邹慧颖,王东薇,樊懿楷,等.基于中红外光谱和机器学习算法的牛奶中β-乳球蛋白快速检测方法[J].华中农业大学学报,2025,44(2):125-133.DOI:10.13300/j.cnki.hnlkxb.2025.02.013

基于中红外光谱和机器学习算法的牛奶中β-乳球蛋白快速检测方法

邹慧颖[1],王东薇[1],樊懿楷[1],刘维华[2],杨俊华[2],余文莉[3],SABEK Ahmed Abdalla Ahmed Ibrahim[3],张淑君[1]

1.华中农业大学动物科学技术学院、动物医学院,武汉 430070;2.宁夏兽药饲料监察所,银川 750000;3.石家庄天泉良种奶牛有限公司,石家庄 050061;

摘要 为建立一种可以快速、批量、高效检测中国荷斯坦牛牛奶中β-乳球蛋白含量的方法,采集501份来自西北、华北和华中主要产奶地区的健康中国荷斯坦牛牛奶样本,采用高效液相色谱法测定牛奶样本中β-乳球蛋白的含量,并同步测定和收集牛奶样本中红外光谱数据(mid-infrared spectroscopy,MIRS)。以MIRS为预测变量,β-乳球蛋白含量为因变量,将12种光谱预处理方法进行连续2次的随机组合,并手动选取特征波段,使用偏最小二乘回归(partial least squares regression,PLSR)作为传统机器学习算法,建立预测牛奶中β-乳球蛋白含量的最优预测模型。结果显示:该模型交叉验证集和测试集的R_C^2和R_P^2分别为0.812 9、0.768 8,均方根误差$RMSE_C$和$RMSE_P$分别为0.476 2、0.524 9 g/L,性能偏差比(ratio of performance to deviation,RPD)为2.076 6,达到畜禽生产性能的测定要求。试验结果表明,可以利用MIRS建立模型预测中国荷斯坦牛牛奶中的β-乳球蛋白含量。

关键词 中红外光谱;牛奶;β-乳球蛋白;机器学习算法;光谱预处理

中图分类号 S823 **文献标识码** A **文章编号** 1000-2421(2025)02-0125-09

牛奶是人类营养物质的重要来源之一,已成为人们生活和农业经济的重要组成部分。牛奶中的营养物质主要是由水分、蛋白质、脂肪、乳糖、维生素和矿物质等组成[1]。牛奶含有30～36 g/L总蛋白质,约占牛奶总量的3.30%,分为酪蛋白和乳清蛋白[2],其中20%的蛋白质主要由α-乳白蛋白(0.60～1.70 g/L)和β-乳球蛋白(2～4 g/L)组成[3]。研究[4-6]表明,β-乳球蛋白具有与脂肪酸或维生素结合的能力,能够作为脂溶性维生素的载体,还具有一定的抗氧化能力和免疫功能。牛奶是人体获取β-乳球蛋白的重要来源,β-乳球蛋白作为一种多功能蛋白具有一定的营养价值。因此,准确测定牛奶中β-乳球蛋白含量和确定牛奶中β-乳球蛋白含量的影响因素对奶牛养殖业和食品产业发展,以及人类的生命健康具有重大意义。

当前用于乳清蛋白含量检测的分析技术主要有高效液相色谱法(HPLC)[7]、酶联免疫吸附测定法(ELISA)[8]、毛细管电泳法(CZE)[9]等。然而,这些方法并不适用于牛奶的规模化和常规化检测。中红外光谱(mid-infrared spectroscopy,MIRS)是一种快速、批量、无损耗、无污染且具有成本效益的技术,广泛应用于常规乳成分检测和奶牛生产性能测定[10](dairy herd improvement,DHI)。目前,许多研究也开始关注牛奶中的精细化成分及奶牛生理状态的测定,如奶牛血液代谢物、牛奶中维生素的含量、脂肪酸组成、矿物质含量、奶牛基因型、奶牛能量状态、奶牛妊娠状态和奶牛甲烷排放等[11-13]。MIRS作为一种高通量、低成本的检测工具,应用于牛奶中β-乳球蛋白含量的快速测定有极大的潜力。国内的DHI测定主要是针对牛奶中乳蛋白、乳脂、乳糖、总固形物

