *et al*., 1999 to predict the number of dominant R alleles in single wheat kernels by VIS-NIRS and by Delwiche *et al*., 1999 to identify different wheat-rye translocation lines by NIRS. We demonstrated the usefulness of PCA on NIRS for classification of genetic and environmental differences in a barley endosperm mutant material (Munck *et al.,* 2001).

We were surprised by the genotype specific patterns of the spectral NIRS patterns behind the

different positions of the samples in a PCA score plot. In the following we will summarise our recent research (Munck, 2005, 2003; Munck *et al*., 2004; Jacobsen *et al*., 2004) using unpublished examples.

Materials and methods

Three data sets of endosperm mutant genotypes and normal barley controls (Munck *et al*., 2004) are introduced with spectroscopic and chemical analyses. The first data set (Figs 1-3, 5) consists of 23 samples of 20 barley genotypes grown in the field in the year 2000. The lines are classified in Fig. 2 as normal (N), protein mutants (P), which are 20-45% increased in lysine and moderately decreased in starch (5-10%), and carbohydrate mutants (C) low or very low in starch (10-40% less than the control) and with a moderate increase in lysine (5-10%). The P mutant Risø genotypes are the alleles *lys3a* (mutant1508), *lys3b* (mutant 18), *lys3c* (mutant 19) and *lys4d* (mutant 8). Mutant *lys3m* induced in Minerva originates from Carlsberg. Lysimax and Lysiba are starch and yield improved recombinants from crosses with *lys3a* and normal barley from Carlsberg. The C mutants are the Risø mutants, mutant 16 and *lys5f* (mutant 13) in Bomi and *lys5g* (mutant 29) in Carlsberg II. Mutants 95 and 449 are in Perga of Italian origin. w1 (line 1201) and w2 (line 841878) of unknown origin were imported to the Carlsberg collection assigned as waxy mutants. Only w2 was waxy (amylose 4.2%). The second data set (Fig. 4, Table 1) consists of four genotypes –Bomi and mutants *lys5f, lys5g* and *lys3a* grown in the field (from data set 1) and in the greenhouse. The third data set (Figs 6-8, Table 2 from Møller 2004b) displays 15 barley lines grown in the field; the N (normal) lines Bomi, Minerva and Triumph, the P (protein) mutants *lys3a, lys3m* and the breeding lines between *lys3a* and normal barleys, Lysimax and Lysimax (positive seed quality selection) and the unselected l*ys3a* breeding lines 502,505,531,538 and 556. The chemical and Near Infrared Reflectance (NIR) spectral analysis (on milled flour 0.5 mm sieve) was carried out by a Foss-NIRSystems (USA) 6500 instrument as described by Munck *et al*., 2004. The raw spectra were multiplicative scatter corrected (MSC) and presented as log 1/R intensity. Chemometric pattern recognition analysis was performed using Principal Component Analysis (PCA) for classification and Partial Least Squares Regression (PLSR) for prediction according to Martens and Næs (1989).

