中国农业科学 2006,39(12):2435-2440
Scientia Agricultura Sinica

# 水稻蛋白质含量 NIR 模型适配范围的研究

吴金红，张洪江，梅捍卫，李 荧，杨 华，王晓珊，林榕辉，罗利军

（上海市农业生物基因中心，上海 201106）

**摘要**:【目的】比较不同类型样品建立水稻蛋白质近红外模型的效果和适配范围。【方法】通过对 178 份来自“II-32B/岳早籼 6 号”的重组自交系和 496 份水稻品种的近红外反射光谱的比较分析，选择其中 59 个株系和 76 份品种作为建模样品，采用偏最小二乘法建立基于品种、重组自交系和混合样品的 3 个蛋白质含量回归模型。【结果】经模型内部交叉验证和对模型外部重组自交系和品种样品的验证结果的比较分析，发现基于分离群体的模型因蛋白质含量范围较窄，样品来源较单一，适应范围仅局限于本群体内样品蛋白质含量预测，而品种和混合模型对群体和品种样品都表现出良好的适应能力，交叉验证决定系数大于 0.90，外部验证决定系数大于 0.89，本试验可为近红外建模的样本集选择提供良好的指导意义。【结论】不同类型样品对建模效果有显著影响，品种模型和混合模型的适配范围显著大于群体模型，研究结果不能支持用背景变异较小的样品建立较高精度回归模型的设想。

**关键词**: 水稻；蛋白质含量；近红外反射光谱分析（NIRS）；群体；品种

# Study on Adaptability of NIR Models of Protein Content in Rice

WU Jin-hong, ZHANG Hong-jiang, MEI Han-wei, LI Ying, YANG Hua,
WANG Xiao-shan, LIN Rong-hui, LUO Li-jun

(*Shanghai Agrobiological Gene Center, Shanghai 201106*)

Abstract: 【Objective】This research was done in order to compare the effect and adaptability of NIR models in different calibration samples.【Method】With NIR spectra of 178 RILs from cross of “II-32B/Yue-Zao-Xian No.6” and 496 rice varieties, 59 RILs and 76 cultivars were selected to develop three PLS regression models. These wre based on RILs, cultivars and mixture samples, respectively. 【Result】 By comparing cross validation results with calibration sets and predictions for both RILs and varieties, we confirmed limitations of RILs based model due to narrow ranges of protein content and lack of diversity. A more reliable prediction method could be established only when the model was being applied to samples from RILs. However, both models that were based on representative varieties and mixture samples exhibited much better adaptability than samples from RILs or varieties with higher determination coefficients in cross validation ($r^2$ >0.90) and testing set validation ($r^2$ >0.89), respectively. This gave suggestions on calibration sample selection. 【Conclusion】Modeling results varied greatly in different types of calibration samples. The adaptability for calibration models in varieties and their mixtures were much broader than that of RILs. Using samples of smaller genetic variance, we were unable to create a regression model with more accuracy.

Key words: Rice; Protein content; Near infrared spectroscopy (NIRS); RILs; Varieties

## 0 引言

【本研究的重要意义】水稻是中国最重要的禾谷类粮食作物。和其它谷物相比，稻米蛋白质的氨基酸平衡和赖氨酸含量是决定水稻营养品质的重要因素之一[1]，提高蛋白质及赖氨酸的水平也是水稻品质育种的重要目标。常规的蛋白质含量分析采用凯氏法（Kjeldahl）（国标 GB2905-82），其操作步骤多，分

**收稿日期**：2005-07-03；**接受日期**：2006-06-02
**基金项目**：上海市农业科学院青年基金项目、农业部“948”项目和国家“973”计划项目（2004CB117204）
**作者简介**：吴金红（1975-），女，浙江平湖人，助研，硕士，研究方向为生物技术。Tel：021-52230526，Fax：210-62204010，E-mail: wjh@sagc.org.cn。张洪江（1979-），男，江苏宿迁人，助研，硕士，研究方向为仪器分析。Tel：021-52230526，Fax：210-62204010，E-mail: zhj@sagc.org.cn。通讯作者罗利军（1961-），男，湖北咸宁人，研究员，博士，研究方向为植物遗传育种，Tel: 021-62200490，Fax：210-62204010，E-mail: lijun@sagc.org.cn。 第一、二作者同等贡献