收稿日期:2024-05-08
基金项目:国家重点研发计划项目(2023YFD1300400);中央高校基本科研业务费专项(2662023DKPY001);石家庄市科技计划项目(221500182A)
邹慧颖,E-mail:1360717697@qq.com
通信作者:张淑君,E-mail:sjxiaozhang@mail.hzau.edu.cn

和尿素氮等5个常规指标，并且使用的预测模型均来自国外，不同国家或地区牛奶的MIRS特征存在较大差异，可能会对测定结果产生一定的影响。

目前针对牛奶中β-乳球蛋白含量的快速、批量检测方法研究较少，国外的学者从2009年开始尝试将传统机器学习方法与测定的牛奶的MIRS结合建立β-乳球蛋白的定量检测模型[14]，并在2016年将该方法应用于多品种奶牛牛奶中β-乳球蛋白含量检测模型的建立[15-17]。然而，目前尚无检测中国荷斯坦奶牛奶中β-乳球蛋白含量的MIRS预测模型。因此，本试验通过在我国采集具有代表性和多样性的牛奶样品，测定牛奶样品的MIRS数据和牛奶中β-乳球蛋白含量，结合二者建立β-乳球蛋白含量的预测模型，同时比较不同预处理方法对牛奶中β-乳球蛋白含量预测的准确性，以期为建立具有我国自主产权的适合我国牛奶中物质成分含量的MIRS定量预测模型提供参考。

1 材料与方法

1.1 牛奶的采集和样品的分装

从我国西北、华北和华中主要产奶地区的5个奶牛场（牧场A、牧场B、牧场C、牧场D、牧场E）采集2022年4—11月健康状况良好的奶牛奶样共343份，从西北4个区域的多个牧场混合罐装奶样中采集2022年3—6月的混合奶样158份，共采集501份牛奶样品。

利用自动挤奶装置完成牛奶采集工作，每份牛奶采集约100 mL，分装到采样瓶中，依次编号，并向每个采样瓶中立即加入溴硝丙二醇防腐剂，缓慢摇晃使其充分溶解。运回途中在奶样周围放置冰袋（2～4 ℃）防止变质。样本到达DHI实验室后，立即进行光谱测定和采集。完成光谱采集后剩余的牛奶样本倒入离心管中（50～55 mL），置于－20 ℃冰箱中保存，用于测定β-乳球蛋白的含量。

1.2 牛奶中β-乳球蛋白含量的检测

1）牛奶样品前处理。将在－20 ℃冰箱冷冻保存的奶样取出，放置在4 ℃冰箱低温过夜解冻，直至奶样完全解冻。将解冻后的奶样取出放置在常温环境，混匀。取1 mL混匀的牛奶样品于50 mL离心管中加入超纯水定容至50 mL，滴加稀释后的乙酸溶液（超纯水与乙酸1∶1混合）使样品的pH值约为4.60，混匀静置1 h。将静置后的奶样以8 000 r/min的转速离心5 min，用1 mL无菌注射器吸取离心后的奶样上清液1 mL，缓慢用微孔滤膜将上清液过滤到进样瓶内，待测。

2）牛奶中β-乳球蛋白含量测定方法。使用高效液相色谱仪对牛奶样本中的β-乳球蛋白含量进行检测，仪器条件和测定步骤参照NY/T 1450—2007《中国荷斯坦牛生产性能测定技术规范》和《高效液相色谱法同时测定巴氏杀菌乳中α-乳白蛋白和β-乳球蛋白》[18-19]。通过高效液相色谱仪测定的β-乳球蛋白含量被称为“真实值”。

1.3 牛奶中红外光谱的采集与测定

牧场A和牧场B采集的所有奶样在宁夏DHI中心进行光谱采集，牧场C、牧场D、牧场E采集的奶样在新疆DHI中心进行光谱采集，宁夏各地牧场采集的奶样在宁夏兽药饲料监察进行光谱采集，均使用丹麦FOSS公司的MilkoScan^{TM}FT＋乳成分分析仪测定。牛奶样品的光谱测定过程：将在保温箱低温保存（2～4 ℃）的新鲜奶样取出后放置在架子上，于45 ℃水浴锅内恒温预热30 min，将预热好的奶样摇匀后放在检测传送带上，打开每个牛奶样品瓶盖，对奶样依次进行检测，检测结果输出牛奶的MIRS、牛奶常规乳成分（乳脂、乳蛋白、乳糖、总固形物和尿素氮）及牛奶体细胞数数据。牛奶样品测定结束后，在乳成分分析仪配套的软件上将采集的MIRS数据导出。

