

DOI: 10.3724/SP.J.1006.2026.51073

裸大麦籽粒 β -葡聚糖含量近红外预测模型的建立

李 赢 石晓旭 刘海翠 石 吕 薛亚光 魏亚凤*

江苏沿江地区农业科学研究所 / 南通市循环农业重点实验室, 江苏南通 226001

摘 要: β -葡聚糖具有调节血糖、降血脂、抗肿瘤、抗氧化等功效, 是裸大麦的重要品质指标之一。建立裸大麦籽粒 β -葡聚糖含量近红外预测模型, 无损、快速地对低世代裸大麦材料籽粒 β -葡聚糖含量进行测定, 可显著提高裸大麦育种效率。本研究采集了 215 份裸大麦籽粒的近红外光谱, 使用酶法测定了 β -葡聚糖含量, 根据光谱扫描结果和蒙特卡洛交叉验证法剔除异常样本 10 个, 利用 SPXY 法将剩余样本划分为校正集和验证集。使用归一化、一阶导数、二阶导数等 14 种方法对光谱进行预处理, 使用偏最小二乘回归(PLSR)、支持向量机(SVM)和主成分回归(PCR)算法构建裸大麦籽粒 β -葡聚糖含量的近红外预测模型, 其中二阶导数处理的 SVM 模型预测效果最好。结合连续投影算法(SPA)、无信息变量剔除法(UVE)、竞争性自适应重加权采样法(CARS)、变量迭代空间收缩法(VISSA)、自举软收缩法(BOSS)和迭代保留信息变量法(IRIV) 6 种特征变量筛选方法对二阶导数处理的 SVM 模型进行优化。其中 CARS-SD-SVM 模型的校正决定系数(R^2_c)、校正标准偏差(RMSEC)、外部验证决定系数(R^2_p)、验证标准偏差(RMSEP)和相对分析误差(RPD)分别是 0.859、0.272、0.877、0.237 和 2.790, 模型预测效果最佳, 且可靠性较高, 适用于裸大麦籽粒 β -葡聚糖含量的分析, 在裸大麦优质育种和品质分析中具有广泛的应用价值。

关键词: 裸大麦; β -葡聚糖含量; 近红外光谱; 预测模型; 特征变量筛选

Establishment of a near-infrared reflectance spectroscopy model for predicting β -glucan content in naked barley grain

Li Ying, Shi Xiao-Xu, Liu Hai-Cui, Shi Lyu, Xue Ya-Guang, and Wei Ya-Feng*

Jiangsu Yanjiang Area Institute of Agricultural Sciences / Key Laboratory of Recycling Agriculture of Nantong City, Nantong 226001, Jiangsu, China

Abstract: β -glucan, known for its roles in regulating blood sugar, lowering blood lipids, and exhibiting anti-tumor and anti-oxidant properties, is one of the most important quality indicators of naked barley. Establishing a near-infrared (NIR) prediction model for β -glucan content enables rapid and non-destructive evaluation of this trait in early-generation naked barley materials, which can significantly enhance breeding efficiency. In this study, NIR spectra were collected from 215 naked barley grain samples, and β -glucan content was determined using enzymatic methods. Based on spectral scanning and Monte Carlo cross-validation, 10 outlier samples were excluded. The remaining samples were divided into calibration and validation sets using the SPXY method. Fourteen preprocessing techniques, including normalization, first derivative, and second derivative, were applied to the spectral data. Partial least squares regression (PLSR), support vector machine (SVM), and principal component regression (PCR) were employed to construct prediction models for β -glucan content. Among these, the SVM model combined with the second derivative preprocessing showed the best predictive performance. To further optimize this model, six feature variable selection methods—successive projections algorithm (SPA), uninformative variable elimination (UVE), competitive adaptive reweighted sampling (CARS), variable iterative space shrinkage algorithm (VISSA), bootstrapping soft shrinkage (BOSS), and iteratively retained informative variables (IRIV) were evaluated. The CARS-SD-SVM model achieved the highest accuracy, with R^2_c , RMSEC, R^2_p , RMSEP, and RPD values of 0.859, 0.272, 0.877, 0.237, and 2.790, respectively. This model not only demonstrated excellent predictive performance but also high reliability, making it suitable for the rapid analysis and prediction of β -glucan content in naked barley grains. It holds great potential for

本研究由江苏省农业科技自主创新资金项目(CX (24) 3089)资助。

This study was supported by the Jiangsu Agricultural Science and Technology Innovation Fund Project (CX (24) 3089).

* 通信作者(Corresponding author): 魏亚凤, E-mail: w-yafeng@163.com

第一作者联系方式: E-mail: 1185077523@qq.com

Received (收稿日期): 2025-08-06; Accepted (接受日期): 2025-11-18; Published online (网络出版日期): 2025-11-20.

URL: <https://link.cnki.net/urlid/11.1809.S.20251120.1110.002>

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

analysis and prediction of β -glucan content in naked barley grains. It holds great potential for application in naked barley breeding and quality assessment.

Keywords: naked barley; β -glucan content; near infrared spectroscopy (NIRS); prediction model; feature variable selection

裸大麦(*Hordeum vulgare* var. *nudum* Hooker f.) 属于禾本科大麦属, 又称青稞、元麦和米麦^[1], 成熟后稃壳容易脱落。相较其他谷物, 裸大麦 β -葡聚糖含量水平较高, 一般为 3.66%~8.62%^[2]。 β -葡聚糖是裸大麦籽粒细胞壁的重要组分, 其分子量高, 结构高度不对称, 导致黏性较高, 影响动物的营养吸收效率及以裸大麦为原料的啤酒的加工性^[3]。然而, β -葡聚糖具有调节血糖、降血脂、抗炎、免疫、抗肿瘤、抗氧化等功效, 经常食用有助于缓解阿尔茨海默病、降低胆固醇和患结肠癌的风险, 因此裸大麦作为 β -葡聚糖的重要来源受到人们的广泛关注, 市场潜力巨大^[4-6]。为满足啤酒加工和保健品加工对不同 β -葡聚糖含量裸大麦的需求, 开展专用型裸大麦育种至关重要。

了解不同裸大麦资源和育种中间材料的 β -葡聚糖含量是进行专用型育种的关键。酶法是测定 β -葡聚糖含量的常用方法, 然而酶法会破坏样本籽粒, 且存在成本较高、操作繁琐、费时费力、安全隐患等弊端, 不适用于育种中间材料的高通量检测^[7]。建立一套无损、高效、低成本、高通量的 β -葡聚糖含量测定方法, 对育种家快速了解不同裸大麦材料的 β -葡聚糖含量水平和开展专用型裸大麦育种至关重要。近红外光谱是一种新兴的无损检测技术, 操作简单、成本较低, 能无损、快速地获得谷物的多个营养成分的含量, 成功应用于花生、谷子、水稻等作物营养成分的快速检测^[8-10]。前人建立了裸大麦 β -葡聚糖含量的近红外预测模型, 如 Schmidt 等^[11]使用 4 种不同的近红外光谱仪扫描 106 份裸大麦和 1 份皮大麦, 分别建立籽粒和谷粉的 β -葡聚糖含量预测模型, 其中籽粒最优模型的决定系数为 0.848, 相对分析误差为 1.82, 不可直接用于育种中间材料 β -葡聚糖含量的定量分析。杨焰婷等^[12]利用近红外光谱仪建立了 76 份裸大麦粉样的特征营养成分含量快速检测模型, 其中 β -葡聚糖含量最优模型的预测决定系数和预测均方根误差分别为 0.9808 和 0.1846, 为加工企业提高合格原料的筛选效率提供了技术指导。Li 等^[13]以 120 份来自西藏、青海和甘肃的裸大麦为材料, 建立了籽粒 β -葡聚糖含量的近红外预测模型, 最优模型的外部验证决定系数为 0.763, 验证

