ELSEVIER

Journal of Cereal Science 42 (2005) 281–299

Journal of CEREAL SCIENCE

www.elsevier.com/locate/jnlabr/yjcrs

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

S. Jacobsen[a], I. Søndergaard[b], B. Møller[a], T. Desler[c], L. Munck[a,*]

[a]*Spectroscopy and Chemometrics Group, Food Quality and Technology, Department of Food Science, The Royal Veterinary and Agricultural University (KVL), Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark*

[b]*Biochemistry and Nutrition Group, BioCentrum-DTU, The Technical University of Denmark, Søltoft Plads, Building 224, DK-2800 Kgs. Lyngby, Denmark*

[c]*Slotsherrensvej 48, DK-2720 Vanløse, Denmark*

Received 5 March 2004; revised 11 February 2005; accepted 20 February 2005

Abstract

Near infrared spectroscopic (NIR; 1100–2500 nm), chemical and genetic data were combined to study the pleiotropic secondary effects of mutant genes on milled samples in a barley seed model. NIR and chemical data were both effective in classifying gene and gene combinations by Principal Component Analysis (PCA). Risø mutants R-13, R-29 high (1→3, 1→4)-β-glucan, low starch and R-1508 (high lysine, reduced starch), near isogeneic controls and normal lines and recombinants were studied. Based on proteome analysis results, six anti-microbial proteins were followed during endosperm development revealing pleiotropic gene effects in expression timing that supporting the gene classification. To verify that NIR spectroscopy data represents a physio–chemical fingerprint of the barley seed, physical and chemical spectral components were partially separated by Multiple Scatter Correction and their genetic classification ability verified. Wavelength bands with known water binding and (1→3, 1→4)-β-glucan assignments were successfully predicted by partial least squares regression giving insight into how NIR-data works in classification. Highly reproducible gene-specific, covariate, pleiotropic classification patterns from NIR and chemical data were demonstrated in PCAs and by visual inspection of NIR spectra. Thus PCA classification of NIR-data gives the classical genetic concept, 'pleiotropy', a new operational definition as a fingerprint from a spectroscopic representation of the phenome carrying genetic, physical and chemical information. It is concluded that barley seed phenotyping by NIR and chemometrics is a new, reliable tool for characterising the pleiotropic effects of mutant gene combinations and other genotypes in selecting barley for quality in plant breeding.

Keywords: Near infrared spectroscopy; Barley mutants; Physico-chemical fingerprinting; Chemometric classification; Phenome; Proteome

1. Introduction

1.1. An exploratory approach to classification of mutant genes and quality characteristics of barley

Near infrared spectroscopy (NIR) (Miller, 2001; Osborne et al., 1993; Williams and Norris, 2001) is an established non-destructive screening method used in plant breeding and in the cereal industry for prediction of a wide range of specific chemical (e.g. water, protein, starch) and physical (e.g. grain hardness) parameters from computerised calibrations using classical statistical and chemometric software. The NIR instrument functions as a 'multimeter' reflecting that NIR spectra contain a wide range of diverse physico-chemical information as demonstrated by the diversity of

Abbreviations: 2-DE; two-dimensional electrophoresis; A; amide % d.m.; A/P; index between A and P content; BG; (1→3, 1→4)-β-glucan; BG+S; sum of BG and starch % d.m.; DBC; dye binding capacity; d.m.; dry matter %; FT; Fourier transform; IR; infrared spectroscopy; MSC; Multiplicative Scatter Correction; NIR; Near Infrared Spectroscopy; P; protein % d.m.; PCA; Principal Component Analysis; PLSR; Partial Least Square Regression; S; starch % d.m.

* Corresponding author. Tel.: +45 35 28 33 58; fax: +45 35 28 32 45.
E-mail address: lmu@kvl.dk (L. Munck).
URLs: http://www.models.kvl.dk (L. Munck), http://www.biocentrum.dtu.dk (L. Munck).

0733-5210/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jcs.2005.02.008

applications in food and agriculture. Campbell et al. (2000) used NIR in transmission mode to classify maize starch mutants by Principal Component Analysis (PCA). The same technology has allowed analysis of the genetic diversity of barley gene banks by spectroscopic data mining (Munck, 2003). In this approach each individual sample is classified by its position in a PCA score plot of the spectral data (Munck, 2003; Munck et al., 2001; 2004; Munck and Møller, 2004) which according to spectroscopic theory (Miller, 2001) reflects the underlying chemistry and physics of the sample.