1.4 异常值的筛选

本试验共采集501份涵盖3个季节（因新冠疫情无法采集冬季奶样）、1～6胎次、泌乳阶段0～305 d等条件的具有代表性和多样性的牛奶样品，每个样品均设置平行重复样；从501份奶样中，剔除MIRS的马氏距离（Mahalanobis distance，MD）＜3、牛奶量不足及真实值异常（$\frac{|X_1-X_2|}{(X_1+X_2)/2}\geqslant 10\%$，$X$为真实值）的无效数据54份，获得用于建模和验证的样品数量如表1所示。

1.5 模型的建立

1）数据集划分。本研究中β-乳球蛋白样品量为447份，先在所有数据中随机抽取5个样品作为模型外部验证集，剩余的数据中75%用于建模，建模过程使用交叉验证调整模型参数，为交叉验证集，25%为测试集，即交叉验证集与测试集的比例为3∶1。结合测定真实值对数据进行异常值剔除，训练模型的过程中进行10折交叉验证，即从验证集中随机移除10%的数据，使用剩余数据建立的模型对移除的数

表 1　有效样品数量及时间分布

Table 1　Effective sample size and time distribution

采样时间/(年-月) Sampling time/(year-month)	牧场A Farm A	牧场B Farm B	牧场C Farm C	牧场D Farm D	牧场E Farm E	宁夏 NX
2022-03	0	0	0	0	0	42
2022-04	15	0	1	10	6	36
2022-05	17	0	10	8	9	52
2022-06	16	0	10	10	10	28
2022-07	15	18	0	0	0	0
2022-08	12	11	0	0	0	0
2022-09	17	20	0	0	0	0
2022-10	17	20	0	0	0	0
2022-11	17	20	0	0	0	0

据进行预测分类。该过程重复10次获得所有记录的预测结果，以确保结果的稳健性和泛化能力。

2）建模光谱数据预处理方法及特征提取。遵循比尔定律，在建立模型之前，通过$A=\lg(1/T)$将以透射率表示的光谱数据转换为吸光度。在建立预测模型前对光谱进行有效预处理，目的是为去除光谱采集过程中环境、仪器及操作引起的系统误差。本研究采用无标准化（none）、归一化（min-max scaling，MMS）、标准化（standardscaler，SS）、均值中心化（mean-center，MC）、标准正态变量变换（standard normal variate transformation，SNV）、移动平均平滑（moving average，MA）、卷积平滑（savitzky golay，SG）、一阶差分（first difference method，D1）、二阶差分（second order difference，D2）、趋势校正（detrend correction，DT）、多元散射校正（multiplicative scatter correction，MSC）、小波变换（wavelet transform，WAVE）12种方法对光谱数据进行连续2次的特征预处理。

3）建模方法。将牛奶中β-乳球蛋白含量的真实值作为因变量，经过预处理及手动选取特征波段后的光谱作为预测变量进行建模，本研究主要使用偏最小二乘回归（partial least squares regression，PLSR）建模算法。

1.6　模型的评价指标

本试验利用交叉验证集决定系数（coefficient of determination of calibration，R_C^2）、交叉验证集均方根误差（root mean square error of calibration，$RMSE_C$，公式中用x表示）、测试集决定系数（coefficient of determination of prediction set，R_P^2）、测试集均方根误差（root mean squared error of prediction，$RMSE_P$，公式中用x表示）和性能偏差比（ratio of performance to deviation，RPD，公式中用m表示）综合评价模型性能以筛选出最佳模型。具体公式如下：

$$R^2=1-\frac{\sum_{n=1}^{N}\left(y_n-\hat{y}_n\right)^2}{\sum_{n=1}^{N}\left(y_n-\bar{y}\right)^2}$$

$$x=\sqrt{\frac{\sum_{n=1}^{N}\left[\left(y_n-\hat{y}_n\right)^2\right]}{N}}$$

$$m=\frac{\mathrm{STD}_{EV}}{x}$$

其中y_n和$\hat{y}_n$分别代表3种乳成分数据集的真实值和预测值。$\bar{y}$是y值的平均值，N代表样本数量，STD_{EV}代表样本的标准差。

2　结果与分析

2.1　牛奶中β-乳球蛋白含量的描述性统计

由表2可知，交叉验证集和测试集牛奶中β-乳球蛋白的含量分别为2.97和2.79 g/L；牛奶中β-乳球蛋白含量具有明显的变异性，变异系数分别为37.04%和39.07%，表明样本具有一定的多样性和代表性，可用于建立模型。