标准偏差为 0.328, 模型预测效果较差, 仍需进一步优化。综上, 前人有关裸大麦籽粒 β -葡聚糖含量的近红外预测模型的研究仍相对较少, 加上模型构建存在使用的样本数量较少、预处理和建模方法较单一等问题, 现有的籽粒预测模型仅可用于粗略估计裸大麦的 β -葡聚糖含量, 难以直接应用于育种。此外, 基于粉样的近红外预测模型的预测效果优于籽粒样品, 然而该方法会破坏籽粒样品, 影响材料的后续种植, 在育种早期世代难以应用于基于粉样的预测模型对单株材料进行的高通量筛选。因此, 需构建预测效果更好的裸大麦籽粒 β -葡聚糖含量的近红外预测模型, 对育种中间材料的 β -葡聚糖含量进行无损快速、低成本测定, 进而为专用型裸大麦育种奠定基础。

本研究选取 215 份国内外裸大麦种质资源, 测定不同资源籽粒的 β -葡聚糖含量, 对其进行近红外光谱扫描, 结合光谱数据和化学测定值构建 β -葡聚糖含量的近红外预测模型并进行优化, 旨在实现对现有裸大麦种质资源和育种中间材料 β -葡聚糖含量的高通量、无损、快速、低成本检测, 提高优异资源和育种中间材料的鉴定挖掘效率, 为裸大麦品质育种、 β -葡聚糖含量相关基因挖掘以及大健康产业发展提供技术支撑。

1 材料与方法

1.1 试验材料

选取 215 份裸大麦资源(附表 1), 国内资源 167 份, 其中江苏 62 份, 西藏 51 份, 四川 21 份, 青海 12 份, 云南 6 份, 其他地区 15 份; 国外资源 48 份, 其中亚洲(除中国) 19 份, 欧洲 10 份, 北美洲 17 份, 国外不知来源 2 份。参试资源籽粒黄色 135 份, 蓝色 35 份, 紫色 32 份, 褐色 8 份, 黑色 5 份。2023 年 11 月播种于江苏省南通市如皋市薛窑基地, 2024 年 5 月收获, 脱粒晒干后, 常温保存备用。

1.2 近红外光谱采集

使用瑞典波通 DA7250 近红外分析仪(DA7250 Diode Array Analyzer)的 Simplicity 软件采集 215 份裸大麦籽粒的光谱数据(图 1)^[14]。分析仪扫描光谱范围为 950~1650 nm, 分辨率设置为 2 nm, 每次扫描获得 351 个波长的吸收强度数据。扫描前, 参试

样本籽粒于恒温箱中放置 48 h, 温度设置为 25℃, 光谱采集时, 室内温度控制在 25℃左右。仪器开机预热 30 min 后, 将裸大麦籽粒均匀放于样品杯中, 使用尺子平整籽粒表面。每份裸大麦籽粒样品装样 3 次, 每次装样扫描 3 次, 取平均光谱值用于近红外预测模型的建立。



图1 本试验中使用的近红外检测仪 DA7250 和小样品盘
Fig. 1 Near-infrared reflectance spectrometer DA7250 and small sample tray used in this experiment

1.3 β -葡聚糖含量测定

利用 Megazyme 公司生产的 β -葡聚糖试剂盒 (K-BGLU 02/17)测定 215 份裸大麦样本的 β -葡聚糖含量。由同一个操作员完成 215 份裸大麦籽粒 β -葡聚糖含量的测定, 每份材料测定 3 个重复, 取平均值记作该样品的化学测定值。

1.4 近红外预测模型的建立与优化

1.4.1 异常样本剔除 由于异常样本数据存在较大误差, 严重影响模型的预测效果。此外, 光谱预处理难以消除异常样本数据中的误差, 因此在模型建立前应先剔除异常样本。使用蒙特卡洛交叉验证法 (monte carlo cross validation, MCCV)同时检测光谱数据和化学测定值的异常值, 计算样本预测误差均值和标准差, 筛选出均值和标准差较高的样本作为异常样本^[15]。

1.4.2 样本集划分 采用基于联合 X-Y 距离的样本集划分法 (samples set partitioning based on joint X-Y distances, SPXY)将剔除异常样本后的样本集按 4 : 1 比例划分为校正集和验证集^[16]。通过计算校正集和验证集样本的平均值、最大值、最小值、极差、标准偏差和变异系数来评估样本划分是否合理。

1.4.3 光谱数据预处理 光谱数据采集过程易受环境、仪器、样本等影响, 出现基线漂移、随机噪声、光谱分散等问题^[17]。因此使用归一化处理、一

阶导数、二阶导数、多元散射校正 (multiplicative scatter correction, MSC)、标准正态校正 (standard normal variate, SNV)、Savitzky-Golay 卷积平滑 (SG)、SG 一阶求导、SG 二阶求导、去散射、一阶导数+MSC、一阶导数+SNV、SG+MSC、SG+SNV、SNV+MSC 共 14 种方法对光谱数据进行预处理^[15,18-19]。

1.4.4 特征变量筛选 每份裸大麦样本包含 351 个近红外光谱变量, 但其中大部分变量可能无法有效反映裸大麦 β -葡聚糖含量, 存在信息冗余的风险, 可能导致模型过拟合。因此, 必须从 351 个波长变量中筛选出能准确表征裸大麦 β -葡聚糖含量的特征变量。采用连续投影算法 (successive projections algorithm, SPA)、无信息变量剔除法 (uninformative variables elimination, UVE)、竞争性自适应重加权采样法 (competitive adaptive reweighted sampling, CARS)、变量迭代空间收缩法 (variable iterative space shrinking algorithm, VISSA)、自举软收缩法 (bootstrapping soft shrinkage, BOSS)、迭代保留信息变量法 (iteratively retained informative variables, IRIV)提取对应 β -葡聚糖含量的特征变量^[20-22]。利用特征变量建立近红外预测模型, 通过对比模型的预测参数, 最终确定最优特征变量选择算法。

1.4.5 近红外预测模型的建立 利用偏最小二乘回归 (partial least squares regression, PLSR)、支持向量机 (support vector machine, SVM) 和主成分回归 (principal component regression, PCR)算法对预处理后的光谱数据进行回归分析, 构建裸大麦籽粒 β -葡聚糖含量的近红外预测模型。以校正决定系数 (R^2_c)、校正标准偏差 (root mean square error of calibration, RMSEC)、外部验证决定系数 (R^2_p)、验证标准偏差 (root mean square error of prediction, RMSEP)和相对分析误差 (ratio of performance to deviation, RPD)作为衡量模型预测效果的主要参数。其中, R^2_c 和 R^2_p 取值范围为 0~1, 越接近于 1, 表明模型准确度越高; RMSEC 和 RMSEP 越低, RPD 越高, 表明模型预测能力越强, 精确度越高。当 RPD > 2.0 时表明该模型具有较高的可靠性, 当 RPD > 2.5 时表明模型预测效果极好, 能够实际应用于定量分析^[23-24]。

1.5 数据处理

采用 Microsoft Excel 2016 进行光谱数据和 β -葡聚糖含量的统计, 使用 Origin 2024 计算裸大麦资源 β -葡聚糖含量的平均值、最大值、最小值、极差、标准偏差、变异系数和正态检验, 并进行验证集预