Recognising sample and genotype specific NIRS patterns in barley

The NIR spectra from the 23 barley seed samples from Material 1 in Fig. 1 depict 1400 wavelength variables with a seemingly narrow variation between samples in absorption value MSC log1/R. Classical statistics of variance based on distributional assumptions cannot extract information from a whole intercorrelated data matrix with thousands of wavelength variables per sample. For this purpose, there is a need for self-calibrating multivariate chemometric data models such as PCA for classification and PLSR for prediction, both based on latent variables. The PCA (PC1 to PC2) in Fig. 2 classifies the spectra in Fig. 1 into three distinct populations: normal (N) barley, endosperm mutants with a high lysine percentage in protein (P) and carbohydrate mutants (C). The C barley lines such as Risø mutant 16, *lys5g* and *lys5f*, which are mutations in the AGP-ase mechanism, have been used by the biochemists to study starch synthesis. It was therefore surprising when we found (Munck *et al*., 2004) that these mutants and three others –mutant 95, mutant 449 and the w1 (line 1201), all in the C cluster (Fig. 2)– compensated the loss in starch by overproducing β-glucan (BG). The mean BG value for the C (12.3%), P (3.7%) and N (4.7%) classes are marked in Fig. 2. Obviously, there is a strong pleiotropic regulative effect where mutations in starch metabolism may channel glucose from α- to β-glucan production. Normally, PCA classifications, which are useful for an overview of large (spectral) data sets, are published in literature without inspecting the underlying data structure. We generated the mean spectra for the C, P and N genotypes and visually screened the whole spectra. Marked differences were found in the narrow area 2260-2360 nm (Fig. 3, marked "a" in Fig. 1) of chemical and genetic significance. The near infrared transmission (NIT) and reflectance (NIR) spectra represent in principle a physical-chemical fingerprint containing repetitive information on the propensity of chemical bonds. A trained spectroscopist can from the first, second or MSC derivatives of the log1/R NIRS data directly explore specific chemical differences between samples and deduce destructive analyses for verification. The P and C genotypes have a characteristic bulb in the area from 2336 to 2352 nm which in spectroscopic literature is assigned to cellulose (2336 and 2352 nm) and fat (2347 nm).

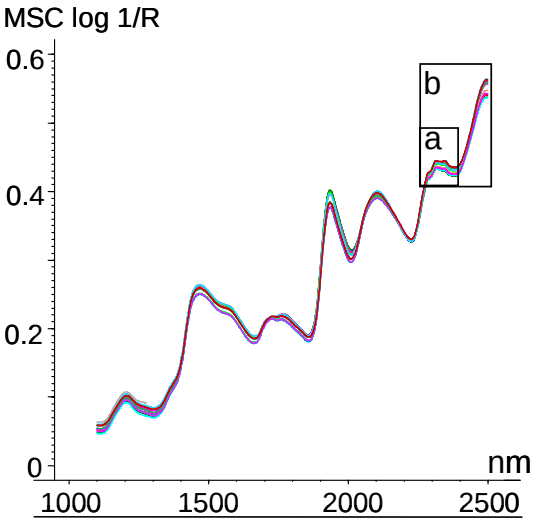

Fig. 1. MSC log 1/R NIRS Spectra 1100-2500 nm of the 23 barley samples in Material 1.

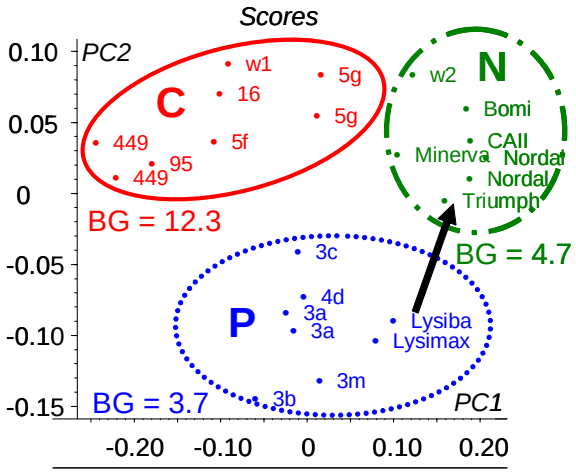

Fig. 2. PCA score plot of the spectra in Fig. 1; see discussion in text.

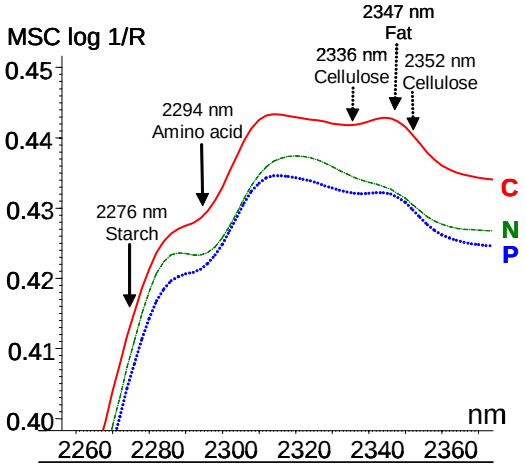

Fig. 3. Mean spectra 2260-2360 nm of clusters N, P and C from the PCA in Fig 2.