表 2　牛奶中β-乳球蛋白的含量

Table 2　The content of β-lactoglobulin in milk

样本集 Sample Set	数据量/个 Number of samples	平均值/(g/L) Mean	标准差/(g/L) SD	变异系数/% CV	最小值/(g/L) Min	最大值/(g/L) Max
交叉验证集 Cross validation set	331	2.97	1.10	37.04	0.25	5.38
测试集 Test set	111	2.79	1.09	39.07	0.37	5.14

2.2　基于牛奶中MIRS的β-乳球蛋白含量预测模型的建立

1）建模MIRS数据预处理方法的筛选。本研究主要采用材料与方法中“1.6第2）部分建模光谱数据预处理方法及特征提取”中提及的12种MIRS数据预处理方法及其相互组合对光谱数据进行特征预处

理。将样品MIRS全波段进行二次预处理，利用PLSR算法建立模型，比较模型效果。如表3所示，选择较优的预处理组合。

表3 较优预处理组合选择结果

Table 3 Optimal preprocessing combination selection results

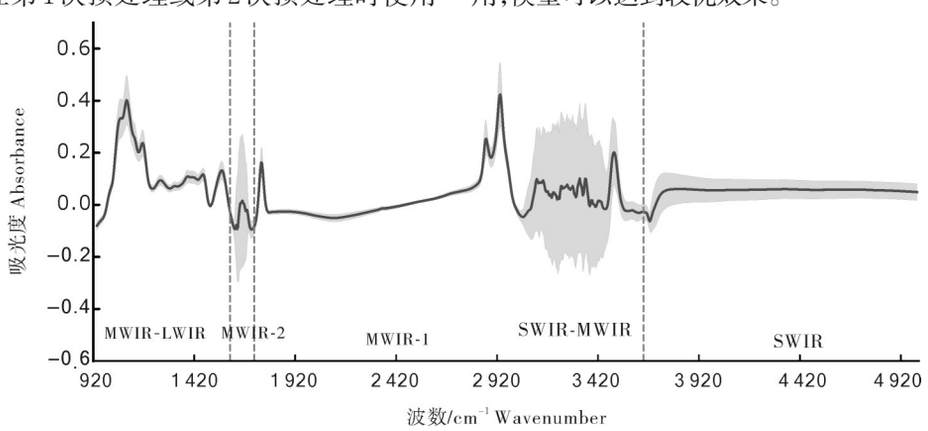
第二次预处理 Second preprocessing	交叉验证集 Cross validation set		测试集 Test set	
	R_C^2	$RMSE_C$/(g/L)	R_P^2	$RMSE_P$/(g/L)
None	0.647 1	0.653 9	0.420 0	0.831 3
MMS	0.647 1	0.653 9	0.420 0	0.831 3
SS	0.647 1	0.653 9	0.420 0	0.831 3
CT	0.658 2	0.643 5	0.434 3	0.821 1
SNV	0.653 3	0.648 0	0.367 1	0.868 4
MA	0.534 5	0.750 9	0.404 6	0.842 3
SG	0.552 2	0.736 5	0.423 3	0.829 0
MSC	0.232 2	0.964 4	-	11.483 9
D1	0.884 9	0.373 4	0.377 7	0.861 1
D2	0.876 7	0.386 5	0.218 5	0.965 0
DT	0.686 1	0.616 7	0.465 3	0.798 3
WAVE	0.644 3	0.656 3	0.390 1	0.852 5

注："-"表示R^2接近于0；第一次预处理均使用"SS"方法。Note: '-' means R^2 is close to 0, the 'SS' method was used for the first preprocessing.