测值与化学值的回归分析,制作相关分析图;采用 Matlab 2024a 软件进行光谱图绘制、异常样本剔除、样本集划分、光谱预处理、特征变量筛选和预测模型建立。

2 结果与分析

2.1 裸大麦籽粒 β -葡聚糖含量的测定

利用酶法获得 215 份裸大麦籽粒的 β -葡聚糖含量值(附表 1)。由图 2 可知,215 份裸大麦籽粒的 β -葡聚糖含量范围为 2.88%~6.98%,平均值为 4.08%,极差为 4.10%,表明本研究选取的裸大麦资源的籽粒 β -葡聚糖含量分布范围较广,具有代表性。利用 Kolmogorov-Smirnov 检验法对裸大麦籽粒 β -葡聚糖含量进行正态检验, P 值为 0.399,大于 0.050,表明该样本集的 β -葡聚糖含量符合正态分布,可用于近红外预测模型的建立。

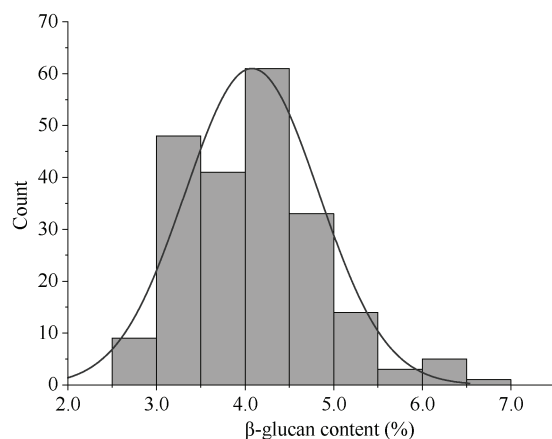


图 2 215 份裸大麦资源 β -葡聚糖含量化学值的分布
Fig. 2 Distribution of β -glucan chemical values in 215 naked barley germplasm resources

2.2 裸大麦籽粒的近红外光谱分析

使用近红外光谱分析仪扫描 215 份裸大麦资源的籽粒,其中,甘垦 5 号、Jet、Tet、22DJ2 和甘垦糯 3 号共 5 份裸大麦材料的籽粒颜色为黑色,扫描结果明显异常,予以剔除。以光谱波长为横轴、吸光度为纵轴绘制剩余 210 份裸大麦籽粒的近红外光谱图。由图 3 可知,参试裸大麦籽粒在 950~1650 nm 区间内的近红外光谱曲线的波峰和波谷变化明显,且变化趋势大致相同,在 992、1204 和 1466 nm 处存在 3 个吸收峰,但不同的样本吸收峰强度不同,表明近红外光谱对裸大麦籽粒 β -葡聚糖含量有较高的区分度,可以用于 β -葡聚糖含量的定量分析。

2.3 裸大麦籽粒 β -葡聚糖含量近红外预测模型的建立

2.3.1 异常值剔除 使用蒙特卡洛交叉验证法来识别光谱数据和 β -葡聚糖含量中潜在的异常值,以平均值为横坐标,标准差为纵坐标来展示所有样本的预测误差分布(图 4)。依据平均值或标准差较高的样本更可能是异常值的原则^[25],以预测误差均值 1.2、标准差 0.4 为界限剔除异常值,编号 45、140、200、201 和 210 共 5 个样本的 β -葡聚糖含量被认定为潜在异常值,需要从数据集中移除,剩余的 205 个样本用于裸大麦籽粒 β -葡聚糖含量近红外预测模型的构建。

2.3.2 样本集划分 剔除异常值后,采用 SPXY 算法按 4 : 1 的比例划分样本集,其中 164 个样本作为校正集,用于建模;剩余 41 个样本作为验证集,用于模型的验证。由表 1 可知,总样本的 β -葡聚糖范围为 2.88%~6.38%,平均值为 4.04%,变异系数为 17.52%,分布范围较广。校正集和验证集的裸大麦

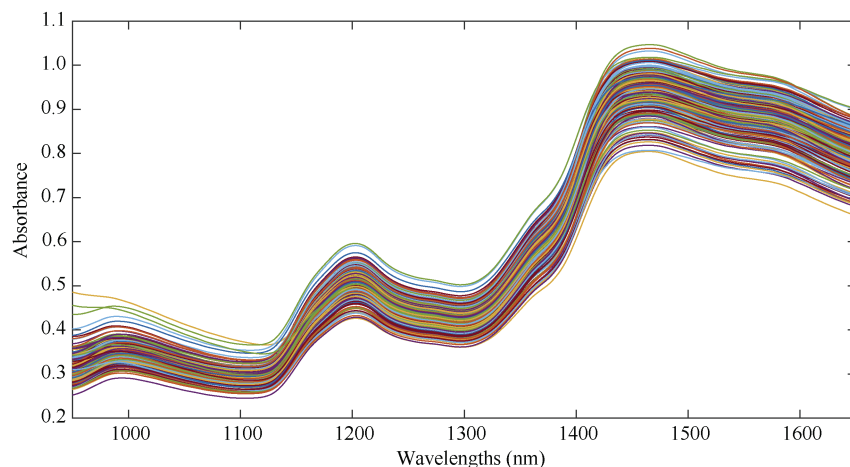


图 3 裸大麦籽粒样品的原始近红外光谱图
Fig. 3 Original near-infrared spectra of 210 naked barley grain samples

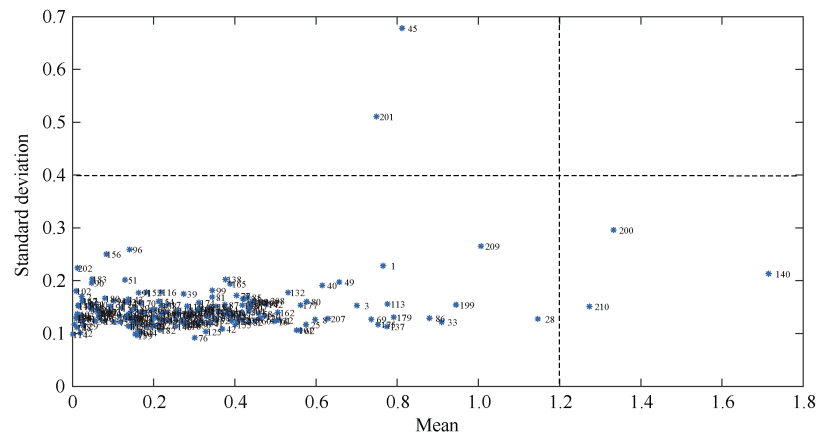


图 4 蒙特卡洛交叉验证法的异常检测结果
Fig. 4 Outlier detection results using the Monte Carlo cross-validation method