The exploratory principle (Munck et al., 1998) i.e. measuring first and generating the preliminary hypotheses afterwards, is an important complement to the classical analytical deductive strategy. Within limits, this NIR method allows explorative identification of new barley samples with deviating compositions by observation without prior hypotheses. Using this approach we have recently demonstrated (Munck et al., 2004) a PCA-based classification of NIR spectra of two barley mutants (Doll, 1983) with a reduced starch content and a compensating increase in (1→3, 1→4)-β-glucan (BG) content. In a separate experiment, four unrelated lines in our spectral data bank were identified independently by PCA to belong to the same cluster and their high BG content was subsequently verified. Specific spectral signatures arising from mutant genes in hundreds of barleys were empirically identified (Munck, 2003; Munck et al., 2001; 2004) with a high reproducibility between sample replicates within the surprisingly narrow range of 0.03 log 1/*R* absorption units, including multiplicative scatter corrections (MSC) e.g. from the NIR region 2260–2380 nm. These specific spectral changes were due to gene specific, pleiotropic changes in chemical composition when compared to near isogenic mother lines. Based on these results we hypothesised (Munck et al., 2001; 2004) that the endosperm phenotypes represented by their multivariate spectral (observational) and chemical (analytical) data, constitute two separate and complementary data sources reflecting the endosperm phenome (Watkins et al., 2001). Both data sources are suitable for genotype classification. In this study multivariate spectral fingerprints from these two data sources showing genetic and environmental effects are described. Here we show how they can be exploited to detect the physical and chemical nature of both known and unknown mutant genes, gene combinations and genotypes. As a side effect, major changes in gene products can be detected as exemplified by (1→3, 1→4)-β-glucan (Munck et al., 2004).

2. Experimental

2.1. Barley lines

Barleys carrying two mutant alleles with a small increase in lysine show the high BG-low starch trait (Munck et al., 2004) were included in the experiments: one extreme (BG: 16.5–19.8%)-*lys*5*f* (Risø mutant 13 in Bomi) and one less extreme (BG: 8.9–13.3%)-*lys*5*g* (Risø mutant 29 in Carlsberg II). They are classified together with the very high lysine mutant (48% increase)-*lys*3*a* (Risø mutant 1508 in Bomi). The double recessive recombinant *lys*3*a*5g was also included and also the *lys*3*a* cultivars, Lysimax, Lysiba and Piggy, bred for improved starch content and yield (Munck, 1992). All seeds were dried and equilibrated to laboratory relative humidity at ∼23 °C before refrigerated storage in plastic bags. Samples were milled and the water content determined shortly after spectra were recorded.

Three barley pools were used:

2.1.1. Pool 1

NIR measurements were obtained on flour from all samples, and dry matter % (d.m.), BG, protein and amide were determined.

A. A total of 26 samples of *lys*3*a* ($n=7$), *lys*5*g* ($n=8$), *lys*3*a*5*g* ($n=1$), *lys*5*f* ($n=4$) genotypes (mutants and segregants) and normal varieties ($n=6$) grown in field trials in 2000.
B. A total of 40 samples of *lys*3*a* ($n=7$), *lys*5*g* ($n=9$), *lys*3*a*5*g* ($n=2$), *lys*5*f* ($n=6$) genotypes (mutants and segregants) and normal varieties ($n=16$) grown in a greenhouse in 1999 and 2000. Average thousand kernel weights for the mutants and mother lines were 45.9 g for *lys*3*a*, 47.0 g for *lys*5*f*, 48.5 g for *lys*5*g*, 51.5 g for *lys*3*a*5*g*, 47.1 g for Carlsberg II and 55.6 g for Bomi.

2.1.2. Pool 2

Consisted of the original *lys*3*a*, *lys*5*g*, *lys*3*a*5*g* genotypes and the control varieties Bomi and Carlsberg II grown at the Carlsberg Research Laboratory farm and analysed by Desler (1987).

2.1.3. Pool 3

NIR measurements were made on flour from all samples in this pool, and in addition d.m., BG, protein, amide, lipid, starch, insoluble and soluble fibre and amino acid composition were determined. A total of 33 samples were analysed: *lys*3*a* ($n=4$), *lys*5*g* ($n=6$), *lys*3*a*5*g* ($n=2$), *lys*5*f* ($n=5$) genotypes (both mutants and segregants) and normal varieties ($n=16$) grown in field trials in 1991 ($n=21$) as well as in a greenhouse 1998–2000 ($n=12$). Due to lack of sample material the analytical data is not complete as indicated in the Figures and Tables and discussed in the text.

2.2. Two-dimensional gel electrophoresis

Two-dimensional electrophoresis was conducted according to Jacobsen et al. (2001).