12种方法建立的牛奶中β-乳球蛋白含量预测模型在交叉验证集上的R_C^2为0.232 2～0.890 4，$RMSE_C$为0.364 4～0.964 4 g/L；测试集上的R_P^2为0～0.466 5，$RMSE_P$为0.797 3～11.483 9 g/L。综合比较各项评价指标，发现在第1次预处理或第2次预处理时使用D1或D2方法交叉验证集效果较好，但测试集效果一般。为避免过拟合现象出现，并保持模型训练结果较好，同时测试结果有较大提升空间，发现使用SS+D1预处理组合时建模效果优于其他组合（R_C^2=0.884 9，$RMSE_C$=0.373 4 g/L，R_P^2=0.377 7，$RMSE_P$=0.861 1 g/L）。

2）建模特征波段的筛选。牛奶样本的平均光谱图见图1。牛奶的MIRS由925～5 008 cm^{-1}范围内的1 060个单独的波点组成，大致分为短波红外区（short-wavelength infrared，SWIR）、中波红外区（mid-wavelength infrared，MWIR）和长波红外区（long-wavelength infrared，LWIR）3个区域[18]。

本研究采用手动选择方法对MIRS特征波段进行选择，在确定预处理为SS+D1的基础上，人工调整选取的波段位置，调整出较优的模型。如图2所示，β-乳球蛋白模型选取了16段特征波段：999.222～1 130.394，1 172.832～1 284.714，1 400.454～1 547.058，1 759.248～1 917.426，1 940.574～2 056.314，2 152.764～2 191.344，2 245.356～2 361.096，2 581.002～2 716.032，3 236.862～3 240.72，3 263.868～3 371.892，3 398.898～3 526.212，3 973.74～3 985.314，4 236.084～4 363.398，4 459.848～4 583.304，4 594.878～4 718.334，4 776.204～4 976.82 cm^{-1}。每一段前后允许有2个波点的差距。结果发现模型中多波段与多波点的协同作用，模型可以达到较优效果。



MWIR-LWIR：中波红外区和长波红外区的过渡区域；MWIR-2：中波红外区区域2；MWIR-1：中波红外区区域1；SWIR-MWIR：短波红外区和中波红外区的过渡区域；SWIR：短波红外区。MWIR-LWIR: The transition region between the medium-wave infrared region and the long-wave infrared region; MWIR-2: Medium-wave infrared region 2; MWIR-1: Medium-wave infrared region 1; SWIR-MWIR: The transition region between the short-wave infrared region and the medium-wave infrared region; SWIR: Short-wave infrared region.