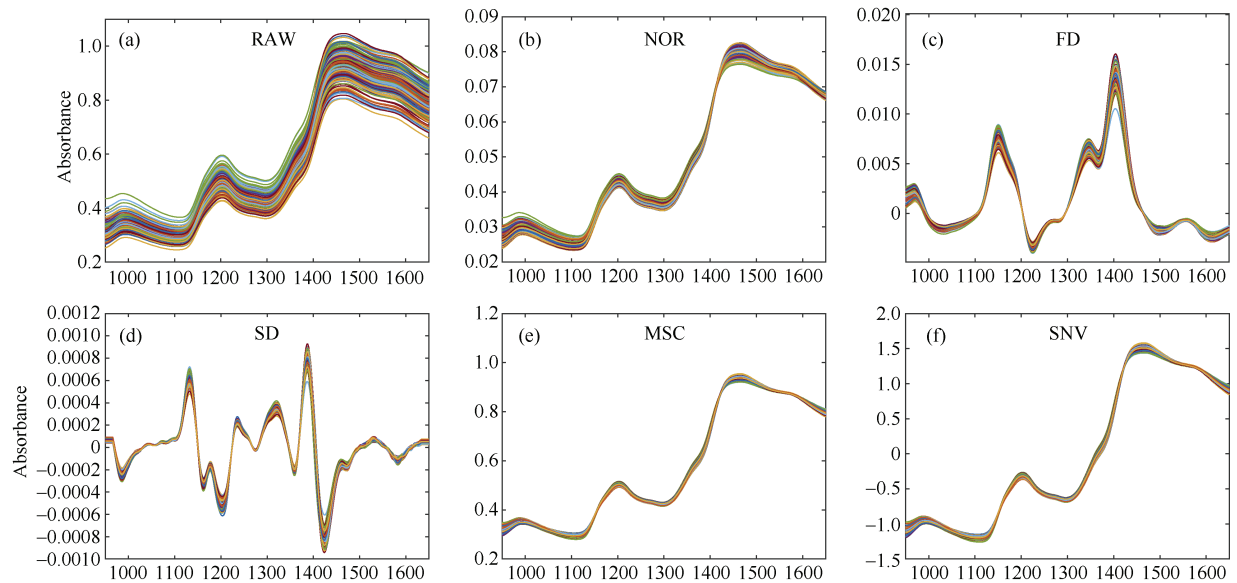
表 1 校正集和验证集样本的 β -葡聚糖含量值
Table 1 Statistics of β -glucan content in calibration and validation sets

样本集 Sample set	样本数 Sample size	平均值 Mean (%)	最大值 Max. (%)	最小值 Min. (%)	极差 Range	标准偏差 SD	变异系数 CV (%)
校正集 Calibration set	164	4.08	6.38	2.88	3.50	0.72	17.58
验证集 Validation set	41	3.90	5.31	2.92	2.39	0.66	16.96
总样本 Total set	205	4.04	6.38	2.88	3.50	0.71	17.52

β -葡聚糖含量范围分别为 2.88%~6.38%和 2.92%~5.31%, 其中校正集包含 β -葡聚糖含量的最大值和最小值, 极差、标准偏差和变异系数均大于验证集, 表明样本集划分合理, 可用于构建预测模型。

2.3.3 光谱数据预处理 使用归一化、一阶导数 (first derivative, FD)、二阶导数(second derivative, SD)、MSC、SNV、SG、SG 一阶求导(SGFD)、SG 二阶求导(SGSD)、去散射、一阶导数+MSC、一阶

导数+SNV、SG+MSC、SG+SNV、SNV+MSC 共 14 种方法对校正集原始光谱进行预处理, 预处理后的光谱图见图 5。与原始光谱相比, SG 卷积平滑处理后的光谱仅与原始光谱存在极小差异, 其余预处理优化后的光谱波峰和波谷更加明显, 光谱曲线更加聚集, 样本间差异性下降。此外, 一阶导数、一阶导数+MSC 和一阶导数+SNV 预处理后的光谱图变化趋势较一致, MSC、SNV、SG+MSC、SG+SNV 和



(图 5)

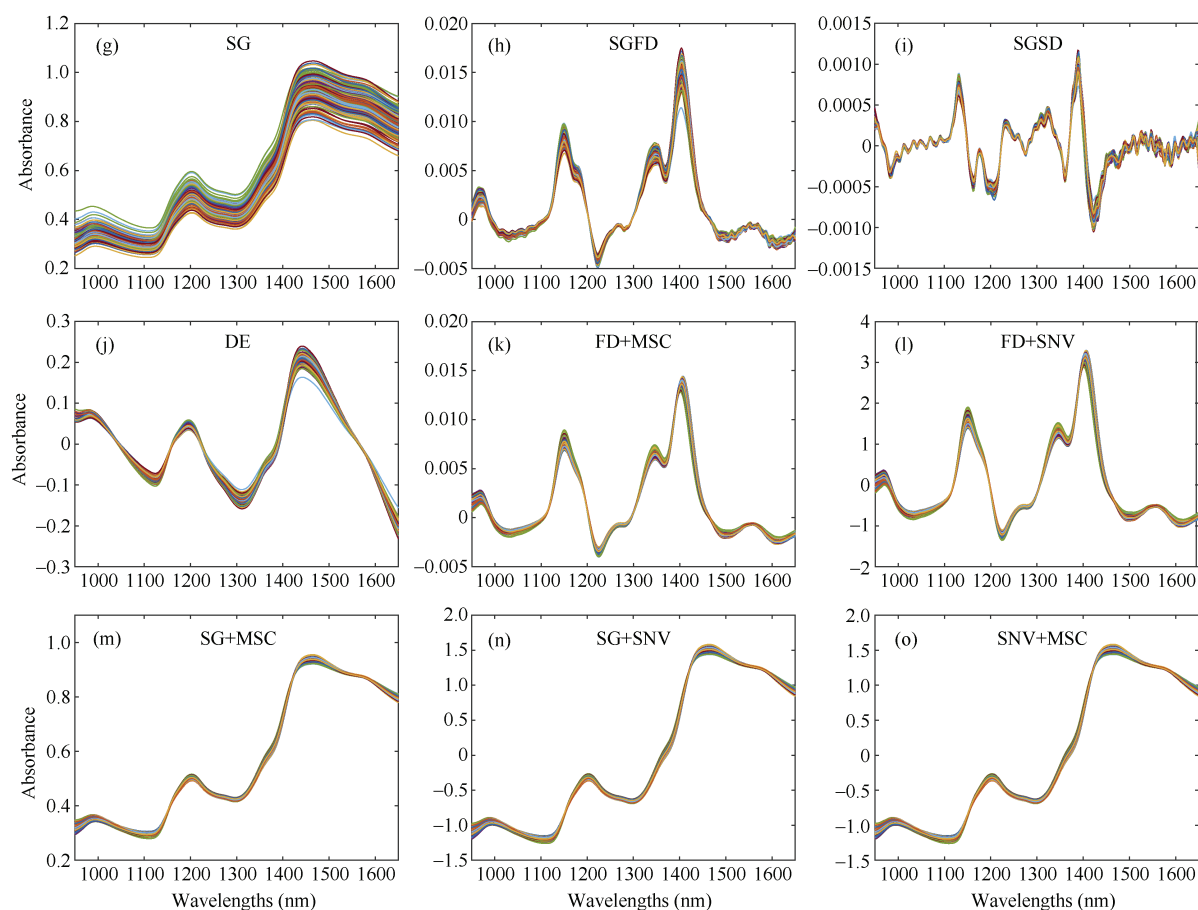


图 5 不同预处理校正集光谱图

Fig. 5 Preprocessed spectra of the calibration set using different methods

a: 原始光谱; b: 归一化; c: 一阶导数; d: 二阶导数; e: 多元散射校正; f: 标准正态校正; g: Savitzky-Golay 卷积平滑; h: SG 卷积平滑一阶求导; i: SG 卷积平滑二阶求导; j: 去散射; k: 一阶导数+多元散射校正; l: 一阶导数+标准正态校正; m: SG 卷积平滑+多元散射校正; n: SG 卷积平滑+标准正态校正; o: 标准正态校正+多元散射校正。

a: raw spectra; b: normalization (NOR); c: first derivative (FD); d: second derivative (SD); e: multiplicative scatter correction (MSC); f: standard normal variate (SNV); g: Savitzky-Golay convolution smoothing (SG); h: Savitzky-Golay convolution smoothing first derivative (SGFD); i: Savitzky-Golay convolution smoothing second derivative (SGSD); j: detrend (DE); k: first derivative + multiplicative scatter correction (MSC); l: first derivative + standard normal variate (SNV); m: Savitzky-Golay convolution smoothing (SG) + multiplicative scatter correction (MSC); n: Savitzky-Golay convolution smoothing (SG) + standard normal variate (SNV); o: standard normal variate (SNV) + multiplicative scatter correction (MSC).

