高光谱技术检测单籽粒小麦粗蛋白含量探索

吴静珠,刘 倩,陈 岩,刘翠玲

(北京工商大学 计算机与信息工程学院 食品安全大数据技术北京市重点实验室,北京 100048)

摘 要:小麦蛋白质含量的性状遗传力较高,通过选择蛋白质含量高的籽粒母本可以达到优质育种的预期效果。研究采用高光谱成像技术结合化学计量学方法建立多籽粒小麦粗蛋白平均模型来实现单籽粒小麦粗蛋白含量的快速预测。实验采集 47 份小麦样本(每份 100 粒)的高光谱图像并提取平均光谱信息,通过联合区间偏最小二乘法筛选特征变量优化建立多籽粒小麦粗蛋白平均模型。模型的相关系数为 0.94,预测均方根误差为 0.28%,相对分析误差为 3.30。通过图像处理提取出待测单籽粒小麦的高光谱图像,应用平均模型预测单籽粒小麦在每个空间像素点的粗蛋白,取其平均值作为该粒麦种的最终粗蛋白含量。经验证,应用上述模型预测同一组样本的单籽粒小麦时,不同籽粒的小麦粗蛋白含量确实存在差异,但蛋白含量均围绕其所在样本的平均值浮动,因此反映出采用平均模型预测单籽粒小麦蛋白的准确性和基本可行性。该方法的研究可以为小麦育种过程中高蛋白籽粒麦种的优选提供一种新思路,推动小麦优质育种的发展。

关键词:高光谱图像; 联合区间偏最小二乘法; 单籽粒小麦; 粗蛋白; 育种

中图分类号:S512.1 **文献标志码**:A **DOI**:10.3788/IRLA201645.S123002

Prediction method of single wheat grain protein content based on hyperspectral image

Wu Jingzhu, Liu Qian, Chen Yan, Liu Cuiling

(Beijing Key Laboratory of Big Data Technology for Food Safety, School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China)

Abstract: The characteristics of wheat protein content has high heritability, so fine-quality breeding can be achieved by selecting the high-protein wheat seed. Combined with chemometric methods′ hyperspectral imaging technique was used to build the average model to achieve fast prediction of single wheat seed protein content. In the experiment, 47 unit wheat seed samples′ hyperspectral images were collected by GaiaChem-NIR system, and the average spectra was obtained by image process methods. Then, synergy interval partial least squares was applied to select the characteristic spectral regions to optimize the prediction model of wheat seed protein content. The optimal models′ determination coefficient is 0.94, the root mean square error of prediction is 0.28%, and the residual predictive deviation (RPD) is 3.30. Finally, the average model was applied to predict the protein content of each pixes of single wheat seed, and calculated the average as the single wheat grain protein content. The experimental results showed that different wheat grain′s protein content value predicted by the optimal model existed difference.

收稿日期:2016-01-15; **修订日期**:2016-02-25

基金项目:土壤植物机器系统技术国家重点实验室开放课题(2014-SKL-05);北京工商大学两科基金培育项目(19008001110)

作者简介:吴静珠(1979-),女,副教授,博士,主要研究领域为基于分子光谱及成像技术的农产品及食品检测。Email:pubwu@163.com

Ueanwhile, the prediction values varied around the average protein content of its sample, which reflected that the average model is accurate and feasible to predict single wheat grain's protein content. Therefore, the studied method provides a new way to select the high-protein wheat seed in the process of breeding, which can promote the development of wheat fine-quality breeding.

Key words: hyperspectral image; SiPLS; single wheat seed; protein; breeding

0 引 言

小麦是人类重要的蛋白质来源。蛋白质含量的高低是决定小麦营养品质和加工品质的重要因素，是小麦国际贸易和品质评价的基本指标。因此，提高籽粒蛋白质含量一直是高产优质小麦新品种选育的主要目标之一[1]。目前，国内外已有许多研究表明，小麦蛋白质含量的性状属于遗传性状且遗传力较高，如李世平[2]指出控制蛋白质含量的基因作用以加性效应为主，F1 代子粒蛋白质含量与双亲平均值高度相关，因此通过选择蛋白质含量高的籽粒母本可以提高后代籽粒蛋白质含量总水平，达到优质育种的预期效果。

如何快速、高效地测定单籽粒小麦蛋白质含量是小麦育种过程中亟待解决的问题。目前，用于小麦蛋白质含量测定的方法主要有国标中规定的凯氏定氮法和近红外光谱法。国标法要求至少研磨 200 g 样品并充分混匀后测得平均蛋白质含量，近红外光谱技术对于小麦等固体颗粒多采用大样品杯装样的测量方式，二者均只能实现对一定数量和质量的小麦样本的平均蛋白质含量的测定，无法测定单粒小麦的蛋白含量。但是同一品种内不同籽粒间的蛋白质含量是有较大差异的[3]，所以寻求一种可以测定单籽粒麦种蛋白质含量的方法对于实现小麦优质育种具有十分重要的现实意义。