析效率低，更为不利的是样品被破坏。在水稻种质资源研究及品质育种中需要对大量材料及时进行分析、鉴定和筛选，同时希望具有优良品质的米粒经分析后能保持完好，以便进一步繁殖。因此，在进行蛋白质含量分析时就迫切需要一种快速、简便、非破坏性的方法。【前人研究进展】近红外光谱分析技术（near infrared spectroscopy，NIRS）是 20 世纪 80 年代后期迅速发展起来的一项物理测试技术，作为一种快速无损分析方法[2,3]，已成为谷物品质分析的重要手段[4~12]。【本研究的切入点】但在对建模样品集对模型影响的评估，以及对不同类型的建模样品集所建立模型的适用范围方面的研究还很少。【拟解决的关键问题】本研究旨在运用傅立叶变换近红外光谱仪的漫反射光谱分析技术，采集水稻糙米粒样品的近红外光谱数据，比较分析不同类型建模样品集所建立的蛋白质含量模型的差异，为近红外建模的样品集选择提供良好指导，并选取合适的模型用于水稻的种质资源研究及品质育种中蛋白质含量的筛选分析。

# 1 材料与方法

## 1.1 仪器

德国 Bruker FT-NIR 近红外光谱仪（Vector 22/N-I）；德国 Bran+Luebbe AA3 连续流动化学分析仪；瑞典 FOSS 消化炉（2012 Digestor）。

## 1.2 试验材料

试验材料共 674 份，包括来自亲本 II-32B/岳早籼 6 号的含有 178 个株系的重组自交系群体和 496 份来自全国各地以及国外引进的代表品种。所有试验材料均在上海市农业科学院良种繁育基地（上海，青浦）繁种获取足量种子。

采集全部材料的近红外光谱，利用 Bruker 公司提供的近红外光谱挑选软件，从中选出光谱特征差异显著的 59 个株系和 79 份代表品种用于蛋白质含量的模型构建。

## 1.3 测定方法

**1.3.1 近红外光谱采集** 以镀金漫反射体为参比，波数范围 4 000 $cm^{-1}$～10 000 $cm^{-1}$，分辨率 8 $cm^{-1}$，扫描次数 64 次，采集糙米粒的近红外光谱。为减少样品粒度差异所引起的光谱不均一性，每个样品重复装样扫描 3 次，扫描时保持样品杯低速旋转，取平均光谱作为该样品的近红外光谱[13,14]。

**1.3.2 蛋白质含量化学值测定** 将已采集了近红外光谱的糙米粒用旋风磨粉碎，过 60 目筛，糙米粉经浓硫酸完全消煮后，使用 AA3 连续流动化学分析仪测定蛋白质含量[15]，并以此化学值代表蛋白质含量的真值。

**1.3.3 NIRS 模型构建** 利用 Bruker 公司的 OPUS（Version 4.0，QUANT Package）软件的 Quant 2 方法，采用偏最小二乘法（PLS）建立了糙米粒近红外光谱与其蛋白质含量之间的定标模型，使用内部交叉验证中各模型的决定系数，均方根等参数评价模型。

# 2 结果与分析

## 2.1 不同建模样本蛋白质含量分布

II-32B 是中国大面积应用的三系不育保持系，岳早籼 6 号的蛋白质含量比较高，经实际测定两者蛋白质含量有明显差异，分别为 11.4%和 14.26%。为了对蛋白质含量进行 QTL 定位分析，笔者利用这两个亲本建立了重组自交系群体。在用于本试验建模的 59 个株系中，糙米粉粗蛋白含量平均值为 13.64%，变幅为 9.97%～16.19%；用于建模的 76 份代表性品种的蛋白质含量平均值为 10.91%，变幅为 7.95%～16.02%；其中蛋白质含量较低一侧比群体样品有所扩展，较高一侧互相比较接近；由重组自交系和代表品种共同构成的 139 份混合样品集的蛋白质平均值为 12.13%，分布范围为前两种不同样品集的总和（表 1）。

由 II-32B/岳早籼 6 号所构建的重组自交群体，样品蛋白质含量分布比较集中，化学测定值主要集中在

**表 1 建模样品蛋白质含量统计参数**

Table 1 Statistical parameters of calibration samples for protein content

| 模型 Model | 样品数 No. of samples | 变幅(%) Range | 平均值(%) Mean | 标准差 Standard deviation |
|---|---|---|---|---|
| 群体 RILs | 59 | 9.97～16.19 | 13.64 | 1.00 |
| 品种 Varieties | 76 | 7.95～16.02 | 10.91 | 1.66 |
| 混合 Mixture | 139 | 7.95～16.19 | 12.13 | 1.93 |