图1 牛奶样本的平均光谱

Fig.1 Mean spectra of milk samples

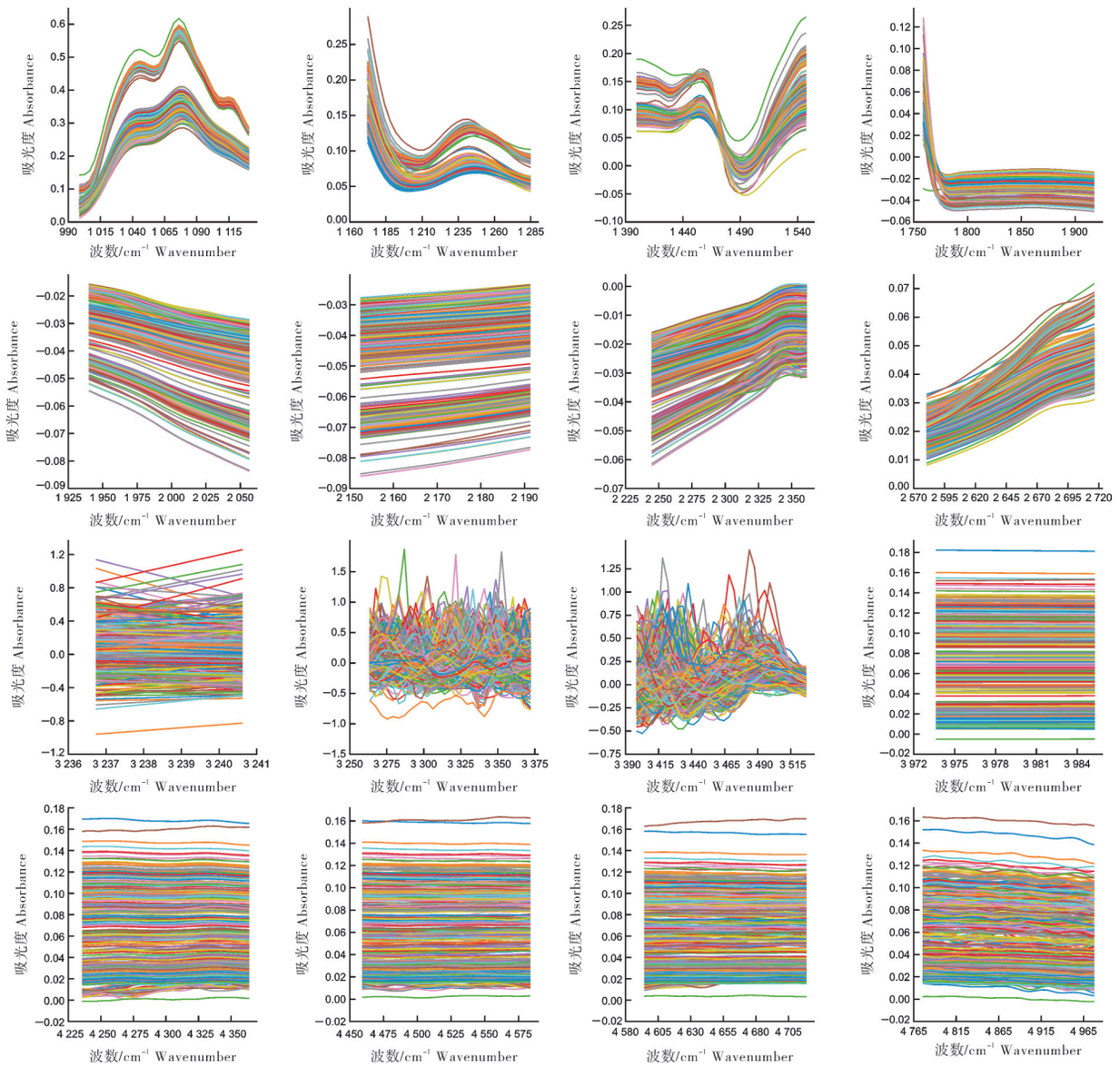


图 2　β-乳球蛋白特征波段光谱

Fig.2　Characteristic spectral map of β-lactoglobulin

3)模型参数及较优模型的筛选。模型参数包括模型预处理方法的参数以及算法的参数，由于本研究选择的较优模型使用 SS+D1 这 2 种无参数的预处理方法，所以模型的主要参数为偏最小二乘回归算法的参数，即主成分(n_component)，参数选择结果对比见表 4。

由表 4 可知，对于 β-乳球蛋白的模型在主成分 n_component=15 时效果最好，R_C^2 和 R_P^2 分别为 0.812 9、0.768 8，$RMSE_C$ 和 $RMSE_P$ 分别为 0.476 2、0.524 9 g/L，RPD 为 2.076 6，即 β-乳球蛋白含量预测较优模型为 SS+D1+PLSR(n_component=15)。

2.3　模型预测效果的验证

利用建立的较优模型对 5 个外部验证样本进行

表 4　不同主成分的建模效果

Table 4　Modeling effect of different principal components

主成分(n_component)	交叉验证集 Cross validation set		测试集 Test set		
	决定系数 R_C^2	均方根误差 $RMSE_C$/(g/L)	决定系数 R_P^2	均方根误差 $RMSE_P$/(g/L)	性能偏差比 RPD
13	0.793 7	0.499 9	0.720 5	0.577 2	1.888 4
14	0.805 0	0.486 2	0.749 8	0.546 0	1.996 3
15	0.812 9	0.476 2	0.768 8	0.524 9	2.076 6
16	0.822 8	0.463 2	0.747 9	0.548 1	1.988 7
17	0.830 5	0.453 2	0.736 7	0.560 1	1.946 1

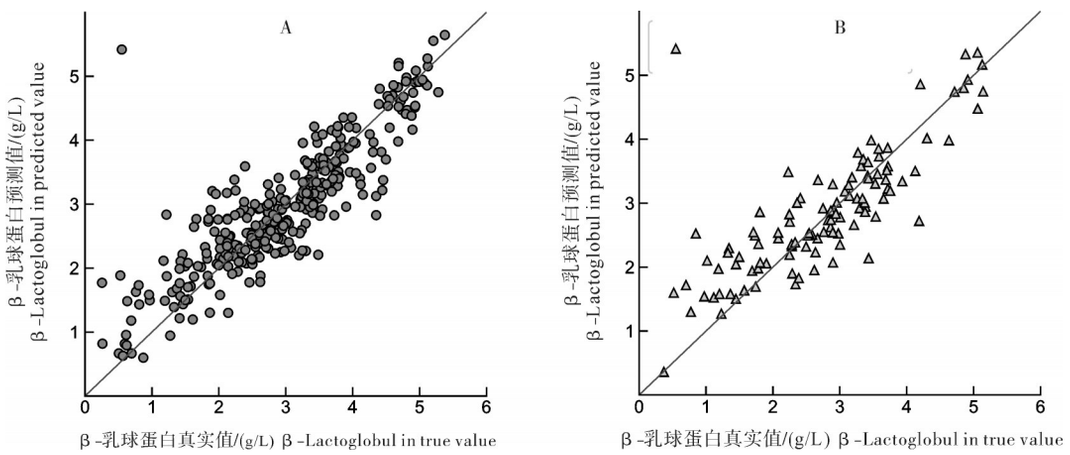
乳球蛋白含量预测，结果见表5。预测偏差的绝对值分别为0.174 5、0.119 0、0.066 2、0.069 3、0.111 9，预测差比分别为4.69%、3.34%、1.93%、2.03%、3.04%，平均预测差比为3.006%。该模型预测的β-乳球蛋白含量准确性较高，可用于牛奶的β-乳球蛋白含量预测。