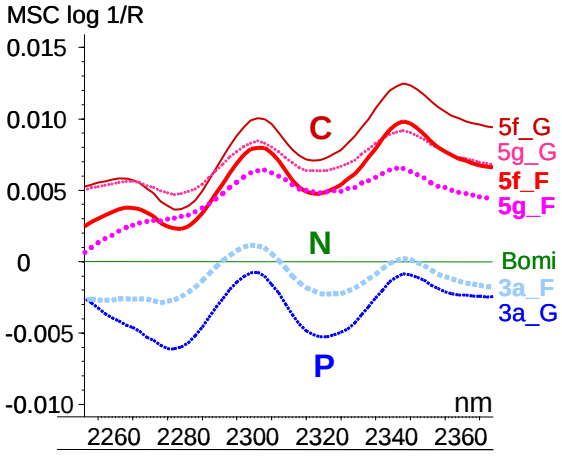

Fig. 4. Differential NIRS mutant spectra 2260-2360 nm of *lys5f*, *lys5g*, *lys3a* to Bomi grown in greenhouse (G) and field (F). Chemical composition in Table 1.

Table 1. Chemical composition (% d.m) of barley mutants and isogenic Bomi control grown in greenhouse (G) and field (F) shown as spectra in Fig. 4

		DM%					
		BG	Starch	Protein	Amide	A/P	Fat
Bomi_G	N	6.8	48.8	14.6	0.38	16.2	1.7
Bomi_F	N	4.9	53.6	11.5	0.29	15.8	1.9
3a_G	P	3.6	41.6	16.6	0.29	10.9	3.5
3a_F	P	3.1	48.5	12.7	0.23	11.4	2.6
5g_G	C	13.5	44.7	16.1	0.39	15.1	2.3
5g_F	C	8.9	47.0	11.8	0.26	13.8	2.3
5f_G	C	20.0	30.5	15.7	0.37	14.7	3.7
5f_F	C	16.5	33.0	14.5	0.31	13.4	3.8

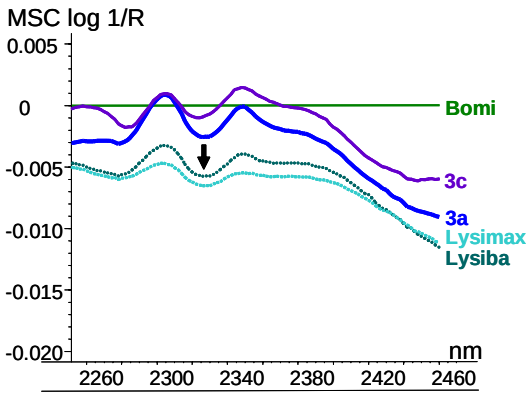

Fig. 5. Differential NIRS mutant spectra 2260-2480 nm to Bomi for lys3a, lys3b, Lysimax and Lysiba.

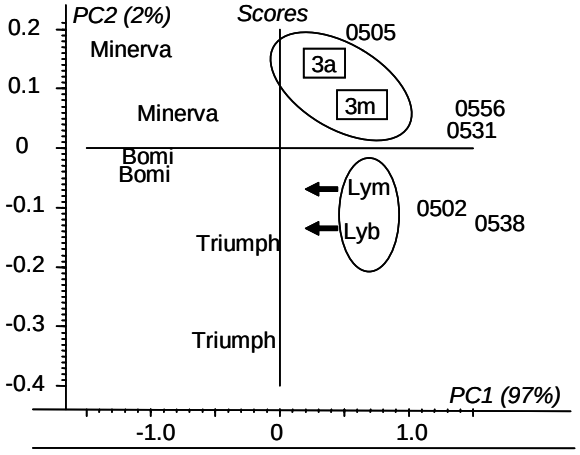

Fig. 6. PCA score plot for 15 barley visual-NIRS spectra (400-2500 nm) MSC log 1/R 400-2500nm demonstrating a case of "data breeding" discussed in text.