高光谱成像技术是近年来出现的一种将光谱与成像科学相结合的无损检测新方法，既能获取被测样品在每个波段处的图像信息，又能获得样品每个空间像素点处的光谱信息，具有更高的分析检测潜质[4]。目前已在农产品的内外部品质检测，损伤识别以及农作物的生产信息获取等领域成为研究热点[5-7]。文中探索采用高光谱成像技术结合化学计量学方法优化建立小麦籽粒粗蛋白平均模型，通过应用粗蛋白平均模型来实现单籽粒小麦种子粗蛋白含量的快速预测，拟为育种过程中高蛋白籽粒麦种的挑选提供一种新思路。

1 材料与方法

1.1 试验材料

47 份对应不同品种的小麦样本由中国农业科学院作物科学研究所提供，置于 4 ℃左右冰箱冷藏。

1.2 蛋白质参考值的测定

每份样本的平均粗蛋白含量值参照 GB/T 5511—2008《谷类和豆类氮含量测定和粗蛋白质含量计算(凯氏法)》[8]测定。

1.3 高光谱图像采集与标定

选用北京卓立汉光仪器有限公司 GaiaSorter 高光谱分选仪采集小麦种子的近红外高光谱图像，示意图如图 1 所示。采集过程及仪器参数设定如下：每个小麦样本取 100 粒，平铺无重叠放置于样品台采集其高光谱图像，图像分辨率 320×256 像素点，光谱扫描范围 876~1 729 nm，曝光时间 25 ms，波段间隔 3.3 nm，波段数 256 个。

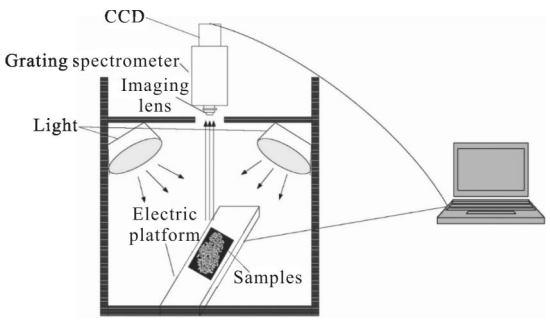

图 1 高光谱图像采集系统

Fig.1 Hyperspectral image system

由于光源的强度在各个波段下分布不均匀，样品的形状不规则以及摄像头中暗电流的存在，造成光源强度分布较弱的波段所获得的图像含有较大的噪音。因此，需要对所采集的高光谱图像按下式进行黑白标定：

$$I_{correction}=\frac{I_{raw}-I_{dark}}{I_{white}-I_{dark}} \quad (1)$$

式中：$I_{correction}$ 为校正后的光谱图像；I_{raw} 为原始光谱图像；I_{white} 为扫描反射率为 99%的标准白板得到的白板标定图像；I_{dark} 为关上光源，拧上镜头盖后采集的黑板标定图像。

1.4 数据处理

1.4.1 平均光谱提取

首先，为使所提取的光谱具有较强的代表性，选取样本全区域作为感兴趣区；然后在样本与背景区分明显的波段下(实验选取波段为 1 026.5 nm)利用最大方差自动取阈法[9]提取样本轮廓；最后依次在 256 个波段下提取样本轮廓范围内的反射率平均值构成该样本的光谱信息，如图 2 所示(以 13 号样本为例)。47 份小麦样本共获得 47 条平均反射光谱用于建模分析。

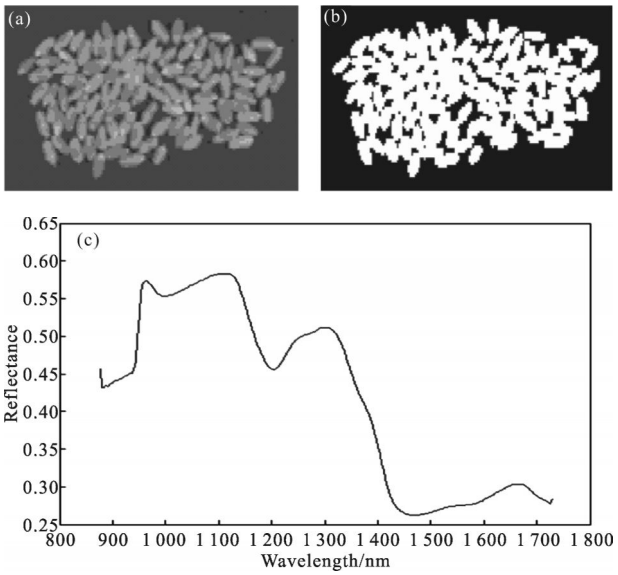

图 2 (a) 1 026.5 nm 波段的灰度图像，(b) 轮廓图像，(c) 提取的高光谱

Fig.2 (a) Grayscale image at 1 026.5 nm, (b) Outline image, (c) Extracted hyperspectra