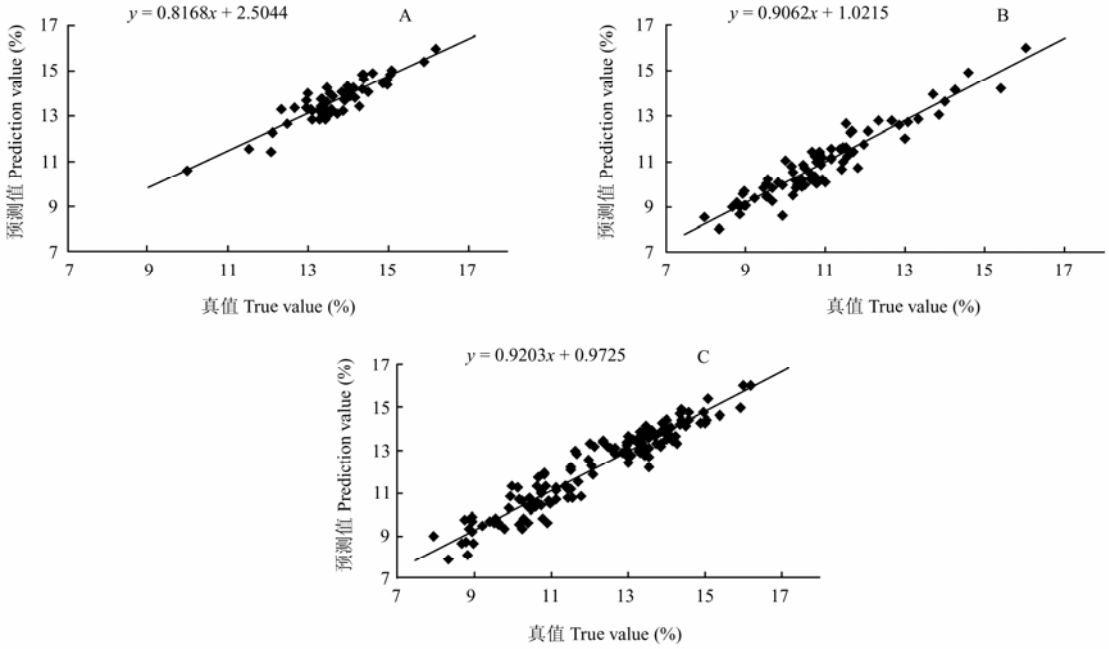
13%～15%这一区间；品种样品分布的均匀度略好于重组自交系，但仍有较多样品的蛋白质含量集中在9%～11.5%范围内；而混合样品的蛋白质含量因两类样品互相弥补而呈现比较均匀的分布（图）。

### 2.2　校正模型的建立及内部交叉验证

经 Quant 2 程序优化，最后分别使用常数偏移去除、最小最大归一化、二阶导数预处理方法，选取 6 102 $cm^{-1}$～4 242.8 $cm^{-1}$、6 102 $cm^{-1}$～4 597.7 $cm^{-1}$、6 102 $cm^{-1}$～4 597.7 $cm^{-1}$ 谱区范围（波数间隔 1），建立了群体、品种和群体品种混合的蛋白质模型（图）。各校正模型的技术参数见表 2。群体模型的决定系数较低，品种和混合模型的决定系数相近，并且达到 0.90 以上；3 种模型的交叉验证均方根分别为 0.436、0.520、0.579，均在较低范围之内。

### 2.3　各模型的外部验证

为检验各预测模型的准确性，笔者各选 9 份重组自交系群体内株系样品和 9 份不同品种样品，对已建立的各模型进行外部验证。外部验证样品的真实蛋白质含量由连续流动化学分析仪测出，其平均值等参数于表 3 中列出。这些外部验证样品的蛋白质含量分布比较宽，极端最小、最大值接近于建模样品的极端值。3 种模型对两类外部样品预测值与化学测定值之间的




A. 重组自交系；B. 代表品种样品；C. 混合样品
A. Recombinant inbred lines (RILs); B. Representative varieties; C. Mixture samples

图　不同校正模型的蛋白质含量真值与预测值的交叉验证结果

Fig.　Cross validation between the true values and the prediction values for protein content using different models