SNV+MSC 预处理后的光谱变化趋势也较为相似。使用原始光谱和预处理后的光谱进行下一步的模型构建。

2.3.4 近红外光谱预测模型的建立 利用 PLSR、SVM 和 PCR 算法对光谱数据进行回归分析, 构建裸大麦籽粒 β -葡聚糖含量的近红外预测模型。由表 2 可知, PLSR 模型中, 除 SGFD 和 SGSD 预处理的模型 RPD 值低于 2, 预测可靠性较差外, 其余预处理后的模型 RPD 值均大于 2, 其中 FD+SNV 处理获得模型的 R^2_p 和 RPD 值最高, 分别是 0.839 和 2.441。所有 PCR 模型的 RPD 值均小于 2, 预测可靠性较低, 仅可用于粗略估计裸大麦籽粒 β -葡聚糖含量, 表明 PCR 算法不适用于参试样本近红外预测模型的建

立。SVM 模型中, SD 处理获得模型的 R^2_c 、 R^2_p 和 RPD 值均最大, 分别是 0.935、0.856 和 2.460, RMSEC 和 RMSEP 最低, 分别为 0.185 和 0.269, 表明该模型可靠性较好, 可用于预测裸大麦籽粒的 β -葡聚糖含量。由于 SD-SVM 模型在所有模型中预测效果最佳, 故将使用不同的特征变量筛选办法对该模型进行进一步优化。

2.3.5 特征变量筛选 为了进一步提高模型的预测能力, 减少数据冗余, 提高运行效率, 采用 SPA、UVE、CARS、VISSA、BOSS 和 IRIV 算法提取 β -葡聚糖含量相关的特征变量。由图 6 可知, 使用 SPA、UVE、CARS、VISSA、BOSS 和 IRIV 算法对二阶导数预处理后的光谱进行特征变量筛选, 分别

表2 不同建模方法预测结果比较

Table 2 Comparison of prediction results using different modeling methods

模型 Model	预处理 Pretreat	校正决定 系数 R^2_c	校正标准 偏差 RMSEC	外部验证决 定系数 R^2_p	验证标准 偏差 RMSEP	相对分析 误差 RPD
PLSR	原始光谱 RAW	0.657	0.419	0.798	0.309	2.143
	归一化 NOR	0.678	0.406	0.799	0.298	2.224
	一阶导数 FD	0.743	0.363	0.808	0.295	2.245
	二阶导数 SD	0.781	0.335	0.779	0.311	2.128
	多元离散校正 MSC	0.710	0.385	0.804	0.306	2.166
	标准正态校正 SNV	0.723	0.376	0.807	0.300	2.204
	Savitzky-Golay 卷积平滑法 SG	0.652	0.422	0.804	0.307	2.157
	Savitzky-Golay 卷积平滑一阶求导 SGFD	0.811	0.311	0.727	0.362	1.830
	Savitzky-Golay 卷积平滑二阶求导 SGSD	0.816	0.307	0.461	0.520	1.274
	去散射 DE	0.717	0.381	0.817	0.292	2.269
	一阶导数+多元离散校正 FD+MSC	0.777	0.338	0.821	0.294	2.252
	一阶导数+标准正态校正 FD+SNV	0.776	0.338	0.839	0.271	2.441
	Savitzky-Golay 卷积平滑+多元离散校正 SG+MSC	0.698	0.393	0.796	0.311	2.131
	Savitzky-Golay 卷积平滑+标准正态校正 SG+SNV	0.711	0.384	0.804	0.301	2.197
	标准正态校正+多元离散校正 SNV+MSC	0.710	0.385	0.804	0.304	2.180
SVM	原始光谱 RAW	0.462	0.536	0.540	0.468	1.414
	归一化 NOR	0.546	0.492	0.601	0.419	1.581
	一阶导数 FD	0.780	0.338	0.826	0.281	2.359
	二阶导数 SD	0.935	0.185	0.856	0.269	2.460
	多元离散校正 MSC	0.928	0.194	0.217	2.932	0.226
	标准正态校正 SNV	0.927	0.195	0.770	0.343	1.931
	Savitzky-Golay 卷积平滑法 SG	0.461	0.536	0.539	0.468	1.413
	Savitzky-Golay 卷积平滑一阶求导 SGFD	0.904	0.226	0.773	0.333	1.990
	Savitzky-Golay 卷积平滑二阶求导 SGSD	0.743	0.375	0.621	0.408	1.624
	去散射 DE	0.614	0.452	0.715	0.351	1.887
	一阶导数+多元离散校正 FD+MSC	0.786	0.334	0.814	0.557	1.188
	一阶导数+标准正态校正 FD+SNV	0.788	0.333	0.812	0.287	2.305
	Savitzky-Golay 卷积平滑+多元离散校正 SG+MSC	0.902	0.225	0.204	2.783	0.238
	Savitzky-Golay 卷积平滑+标准正态校正 SG+SNV	0.902	0.225	0.763	0.346	1.915
	标准正态校正+多元离散校正 SNV+MSC	0.928	0.194	0.768	0.346	1.913
PCR	原始光谱 RAW	0.096	0.680	0.102	0.641	1.032
	归一化 NOR	0.319	0.590	0.429	0.506	1.307
	一阶导数 FD	0.487	0.512	0.543	0.461	1.436
	二阶导数 SD	0.714	0.382	0.641	0.401	1.650
	多元离散校正 MSC	0.454	0.528	0.510	0.471	1.406
	标准正态校正 SNV	0.453	0.529	0.510	0.470	1.409
	Savitzky-Golay 卷积平滑法 SG	0.096	0.680	0.102	0.641	1.032
	Savitzky-Golay 卷积平滑一阶求导 SGFD	0.708	0.387	0.683	0.388	1.705
	Savitzky-Golay 卷积平滑二阶求导 SGSD	0.880	0.248	0.507	0.521	1.271
	去散射 DE	0.407	0.551	0.487	0.490	1.351

(续表 2)

模型 Model	预处理 Pretreat	校正决定 系数 R^2_c	校正标准 偏差 RMSEC	外部验证决 定系数 R^2_p	验证标准 偏差 RMSEP	相对分析 误差 RPD
PCR	一阶导数+多元离散校正 FD+MSC	0.533	0.489	0.623	0.429	1.544
	一阶导数+标准正态校正 FD+SNV	0.609	0.447	0.676	0.382	1.730
	Savitzky-Golay 卷积平滑+多元离散校正 SG+MSC	0.454	0.528	0.510	0.471	1.406
	Savitzky-Golay 卷积平滑+标准正态校正 SG+SNV	0.454	0.528	0.510	0.470	1.409
	标准正态校正+多元离散校正 SNV+MSC	0.454	0.528	0.510	0.470	1.408

RAW: 原始光谱; NOR: 归一化; FD: 一阶导数; SD: 二阶导数; MSC: 多元离散校正; SNV: 标准正态校正; SG: Savitzky-Golay 卷积平滑; SGFD: Savitzky-Golay 卷积平滑一阶求导; SGSD: Savitzky-Golay 卷积平滑二阶求导; DE: 去散射; PLSR: 偏最小二乘回归; SVM: 支持向量机; PCR: 主成分回归。

RAW: raw spectra; NOR: normalization; FD: first derivative; SD: second derivative; MSC: multiplicative scatter correction; SNV: standard normal variate; SG: Savitzky-Golay convolution smoothing; SGFD: Savitzky-Golay convolution smoothing first derivative; SGSD: Savitzky-Golay convolution smoothing second derivative; DE: detrend; PLSR: partial least squares regression; SVM: support vector machine; PCR: principal component regression; R^2_c : correction determination coefficient; RMSEC: root mean square error of calibration; R^2_p : coefficient of determination for external validation; RMSEP: root mean square error of prediction; RPD: ratio of performance to deviation.