高光谱仪在其测量临界区有较强的机器噪声，因此截去两端噪声严重波段，取 939~1 692 nm 范围内 215 个波段的高光谱进行分析。光谱提取过程通过 MATLAB 软件编程实现。

1.4.2 特征变量筛选

高光谱数据具有波段多、数据量大、冗余性强等特点，若直接对原始光谱数据进行建模可能会导致数据建模效率低、模型性能差，故需要对全光谱进行特征变量筛选。采用组合间隔偏最小二乘法[10](Synergy interval Partial Least Squares，SiPLS)筛选特征变量，即将全光谱等分为 n 子个区间后，按照排列组合 C_n^m 的思想依次联合其中 m 个区间建模，筛选相关系数最大且误差最小的一个组合区间作为特征变量。

1.4.3 模型建立与评价

对 SiPLS 方法筛选出的光谱区间，采用 5 折偏最小二乘法(Partial Least Squares，PLS)建立小麦蛋白质定量分析模型。根据模型的主成分数 nF，校正集相关系数 Rc，交叉校验均方根差 (Root Mean Square Rrror of Cross Validation，RMSECV)，预测集相关系数 Rp，预测均方根误差 (Root Mean Square Error of Prediction，RMSEP) 和相对分析误差(Residual Predictive Deviation，RPD)评价模型的预测精度和稳健性。

2 结果与分析

2.1 异常样本剔除和样本集划分

为避免异常样本对模型精度的干扰，采用蒙特卡洛异常样本剔除法剔除异常样本 1 个，对剩余 46 份样本按照 Kennard-Stone 法以 3∶1 比例划分，得到建模集样本 35 个，预测集样本 11 个，如表 1 所示。

表 1 样品集统计信息

Tab.1 Statistic information of samples

Data set	Samples	Min	Max	Mean	Standard deviation
Calibration set	35	11.01%	16.87%	14.14%	1.18%
Prediction set	11	13.16%	16.45%	14.38%	0.91%

2.2 小麦籽粒粗蛋白平均模型建立与优化

确定合适的子区间个数和联合区间数是采用 SiPLS 法筛选特征变量的关键。对样本集的 46 条平均光谱进行多元散射校正预处理后，试验将全光谱等分成 6~25 个子区间，对于每个确定的子区间个数 n，分别建立和比较了联合区间数为 2、3、4 的最佳 PLS 模型，结果如表 2 所示。当划分区间数为 11，建模子区间组合为[4 8 11]时，所建模型指标最优，且采用的波长点数从 215 减少到 32 个，极大降低了模型复杂度。

表 2 SiPLS 的区间组合建模

Tab.2 Models based on SiPLS

Joint interval number	Divided interval number	Interval combination	Wavelength number	*nF*	*Rc*	RMSECV	*Rp*	RMSEP	RPD
2	6	1 2	71	8	0.92	0.45%	0.94	0.37%	2.49
3	11	4 8 11	32	10	0.94	0.40%	0.95	0.28%	3.30
4	16	1 2 4 16	56	8	0.94	0.40%	0.95	0.33%	2.78

2.3 单籽粒小麦粗蛋白含量预测

文中所述小麦籽粒粗蛋白平均模型是基于高光谱图像感兴趣区内的平均光谱建立的，而高光谱成像技术的最大优势在于它在感兴趣区内每个空间像素点处都包含了丰富的光谱信息，因此可以应用平均模型预测每个空间像素点的粗蛋白含量。

从样本集中挑选出蛋白含量差异较大的两个样本 A28 和 A47，样本平均粗蛋白含量分别为 12.42% 和 16.73%，各抽选 3 粒构成待测单籽粒样本集。获取待测单粒小麦种子的高光谱图像，提取图像中每个像素点的光谱信息带入 2.2 节中建立的最优 PLS 定量模型，获得单粒小麦种子每个空间像素点的蛋白预测值，取其平均值作为该粒麦种的粗蛋白含量。

图 3 给出了 6 粒待测单籽粒小麦种子的粗蛋白含量预测结果，图中第二列为每粒小麦经旋转放大后的 RGB 图像，第三列以伪彩色图的方式直观展示单粒小麦每个空间像素点的蛋白含量，并标出每粒麦种的蛋白含量平均值(%)。可以看出，同一样本的不同小麦籽粒的蛋白预测值存在差异性，但同时又围绕其所在样本平均粗蛋白含量值浮动，说明采用高光谱成像技术预测单籽粒小麦粗蛋白含量的方法是可行的，可为小麦育种的优选提供参考。