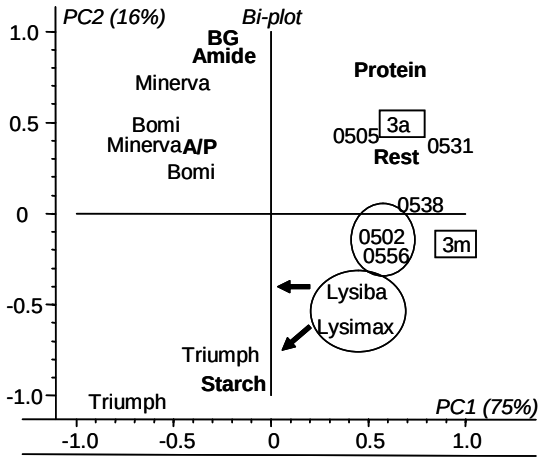

Fig. 7. PCA bi plot on the chemical composi-tion of 15 barley varieties displayed in Table 2.

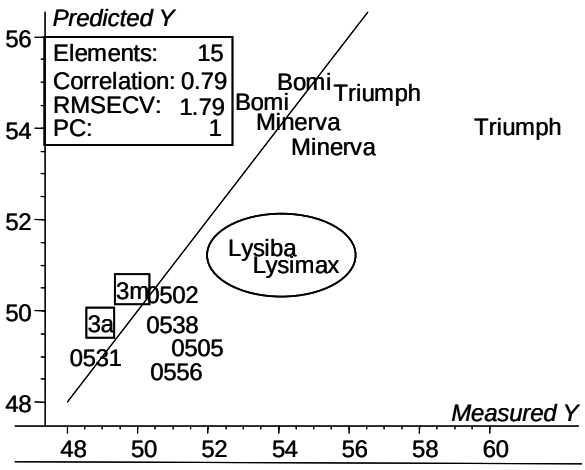

Fig. 8. PLSR correlation plot for prediction of starch in the material in Table 2 by visual-NIRS as ordinate and the chemical starch analysis as abscissa.

Table 2. Chemical composition of the 6 normal and 9 *lys3a* genotypes presented in Figs 6, 7 and 8

	Normal	Group 1	Group 2	Group 3	Group 4
	n = 6	Lysiba, Lysimax	502, 556	505, 531, 538	*lys3a, lys3m*
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
β-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

The substantial increases of the fat components anticipated from the spectral patterns of *lys3a* and *lys5* barleys are verified by chemical analyses in Table 1. Further studies (Jacobsen *et al*., 2004; Munck, 2005) demonstrate that the spectral patterns do not only have a chemical interpretation, but also a genetic significance as a phenomenological trait expressing the whole active genome as a pattern of chemical bonds represented by the spectral phenome (Munck *et al*., 2004; Munck, 2005).