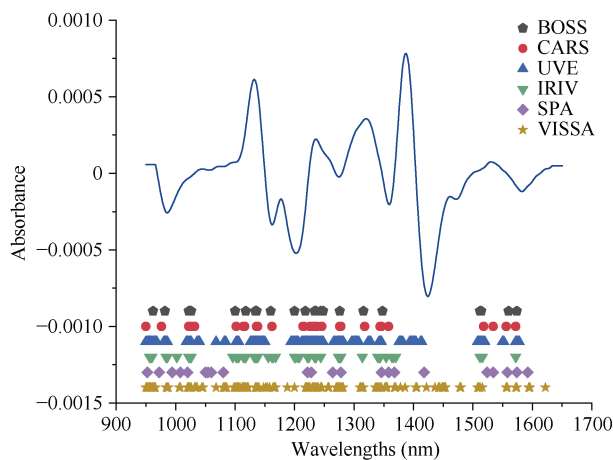


图 6 基于 6 种算法的特征波长筛选
Fig. 6 Feature wavelength selection results based on six algorithms

SPA: 连续投影算法; UVE: 无信息变量剔除法; CARS: 竞争性自适应重加权采样法; VISSA: 变量迭代空间收缩法; BOSS: 自举软收缩法; IRIV: 迭代保留信息变量法。
SPA: successive projections algorithm; UVE: uninformative variables elimination; CARS: competitive adaptive reweighted sampling; VISSA: variable iterative space shrinking algorithm; BOSS: bootstrapping soft shrinkage; IRIV: iteratively retained informative variables.

筛选到 22、114、28、115、23 和 47 个波段, 分别占原始波段的 6.27%、32.48%、7.98%、32.76%、6.55% 和 13.39%。不同算法筛选获得的特征变量主要集中于 950~1026 nm、1100~1138 nm、1194~1248 nm、1270~1282 nm、1340~1354 nm、1508~1518 nm 和 1572~1578 nm。

2.3.6 近红外光谱预测模型的优化 利用 SVM

算法对特征变量筛选后的光谱数据进行回归分析, 优化后的模型预测效果见表 3。其中 SPA-SD-SVM、UVE-SD-SVM、VISSA-SD-SVM、BOSS-SD-SVM 和 IRIV-SD-SVM 模型的 R^2_p 和 RPD 值均低于全波长的 SVM 模型, 预测效果出现不同程度的下降。与 FULL-SD-SVM 模型相比, CARS-SD-SVM 模型的 R^2_p 从 0.856 提高到 0.877, 提高了 2.45%, RMSEP 从 0.269 降低到 0.237, 降低了 11.90%, RPD 值从 2.460 提高到 2.790, 提高了 13.41%, 预测效果优于 FULL-SD-SVM, 可用于定量分析裸大麦籽粒的 β -葡聚糖含量。CARS 筛选特征光谱的过程见图 7, 在 25 次采样中, 交互验证均方根误差(root mean square error of cross-validation, RMSECV)最小为 0.3291, 将原始的 351 个变量减少到 28 个, 减少了 92.02%, 极大地提高了运行效率。同时, 对 CARS-SD-SVM 模型的预测结果与化学测量结果做线性图(图 8), 发现验证集各样本 β -葡聚糖含量的预测值与化学测定值之间具有较好的线性关系, 能满足近红外用于裸大麦籽粒 β -葡聚糖含量的高通量、无损、快速检测。

3 讨论

β -葡聚糖含量极大影响裸大麦的用途, 是培育专用型裸大麦品种的重要改良指标。由于 β -葡聚糖含量的常用化学测定方法酶法成本较高、存在安全隐患, 且会对籽粒造成破坏, 分析一批样品须 4 h 以上, 难以对裸大麦资源和育种中间材料的 β -葡聚糖

表 3 不同特征波长筛选方法建立的 SVM 模型预测结果
Table 3 Prediction results of SVM models based on different feature selection methods

方法 Method	变量数 Variable number	校正决定系数 R^2_C	校正标准偏差 RMSEC	外部验证决定系数 R^2_P	验证标准偏差 RMSEP	相对分析误差 RPD
FULL-SD-SVM	351	0.935	0.185	0.856	0.269	2.460
SPA-SD-SVM	22	0.786	0.334	0.772	0.321	2.060
UVE-SD-SVM	114	0.789	0.338	0.817	0.281	2.357
CARS-SD-SVM	28	0.859	0.272	0.877	0.237	2.790
VISSA-SD-SVM	115	0.818	0.311	0.822	0.287	2.304
BOSS-SD-SVM	23	0.809	0.316	0.817	0.294	2.249
IRIV-SD-SVM	47	0.883	0.250	0.803	0.316	2.097

FULL: 全波长光谱。缩写同表 2 和图 6。
FULL: full length spectra. Abbreviations are the same as those given in Table 2 and Fig. 6.

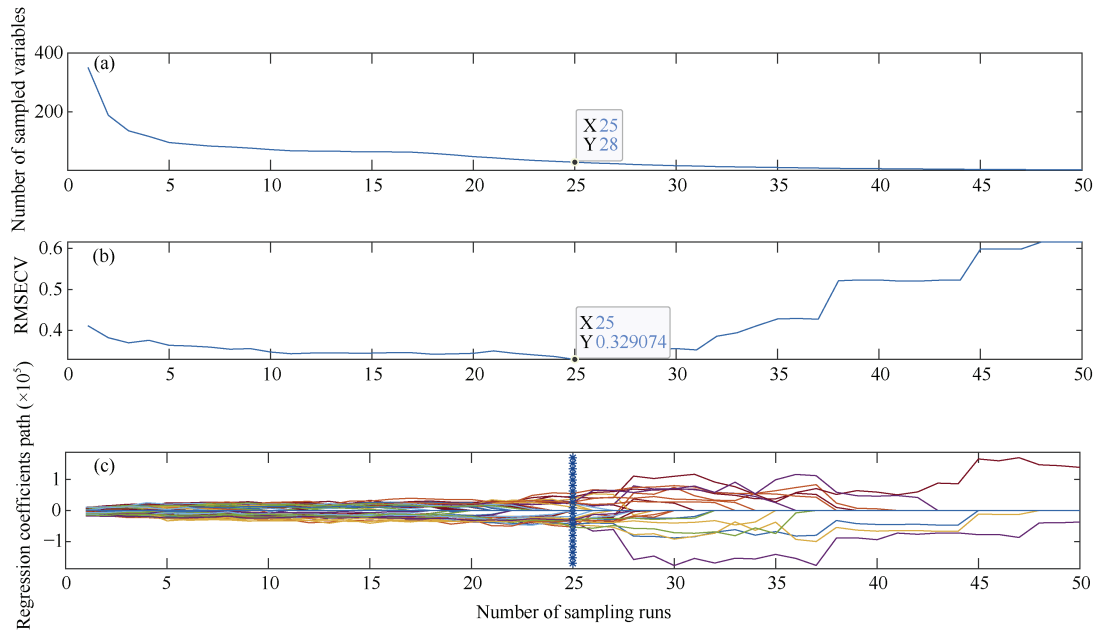


图 7 二阶导数预处理后 CARS 筛选特征变量的过程和变量分布图
Fig. 7 Process and variable distribution of CARS-based feature selection after second derivative preprocessing
(a): 选择变量数量的变化; (b): RMSECV 的变化; (c): CARS 算法计算过程中各变量的回归系数。RMSECV: 交互验证均方根误差。
(a): variation in selected variable count; (b): changes in RMSECV; (c): regression coefficients of selected variables during the CARS process.
RMSECV: root mean square error of cross-validation.