表 2　3 种模型的交叉验证结果

Table 2　Results of the cross validation of three models

| 模型 Model | 谱区 ($cm^{-1}$) Frequency region | 预处理 Preprocessing | 维数 Ranks | 决定系数 $r^2$ | 交叉验证均方根 RMSECV |
|---|---|---|---|---|---|
| 群体 RILs | 6102～4242.8 | COE | 7 | 0.8050 | 0.436 |
| 品种 Varieties | 6102～4597.7 | MMN | 7 | 0.9001 | 0.520 |
| 混合 Mixture | 6102～4597.7 | SD | 8 | 0.9090 | 0.579 |

$r^2$=coefficient of determination, RMSECV=root mean square error of cross validation; COE=constant offset elimination 常数偏移去除; MMN=Min-Max normalization 最小最大归一化; SD=second derivative 二阶导数

**表 3 外部样品蛋白质含量统计参数**

Table 3 Statistical parameters of testing samples for protein content

| 外部样品 Samples | 样品数 No. of samples | 变幅(%) Range | 均值(%) Mean | 标准差 S.D. |
|---|---|---|---|---|
| 群体 RILs | 9 | 11.40～15.74 | 13.31 | 1.43 |
| 品种 Varieties | 9 | 8.76～13.37 | 11.11 | 1.63 |

平均偏差除群体模型对品种样品预测时稍大外，其余都为±0.1 左右，说明模型的预测结果与常规化学分析方法接近；*t* 检验值在 0.12～0.42 之间，都远小于 $P \leqslant 0.05$ 的临界值（$t_{0.05}$=2.31），进一步说明近红外预测与常规化学分析方法无明显差异。

在对群体样品的外部验证中，3 个模型都有较高的决定系数，特别是品种模型有最高的决定系数，预测均方根低于或接近于 0.5（表 4）。

在对品种样品的外部验证中，品种和混合校正模型得到满意结果，而群体模型的决定系数偏小，对比图 A 可以发现，群体建模样品中蛋白质含量在 10%～12.5%区间内样品数很少，而误差偏大的品种样品中 5 个样品的蛋白质含量真值都落在这一区间里（数据未列出），且真值最小的外部样品已经超出了群体的建模样品集的变幅。因为群体建模样品集里缺乏低蛋白的参试样品，模型对此类样品的认知能力差，所以很难获得准确的预测结果；混合模型对外部品种样品的预测结果又明显好于品种模型，可见在建模样品中加入重组自交系材料后，由于弥补了样品分布的不均一，进一步提高了校正模型的适应性。

综上所述，以变幅大，分布均一的建模样品集得到的校正模型有良好的适应能力，对外部群体和品种样品的预测比较稳定、准确。一旦建模样品集变幅小、分布不够均一，则模型只能识别同类型的未知样品，不能适应变异广泛的其它外部样品。

**表 4 外部样品的验证结果参数统计**

Table 4 Statistical parameters of testing set for protein content

| 模型 Model | 群体样品 Samples from RI population | | | | 品种样品 Samples from varieties | | | |
|---|---|---|---|---|---|---|---|---|
| | 决定系数 $r^2$ | 预测均方根 RMSEP | 平均偏差 Bias | *t* 测验 *t* test | 决定系数 $r^2$ | 预测均方根 RMSEP | 平均偏差 Bias | *t* 测验 *t* test |
| 群体 RILs | 0.8740 | 0.508 | -0.11 | 0.29 | 0.7861 | 0.777 | 0.31 | 0.12 |
| 品种 Varieties | 0.9104 | 0.432 | 0.08 | 0.29 | 0.8980 | 0.535 | -0.04 | 0.42 |
| 混合 Mixture | 0.8946 | 0.495 | -0.12 | 0.25 | 0.9315 | 0.448 | -0.14 | 0.19 |

RMSEP=root mean square error of prediction

## 3 讨论

谷物籽粒样品的近红外光谱成功应用于各类成份含量的快速预测，其精确度很大程度上取决于仪器定标模型的质量。红外光谱特征与水分、蛋白、脂肪、直链淀粉等含量的相关也受到样品基本状态的影响，如品种类型、籽粒成熟度、产地（成熟时气候环境）、收获后干燥储存条件等[16~20]。因此，一种比较可靠的建模策略是尽量扩大建模样品的变异范围，例如要求待测指标的变幅很宽、尽可能包括各种类型的品种、从不同产地和种植季节收集样品等等。把这些背景因素对近红外光谱的干扰纳入到模型优化程序之中，那么所得到的模型就可以得到很好的校正，从而使得模型具有广泛的适应范围，可应用于多种多样未知样品的预测。目前报道的谷物品质近红外定标工作大多采用上述建模策略，其模型具有比较好的适应能力[18~20]。