The reproducibility, fine-tuning and informative capacity of NIRS spectra are indeed impressive. The MSC log1/R absorption range is 0.04 units for classification of C, P and N barleys in Fig. 3. However, the range needed is 100 times less for classifying the C versus P+N groups in the 1890-1920 nm area for dry matter (DM) content within the narrow response of 89-93% d.m. (Munck, 2005). A high BG content of the C group conditions a mean difference in DM of 1.5% between these groups. The precision of NIRS allows individual barley samples (genotypes) to be differentiated by their spectral patterns. This is demonstrated by the spectra in Fig. 4 with chemical evaluation shown in Table 1. Four samples representing three mutants *lys3a, lys5f, lys5g* and the normal control Bomi grown in the field are selected from Material 1 and compared with the corresponding genotypes grown in greenhouse. Figure 4 displays the differential spectra where the mutant spectra are subtracted from those of the Bomi control. A well-conserved genetic pattern is demonstrated in the spectral area 2260-2360 nm for the two different environments. There is some offset and a minor effect on the spectral form due to environment. Bomi is near isogenic for the *lys3a* and *lys5f* mutants. The differential spectrum to Bomi of these mutants constitutes a spectral representation of pleiotropy involving all expression effects of the mutant on the level of chemical bonds (Table 1) in the endosperm (Munck, 2005; Jacobsen *et al*., 2004). While protein alone has a low power in discriminating between the genotypes (Table 1), the amide-to-protein (A/P) index clearly separates the high-lysine P mutant *lys3a* from the others. On the other hand, the BG-compensated starch mutants (C) show a very high level of BG (16.5-20% DM in *lys5f*) when Bomi is 4.9-6.8 % DM and *lys3a* is reduced to 3.1-3.6%. There is a corresponding reduction in starch from 48.8-53.6% DM in Bomi down to 30.5-33.0 % DM in *lys5f.* It is thus the pleiotropic differences in expression of chemical composition between the mutants in Table 1 that explain the unique spectral patterns of the same samples displayed in Fig. 4.

"Data breeding" for complex quality traits by NIRS selecting improved segregants from a PCA score plot

We will now test if an improvement in breeding for plump starch rich seeds in high-lysine *lys3a* barley at Carlsberg 1973-1988 can be followed by NIRS technology to be further exploited for other purposes in plant breeding. In the PCA score plot from Material 1, the improved *lys3a* genotypes Lysiba with starch (S 52.2%) and Lysimax (S 52.9%) are classified in between the original *lys3a* (S 48.5%) mutant and the normal barley Triumph high in starch (S 58.5%), indicating a change in chemical composition. The differential spectra 2260 to 2360 nm to Bomi (S 53.5%) of the improved Lysiba and Lysimax genotypes and of the *lys3a* (BG 3.1%) and *lys3c* (BG 6.4%) allele are presented in Fig. 5. The spectral differences between the alleles *lys3a* and *lys3b* are mainly due to the difference in BG. The spectra from the starch-improved *lys3a* lines are moved downwards to the baseline and the mutant characteristics are flattened out. The area between the *lys3a* spectrum and the Lysiba/Lysimax recombinants marked by the arrow gives a spectral representation of 15 years of

breeding work to improve seed quality. Table 2 outlines the chemical composition (six variables) of material 3 with falling starch content from 54.6 to 48.7% consisting of: normal barley, improved *lys3a* breeding lines (group 1), unselected recombinants (groups 2 and 3) and original mutants (group 4). The PCA biplot in Fig. 7 gives a convenient overview of how the chemical analyses influence genotype classification that is comparable with the corresponding VIS-NIRS (400-2500 nm) PCA classification plot in Fig. 6. In the biplot in Fig. 7 the variable "Starch" is positioned near Triumph, indicating a high level of starch in this cultivar. The move in both PCA´s (Figs 6 and 7) of the Lysiba and Lysimax improved genotype from the position of the original low starch *lys3a* mutant towards the high-starch variety Triumph is clearly demonstrated in both PCA´s. The NIRS and chemical data sets are combined in the PLSR starch prediction plot in Fig. 8. This is how NIR and NIT spectroscopy are utilised today by plant breeders for chemical prediction. But because near infrared spectroscopy gives a total estimate of the chemical composition of a barley sample in a PCA score plot it is possible empirically, by comparison to a high-quality genotype, to evaluate and to select *the whole* expression of the genotype on the spectral level in a cross-breeding program by "data breeding". It is thus possible to represent a complex quality trait such as nutritional value (described here) or malt quality (Møller 2004a) *as a whole* by NIR and NIT spectroscopy. This can be done without chemical analysis, except for the evaluation of the final varieties. It is clear that NIR and NIT spectroscopy combined with PCA and PLSR data analysis (chemometrics) is a revolution in cost-effective breeding for quality. It reflects in a reproducible way the above-described changes in chemical composition down to 1-2 percentage points of each component by PLSR and summarises by PCA and visual inspection of spectral intervals subtle differences in quality characteristics *as a whole* for each genotype and sample representing patterns of chemical bonds.