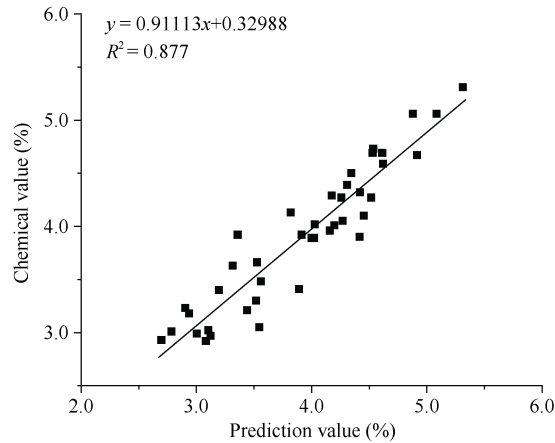


图 8 验证集 β -葡聚糖含量预测值与化学值的相关性
Fig. 8 Correlation between predicted and chemically measured β -glucan content in the validation set

含量进行高通量测定。近红外预测模型因其成本低、测定速率快、操作简便等优势已经成功应用于大麦蛋白质、酚类、花青素等多个营养指标的测定^[26-28]。本研究利用波通 DA7250 近红外分析仪构建了裸大麦籽粒 β -葡聚糖含量的预测模型, 结合特征变量筛选对模型进行优化, 其中 CARS-SD-SVM 模型的相对分析误差(RPD)为 2.790, 模型预测质量最高, 可用于高通量测定裸大麦资源和育种中间材料的 β -葡聚糖含量^[24]。

虽然前人已经构建了裸大麦籽粒 β -葡聚糖含量的近红外预测模型, 如 Schmidt 等^[11]和 Li 等^[13]分别使用 106 份和 120 份裸大麦样品建立了籽粒 β -葡聚糖含量的预测模型, 但是这些研究使用的样本量相

对较少,可能无法较好地反映大量样本的情况。本研究以 215 份裸大麦资源为材料构建了近红外预测模型,样本量较大,且 β -葡聚糖含量范围为 2.88%~6.38%,基本覆盖了常规资源,适用于近红外预测模型的建立。本研究使用 SPXY 法对样本集进行划分,SPXY 法是从 Kennard-Stone 法的基础上发展而来的,划分样本时能综合考虑光谱值和化学测定值,使所筛选出的样本对于 β -葡聚糖含量的校正集样本更具有代表性和适用性^[29]。由于样本种类、营养指标、仪器等差异,不同预处理方法的处理效果存在较大差异^[30]。杨焰婷等^[12]使用 3 种单一的预处理方法对青稞光谱数据进行处理, Xiong 等^[31]使用 6 种单一的预处理方法对大麦光谱数据进行预处理。本研究使用归一化、一阶导数、二阶导数等 14 种单一或组合的预处理方法对原始光谱进行处理,使用的预处理方法较多,可能在减小或消除光谱数据的噪声、散射的干扰和背景基线漂移等的影响,提高预测模型的质量和预测准确度等方面存在较大的优势。本研究构建的 CARS-SD-SVM 模型的外部验证决定系数为 0.877,高于 Schmidt 等^[11]的 0.848 和 Li 等^[13]的 0.763,模型预测质量表现较好,可用于裸大麦育种中间材料的 β -葡聚糖含量的高通量无损测定,这可能是参试样品数量较多、样本划分方法合理、预处理和特征筛选方法更多样所致。

本研究发现黑色裸大麦籽粒样本光谱扫描结果明显异常,这可能是由于深色物体对近红外光的吸收能力较强,导致反射光不足,进而影响了扫描精度^[14]。根据不同的籽粒颜色建立不同的预测模型,可能是提高模型预测精度的有效途径^[32]。后续可继续收集黑色裸大麦资源,通过建立黑色籽粒专用的 β -葡聚糖含量预测模型对现有的模型进行丰富和补充。此外,将黑色裸大麦籽粒磨成粉样或使用更新型的近红外分析仪也可能减少深色种皮对近红外光的干扰。本研究使用的波通 DA7250 近红外分析仪的光谱范围为 950~1650 nm, Yang 等^[33]筛选获得与燕麦 β -葡聚糖含量相关的特征变量(1000~1023 nm、1075~1136 nm、1316~1436 nm)均在该光谱范围内,然而 Albanell 等^[28]发现 2260~2380 nm 也对大麦 β -葡聚糖含量的贡献较大,后续还需使用光谱范围更广的近红外分析仪对本研究所构建的近红外模型进行进一步优化,以提高模型预测的准确性。本研究参试材料的 β -葡聚糖含量范围仅为 2.88%~6.38%, β -葡聚糖含量极高的裸大麦样本较少,因此构建的预

测模型在含量分布两端(尤其是高值区)的预测性能尚不明确,无法较好地服务专用型裸大麦品种的选育。后续还需补充一批 β -葡聚糖含量极高的裸大麦材料对现有模型进行补充完善,以提高该模型的适用性,进而为裸大麦专用型品种选育奠定基础。

4 结论

本研究利用 215 份国内外裸大麦资源建立了籽粒 β -葡聚糖含量的近红外预测模型,其中 CARS-SD-SVM 模型的校正决定系数(R^2_c)为 0.859,校正标准偏差(RMSEC)为 0.272,外部验证决定系数(R^2_p)为 0.877、验证标准偏差(RMSEP)为 0.237,相对分析误差(RPD)为 2.790,预测效果最佳且可靠性较高,可对裸大麦籽粒的 β -葡聚糖含量进行定量分析,为推进裸大麦的资源评价、专用型品种选育及品质检测提供了有力的技术支持。

附表 请见网络版: 1) 本刊网站 <http://zwxb.chinacrops.org/>; 2) 中国知网 <http://www.cnki.net/>; 3) 万方数据 <http://c.wanfangdata.com.cn/Periodical-zuowxb.aspx>。

References

- [1] 白羿雄, 郑雪晴, 姚有华, 等. 青稞种质资源表型性状的遗传多样性分析及综合评价. 中国农业科学, 2019, 52: 4201–4214. Bai Y X, Zheng X Q, Yao Y H, et al. Genetic diversity analysis and comprehensive evaluation of phenotypic traits in hulless barley germplasm resources. Sci Agric Sin, 2019, 52: 4201–4214 (in Chinese with English abstract).
- [2] 侯殿志, 沈群. 我国 29 种青稞的营养及功能组分分析. 中国食品学报, 2020, 20: 289–298. Hou D Z, Shen Q. Analysis of nutrition and functional components of 29 kinds of highland barley in China. J Chin Inst Food Sci Technol, 2020, 20: 289–298 (in Chinese with English abstract).
- [3] Meints B, Hayes P M. Breeding naked barley for food, feed, and malt. Plant Breed Rev, 2019, 43: 95–119.
- [4] Iaccarino N, Khakimov B, Mikkelsen M S, et al. Structurally different mixed linkage β -glucan supplements differentially increase secondary bile acid excretion in hypercholesterolaemic rat faeces. Food Funct, 2020, 11: 514–523.
- [5] Zhang J J, Zheng M J, Zhou L Y, et al. Oat β -glucan alleviates muscle atrophy via promoting myotube formation and suppressing protein degradation. J Sci Food Agric, 2023, 103: 6252–6262.
- [6] Wu Y S, Shao Y, Shao X Y, et al. Qingke β -glucan and Lactobacillus mitigate neuroinflammation and enhance cognitive function in an Alzheimer's disease mouse model. Int J Biol Macromol, 2025, 319: 145427.
- [7] McClear B V, Glennie-Holmes M. Enzymic quantification of (1 to 3) (1 to 4)- β -D-glucan in barley and malt. J Inst Brew, 1985, 91: 285–295.