与此相反，如果采用遗传背景和产地环境比较一致的样品进行近红外建模，那么可以预期所建模型将具有以下特征：（1）该模型的适应范围狭窄，仅能用于相类似的未知样品的预测；（2）模型可能有比较好的技术指标，交叉验证的误差较小；（3）对于同类型未知样品的预测准确度应有所提高。但这些推断有待实验验证，至今未见上述两种策略建模效果对比试验的研究报道。

本试验以水稻糙米粗蛋白含量为例，利用一套重组自交系群体代表变异较少的样品集；同时以类型、来源广泛的水稻品种构成变异较大的样品集，以两者共同组成的混合样品为第三种样品集，分别建立近红

外回归模型，通过内部交叉验证和外部未知样品的预测准确度，对比分析了 3 种样品模型的适应范围。

试验结果表明，不同类型的建模样品集对模型质量的影响极为显著。首先，建模样品蛋白质含量的变幅，决定了校正模型的适宜分析区间，超出这个区间后模型不能够准确预测；其次，样品分布不能过度集中，以保持不同类型样品权重的一致。例如，对于群体模型而言，蛋白质含量即使落在校正样品集变幅内，但由于在整个建模样品中该区间的样品数量偏少，也使得预测结果明显偏离真实值；第三，品种和混合样品集保持了较大变幅和较均一分布，所建模型比群体模型有明显的改善；第四，由于重组自交系群体样品蛋白质含量分布的局限性，群体模型的基本参数不涉及品种和混合模型，对于未知品种样品的预测效果差，即使针对群体内的其它株系，群体模型的预测效果相对于品种模型而言，也没有任何精确度上的优势。

由于样品近红外光谱中包含来自样品各种理化特征的复杂信息，因此一个可靠的回归预测模型需要用大量代表性样品来建立和校正。本试验采用的建模样品数，尤其是外部验证样品偏少，虽然可以比较清晰地反映出不同类型样品对于建成的近红外预测模型的影响，但要把它作为实际应用的检测手段，还需进一步扩大样品集并在分析测试中不断优化模型，进一步提高整个模型的可靠性和精确度。

## 4　结论

本试验采用一个水稻重组自交系群体（RILs）和一些水稻品种的糙米样为材料，建立群体、品种和两者混合等三种蛋白质的近红外校正模型。比较分析发现不同类型样品对建模效果有显著影响，品种模型和混合模型能适应不同类型未知样品的预测，效果显著高于群体模型；群体模型因化学值分布较窄而缺乏对品种样品的适应性；研究结果不能支持用背景变异较小的样品建立较高精度回归模型的设想。

**致谢**：感谢 Bruker 公司（北京）技术人员在光谱特征分析和样品筛选工作中给予的帮助。

## References

[1] 罗利军, 应存山, 汤圣祥. 稻种资源学. 湖北科学技术出版社, 2002: 408-414.
Luo L J, Ying C S, Tang S X. *Rice Germplasm Resources*. Hubei Science and Technology Press, 2002: 408-414. (in Chinese)

[2] Blanco M, Villarroya I. NIR spectroscopy: a rapid-response analytical tool. *Trends in Analytical Chemistry*, 2002, 21 (4): 240-250.

[3] Geladia P, Sethson B, Nyström J, Lillhonga T, Lestander T, Burger J. Chemometrics in spectroscopy Part 2. Examples. *Spectrochimica Acta Part B*, 2004, 59: 1347-1357.

[4] 唐绍清, 石春海, 焦桂爱, 胡培松, 王海莲, 万建民. 利用近红外反射光谱技术测定稻米中脂肪含量的研究初报. 中国水稻科学, 2004, 18: 563-566.
Tang S Q, Shi C H, Jiao G A, Hu P S, Wang H L, Wan J M. Analysis of fat content in milled rice by near infrared reflectance spectroscopy. *Chinese Journal of Rice Science*, 2004, 18: 563-566. (in Chinese)

[5] 王秀荣, 廖　红, 严小龙. 应用近红外光谱分析法测定大豆种子蛋白质和脂肪含量的研究. 大豆科学, 2005, 24: 199-201.
Wang X R, Liao H, Yan X L. Study on analyzing soybean protein and oil contents by near-infrared spectroscopy. *Soybean Science*, 2005, 24: 199-201. (in Chinese)

[6] Delwiche S R. Protein content of single kernels of wheat by near-infrared reflectance spectroscopy. *Journal of Cereal Science*, 1998, 27: 241-254.