In genetics and plant breeding there is a need for a new multivariate way of thinking (Munck, 2005) in re-defining the biological individual, or zygote*. Now the self-organising (endosperm) tissue can be considered as a response interface on the level of chemical bonds for the genome*. It is, as demonstrated here, read as a pattern by spectroscopy, classified by multivariate analysis and interpreted by genetics, chemistry and technology. Classical genetic variance statistics is focused on individual genes and traits, assuming more or less free distribution. The current Quantitative Trait Loci (QTL) analysis aims at revealing complex genome-phenome-quality relationships. It combines trait and genome information and has until recently employed the traditional analysis of variance, despite the fact that variables in most gene, trait and quality complexes are strongly dependent on each other. This is why QTL analysis has such a mixed reputation. Using multivariate pattern recognition analysis PCA and PLSR can overcome this problem, as recently demonstrated by Bjørnstad *et al*. (2004) with PLSR. A new high-precision QTL analysis on the spectral phenome level for use by plant breeders will be possible when comparing (by PCA) and combining (by PLSR) NIRS data with RFLP data. It is likely that NIR and NIT fingerprinting is able to function as a stand-alone analysis, if genetically (by DNA), chemically and technologically defined controls are provided.

References

Bjørnstad, Å., Westad, F. and Martens, H. (2004). Analysis of genetic marker-phenotype relationships by jack-knifed partial least squares regression (PLSR). *Hereditas,* 141: 149-165.

Campbell, M.R., Sykes, J. and Glover, D.V. (2000). Classification of Single and Double-Mutant Corn Endosperm Genotypes by Near-Infrared Transmittance Spectroscopy. *Cereal Chemistry,* 77(6): 774-778.

Delwiche, S.R., Graybosch, R.A. and Peterson, J. (1999). Identification of Wheat Lines Possessing the 1AL.1RS or 1BL.1RS Wheat-Rye Translocation by Near-Infrared Reflectance Spectroscopy. *Cereal Chemistry,* 76(2): 255-260.

Jacobsen, S., Søndergaard, I., Møller, B., Desler, T. and Munck, L. (2005). A chemometric evaluation of the underlying physical and chemical patterns that support near Infrared spectroscopy of barley seeds as a tool for explorative classification of endosperm genes and gene combinations. *Journal of Cereal Science,* 42(3): 281-299.

Munck, L. (2005). *The Revolutionary Aspect of Chemometric Technology. The Universe and Biological Cells as Computers. A Plea for Cognitive Flexibility in Mathematical Modelling*. lmu@kvl.dk. Samfundslitteratur Grafik. The Royal Veterinary and Agricultural University, Denmark, p. 352.

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

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

Møller, B. (2004a). Near infrared spectra of barley of malting grade represent a physical-chemical fingerprint of the sample which is able to predict germinative vigour in a multivariate data evaluation model. *Journal Institute of Brewing,* 110(1): 18-33.

Møller, B. (2004b). *Screening Analyses for Quality Criteria in Barley*. Ph.D Thesis. Department of Food Science, The Royal Veterinary and Agricultural University, Denmark.

Wang, D., Dowell, F.E. and Lacey, R.E. (1999). Predicting the Number of Dominant R Alleles in Single Wheat Kernels Using Visitble and Near Infrared Reflectance Spectra. *Cereal Chemistry,* 76(1): 6-8.

Williams, P. and Norris, K. (eds) (2001). *Near-Infrared Technology in the Agriculture and Food Industries*. American Association of Cereal Chemists, St Paul. Minn., U.S.A.