- [8] 胡美玲, 鄧晨阳, 薛晓梦, 等. 单粒花生蔗糖含量近红外预测模型的建立. 作物学报, 2023, 49: 2498–2504.
Hu M L, Zhi C Y, Xue X M, et al. Establishment of near-infrared reflectance spectroscopy model for predicting sucrose content of single seed in peanut. Acta Agron Sin, 2023, 49: 2498–2504 (in Chinese with English abstract).
- [9] 王若楠, 张颖星, 于筱菡, 等. 基于近红外快速检测技术的谷子淀粉多样性分析及模型构建. 作物学报, 2025, 51: 1757–1768.
Wang R N, Zhang Y X, Yu X H, et al. Near-infrared spectroscopic evaluation of starch diversity and model construction in foxtail millet. Acta Agron Sin, 2025, 51: 1757–1768 (in Chinese with English abstract).
- [10] Xu Y H, Dong Y, Liu J M, et al. Combination of near infrared spectroscopy with characteristic interval selection for rapid detection of rice protein content. J Food Compos Anal, 2025, 137: 106995.
- [11] Schmidt J, Gergely S, Schönellechner R, et al. Comparison of different types of NIR instruments in ability to measure β -glucan content in naked barley. Cereal Chem, 2009, 86: 398–404.
- [12] 杨焰婷, 江谧, 詹瑞玲, 等. 青稞特征营养成分含量快速检测模型的建立及优化. 食品工业科技, 2024, 45(18): 228–238.
Yang Y T, Jiang M, Zhan R L, et al. Development and optimization of a rapid detection model for characteristic nutrient content in highland barley. Sci Technol Food Ind, 2024, 45(18): 228–238 (in Chinese with English abstract).
- [13] Li L L, Li L, Gou G Y, et al. Feasibility analysis of nondestructive detection of multiple parameters of highland barley by near-infrared spectroscopy. J Sci Food Agric, 2025, 105: 5194–5206.
- [14] 李鑫瑜, 侯名语, 崔顺立, 等. 红色和黑色花生籽仁黄酮含量近红外模型的构建. 中国农业科学, 2025, 58: 1284–1295.
Li X Y, Hou M Y, Cui S L, et al. Construction of near infrared spectrometry model for flavonoids content of peanut with red and black testa. Sci Agric Sin, 2025, 58: 1284–1295 (in Chinese with English abstract).
- [15] Liu J M, Luo X, Zhang D J, et al. Rapid determination of rice protein content using near-infrared spectroscopy coupled with feature wavelength selection. Infrared Phys Technol, 2023, 135: 104969.
- [16] Galvão R K H, Araujo M C U, José G E, et al. A method for calibration and validation subset partitioning. Talanta, 2005, 67: 736–740.
- [17] Zhang W, Feng W W, Cai Z Q, et al. A deep one-dimensional convolutional neural network for microplastics classification using Raman spectroscopy. Vib Spectrosc, 2023, 124: 103487.
- [18] Deng G M, Liu H G, Li J Q, et al. Rapid prediction of nucleosides content and origin traceability of *Boletus bainiugan* using Fourier transform near-infrared spectroscopy combined with chemometrics. Spectrochim Acta A Mol Biomol Spectrosc, 2025, 327: 125421.
- [19] Khatun M S, Al Masum A, Islam M H, et al. Short wave-near infrared spectroscopy for predicting soluble solid content in intact mango with variable selection algorithms and chemometric model. J Food Compos Anal, 2024, 136: 106745.
- [20] Jiang H, Wang J N, Chen Q S. Comparison of wavelength selected methods for improving of prediction performance of PLS model to determine aflatoxin B₁ (AFB₁) in wheat samples during storage. Microchem J, 2021, 170: 106642.
- [21] 李杨, 李翠玲, 王秀, 等. 高光谱成像的黄瓜病虫害识别和特征波长提取方法. 光谱学与光谱分析, 2024, 44: 301–309.
Li Y, Li C L, Wang X, et al. Identification of cucumber disease and insect pest based on hyperspectral imaging. Spectrosc Spectr Anal, 2024, 44: 301–309 (in Chinese with English abstract).
- [22] Deng B C, Yun Y H, Cao D S, et al. A bootstrapping soft shrinkage approach for variable selection in chemical modeling. Anal Chim Acta, 2016, 908: 63–74.
- [23] 邵园园, 王永贤, 玄冠涛, 等. 基于高光谱成像的肥城桃品质可视化分析与成熟度检测. 农业机械学报, 2020, 51(8): 344–350.
Shao Y Y, Wang Y X, Xuan G T, et al. Visual detection of sse and firmness and maturity prediction for Feicheng peach by using hyperspectral imaging. Trans CSAM, 2020, 51(8): 344–350 (in Chinese with English abstract).
- [24] Li P, Jin Q T, Liu H M, et al. Determination of soluble solids content in loquat using near-infrared spectroscopy coupled with broad learning system and hybrid wavelength selection strategy. LWT, 2024, 206: 116570.
- [25] Zhang H L, Zhan B S, Pan F, et al. Determination of soluble solids content in oranges using visible and near infrared full transmittance hyperspectral imaging with comparative analysis of models. Postharvest Biol Technol, 2020, 163: 111148.
- [26] Lin C, Chen X, Jian L, et al. Determination of grain protein content by near-infrared spectrometry and multivariate calibration in barley. Food Chem, 2014, 162: 10–15.
- [27] Han Z G, Cai S G, Zhang X L, et al. Development of predictive models for total phenolics and free p-coumaric acid contents in barley grain by near-infrared spectroscopy. Food Chem, 2017, 227: 342–348.
- [28] Albanell E, Martínez M, De Marchi M, et al. Prediction of bioactive compounds in barley by near-infrared reflectance spectroscopy (NIRS). J Food Compos Anal, 2021, 97: 103763.
- [29] 李栓明, 郭银巧, 王克如, 等. 小麦籽粒蛋白质光谱特征变量筛选方法研究. 中国农业科学, 2015, 48: 2317–2326.
Li S M, Guo Y Q, Wang K R, et al. Research on variable selection of wheat kernel protein content with near-infrared spectroscopy. Sci Agric Sin, 2015, 48: 2317–2326 (in Chinese with English abstract).
- [30] Yu Y, Chai Y H, Li Z T, et al. Quantitative predictions of protein and total flavonoids content in Tartary and common buckwheat using near-infrared spectroscopy and chemometrics. Food Chem, 2025, 462: 141033.
- [31] Xiong C H, She Y X, Jiao X, et al. Rapid nondestructive hardness detection of black highland barley kernels via hyperspectral imaging. J Food Compos Anal, 2024, 127: 105966.
- [32] 陈森, 侯名语, 崔顺立, 等. 不同种皮颜色花生糖含量近红外模型的构建. 光谱学与光谱分析, 2022, 42: 2896–2902.
Chen M, Hou M Y, Cui S L, et al. Construction of near-infrared model of peanut sugar content in different seed coat colors. Spectrosc Spectr Anal, 2022, 42: 2896–2902 (in Chinese with English abstract).
- [33] Yang Z Y, Cheng Z, Su P Y, et al. A model for the detection of β -glucan content in oat grain based on near infrared spectroscopy. J Food Compos Anal, 2024, 129: 106105.