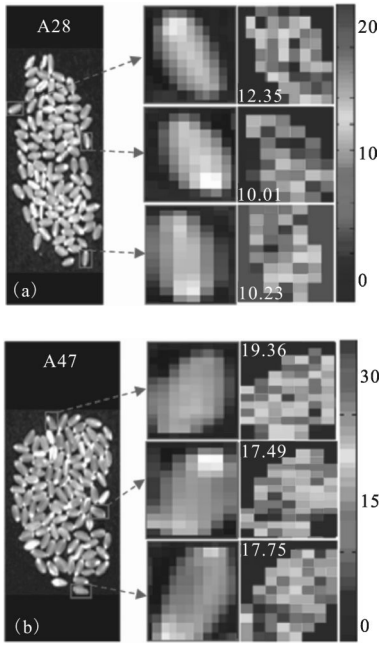

图 3 图(a)A28 与图(b)A47 单粒小麦种子粗蛋白含量预测

Fig.3 Protein content prediction results of sample (a) A28 and (b) A47

3 结束语

以 47 份不同品种的小麦样本为研究对象，每个样本取 100 粒，采集其高光谱图像并提取平均反射率光谱；利用 SiPLS 筛选特征波段，将波长变量数从 215 降低到 32 个；结合 PLS 方法建立小麦籽粒粗蛋白平均模型，模型的校正集及预测集相关系数分别为 0.94、0.95，预测均方根误差 RMSEP 为 0.28%，相对分析误差 RPD 为 3.30；最后将待测单籽粒小麦种子的高光谱图像简化为特征波长下的图像，带入上述模型，获得单籽粒每个空间像素点的蛋白预测值，取平均后作为该粒麦种的粗蛋白含量，并通过伪彩色图方式直观展示。结果表明，同一样本的不同小麦籽粒的蛋白预测值存在差异，但同时又围绕其所在样本平均粗蛋白含量值浮动，说明利用高光谱成像技术结合偏最小二乘法预测单籽粒小麦种子粗蛋白含量的方法是可行的，能够为挑选高蛋白籽粒母本实现小麦优质育种提供一种新思路。

参考文献：

[1] Qi Linjuan, Hu Xuexu, Zhou Guiying, et al. Analysis of wheat protein quality in the main province of China in 2004-2011 [J]. ***Scientia Agricultura Sinica***, 2012, 45(20): 4242-4251. (in Chinese)

[2] Li Shiping, Wang Suibao, Yang Yujing, et al. Research on wheat protein content genetic law and quality improvement [J]. ***Chinese Agricultural Science Bulletin***, 2005, 21(2): 126-128. (in Chinese)

[3] Feng Hui. Analysis of different wheat grain's protein and

starch content variance and the sowing time effect of the quality traits [D]. Zhengzhou: Henan Agricultural University, 2009. (in Chinese)

[4] Zhang Baohua, Li Jiangbo, Fan Shuxiang, et al. Principle and application of hyperspectral image technology in fruit and vegetable quality and safety nondestructive testing [J]. ***Spectroscopy and Spectral Analysis***, 2014, 34(10): 2743-2751. (in Chinese)

[5] Li Ziyang, Qian Yonggang, Shen Qingfeng, et al. Leaf area index retrieval from remotely sensed hyperspectral data [J]. ***Infrared and Laser Engineering***, 2014, 43(3): 944-949. (in Chinese)

[6] Sun Mei, Chen Xinghai, Zhang Heng, et al. Nondestructive inspect of apple quality with hyperspectral imaging [J]. ***Infrared and Laser Engineering***, 2014, 43(4): 1272-1277.(in Chinese)

[7] Li Dan, He Jianguo, Liu Guishan, et al. Non-destructive detection of moisture content in gherkin using hyperspectral imaging [J]. ***Infrared and Laser Engineering***, 2014, 43(7): 2393-2397. (in Chinese)

[8] GB/T 5511-2008. The kjeldahl method for determination of nitrogen content and crude protein content of grains and legumes[S]. ISO20483, IDT, 2006. (in Chinese)

[9] Rafael C G, Richard E W, Steven L E. Digital Image Processing Using Matlab [M]. Beijing: Publishing House of Electronics Industry, 2005.

[10] Chen Quansheng, Jiang Pei, Zhao Jiewen. Measurement of total flavone content in snow lotus (Saussurea involucrate) using near infrared spectroscopy combined with interval PLS and genetic algorithm [J]. ***Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy***, 2010, 76 (1): 50-55. (in Chinese)