[7] Pettersson H, Aberg L. Near infrared spectroscopy for determination of mycotoxins in cereals. *Food Control*, 2003, 14: 229-232.

[8] Wu J G, Shi C H, Zhang X M. Estimating the amino acid composition in milled rice by near infrared reflectance spectroscopy. *Field Crops Research*, 2002, 75: 1-7.

[9] Wu J G, Shi C H. Prediction of grain weight, brown rice weight and amylose content in single rice grains using near-infrared reflectance spectroscopy. *Field Crops Research*, 2004: 87: 13-21.

[10] Blakeney A B, Flinn P C. Determination of non-starch polysaccharides in cereal grains with near-infrared reflectance spectroscopy. *Molecular Nutrition and Food Research*, 2005, 49: 546-550.

[11] Baye T M, Pearson T C, Settles A M. Development of a calibration to predict maize seed composition using single kernel near infrared spectroscopy. *Journal of Cereal Science*, 2006, 43: 236-243.

[12] Delwiche S R, Graybosch R A. Identification of waxy wheat by near-infrared reflectance spectroscopy. *Journal of Cereal Science*, 2002, 35: 29-38.

[13] 赵丽丽, 赵龙莲, 李军会, 张录达, 严衍禄. 傅立叶变换近红外光谱仪扫描条件对数学模型预测精度的影响. 光谱学与光谱分析, 2004, 24: 41-44.
Zhao L L , Zhao L L, Li J H, Zhang L D, Yan Y L. Influence of FT-NIR spectrometer requirements on the math model's precision. *Spectroscopy and Spectral Analysis*, 2004, 24: 41-44. (in Chinese)

[14] 李 宁, 闵顺耕, 覃方丽, 于飞健, 叶升锋. 近红外光谱法非破坏性测定黄豆籽粒中蛋白质、脂肪的含量. 光谱学与光谱分析, 2004, 24: 45-49.
Li N, Min S G, Qin F L, Yu F J, Ye S F. Nondestructive analysis of protein and fat in whole kernel soybean by NIR. *Spectroscopy and Spectral Analysis*, 2004, 24: 45-49. (in Chinese)

[15] Total Kjeldahl nitrogen in acid digests. Bran Luebbe, Inc. AutoAnalyzer Applications: Method No. G-188-97 Rev. 3, 2002.

[16] Shuso K, Motoyasu N, Kazuhiro T, Kazuhiko I. Development of an automatic rice-quality inspection system. *Computers and Electronics in Agriculture*, 2003, 40: 115-126.

[17] 魏良明, 严衍禄, 戴景瑞. 近红外反射光谱测定玉米完整籽粒蛋白质和淀粉含量的研究. 中国农业科学, 2004, 37: 630-633.
Wei L M, Yan Y L, Dai J R. Determining protein and starch content of whole maize kernel by NIRS. *Scientia Agricultura Sinica*, 2004, 37: 630-633. (in Chinese)

[18] 李君霞, 张洪亮, 严衍禄, 闵顺耕, 李自超. 水稻蛋白质近红外定量模型的创建及在育种中的应用. 中国农业科学, 2006, 39: 836-841.
Li J X, Zhang H L, Yan Y L, Min S G, Li Z C. Establishment of math models of NIRS analysis for protein contents in seed and it′s application in rice breeding. *Scientia Agricultura Sinica*, 2006, 39: 836-841. (in Chinese)

[19] 陈 斌, 赵龙莲, 李军会, 严衍禄. 近红外光谱法快速分析葛根中的有效成分. 光谱学与光谱分析, 2002, 22: 976-979.
Chen B, Zhao L L, Li J H, Yan Y L. The rapid analysis of functional components of P. lobata by near infrared spectrum. *Spectroscopy and Spectral Analysis*, 2002, 22: 976-979. (in Chinese)

[20] 甘 莉, 孙秀丽, 金 良, 王高全, 徐久伟, 魏泽兰, 傅廷栋. NIRS 定量分析油菜种子含油量、 蛋白质含量数学模型的创建. 中国农业科学, 2003, 36: 1609-1613.
Gan L, Sun X L, Jin L, Wang G Q, Xu J W, Wei Z L, Fu T D. Establish of math models of NIRS analysis for oil and protein contents in seed of Brassica napus. *Scientia Agricultura Sinica*, 2003, 36: 1609-1613. (in Chinese)

（责任编辑 于 竞）