图3是β-乳球蛋白较优模型得出的真实值和预测值的线性拟合图，可以看出交叉验证集、测试集的真实值和预测值的散点图均围绕在$y=x$（真实值等于预测值）线附近，表明预测含量和真实含量之间存在明显的相关性，这个模型在试验数据范围内预测能力较好。

表5 预测模型的预测验证效果

Table 5 Effects of prediction model

外部验证集 External validation	真实值/(g/L) Reference value	预测值/(g/L) Predicted value	预测偏差/(g/L) Prediction bias	预测差比/% Prediction difference ratio
1	3.720 9	3.895 4	0.174 5	4.69
2	3.564 0	3.683 0	0.119 0	3.34
3	3.425 1	3.491 3	0.066 2	1.93
4	3.422 0	3.352 7	0.069 3	2.03
5	3.686 8	3.574 9	0.111 9	3.04



A：交叉验证集；B：测试集。A：Cross-validation set；B：Test set.

图3 β-乳球蛋白模型真实值与预测值的线性拟合

Fig.3 Linear fitting of true and predicted values of β-lactoglobulin

3 讨 论

3.1 最优模型的选择

中红外光谱数据不仅包含生物样品中的化合物相关信息信号，还包括来自环境背景、高频噪声、基线偏移和重叠谱带的非信息信号[20]。因此，要准确获得样品本身的光谱信息不仅需要在进行试验前注意基线漂移问题，还需要在建立预测模型前对光谱进行预处理及特征波段的选择，以减弱各种非目标因素对光谱的影响及简化后续建模处理运算过程，提高预测准确度[21]。本研究中对光谱进行了12种预处理方法的144种组合预处理，广泛比较了多种预处理之间的优劣性，选择SS+D1的预处理方式。标准化处理能够消除尺度差异过大带来的不良影响，而一阶导数通过消除恒定基线来提高光谱分辨率，减小仪器背景或偏移对信号的影响。

PLSR算法是对MIRS这类线性强、多特征数据建模最有效的算法之一，同时这种算法很少出现过拟合情况，因此PLSR是利用MIRS建立定量预测模型最广泛的传统机器学习算法[22-24]，它通过特征光谱降维和线性回归构建特征矩阵并对奶成分进行预测。

选取特征波段的方法有很多，主要包括算法选取特征与手动选取特征两类[25]。笔者所在实验室前期的研究结果[26]表明针对乳清蛋白进行建模时采用手动选取特征的方法能够达到更好的建模效果，其优点在于选择的过程中可以强化波段（即相邻波点）的作用，兼顾波点之间的共线性问题，同时可以在提升模型的过程中保留更多的光谱原始信息状态，包容性与泛化能力更强，选取波段准确；缺点是选择速度慢，效率低。在目前的研究中，为了提高特征选择的效率，很多学者采用算法选取特征的方法，但这种方法忽略了相邻波点之间的协同作用，思路较为单一。因此，为了保证模型的性能能够达到更高水平，本研究采取了手动选取特征波段的方法。

3.2　基于牛奶中 MIRS 的β-乳球蛋白含量预测模型的准确性

本研究首次建立了基于MIRS的中国荷斯坦牛牛奶中β-乳球蛋白含量的预测模型，模型的R_C^2和R_P^2分别为0.812 9、0.768 8，$RMSE_C$和$RMSE_P$分别为0.476 2、0.524 9 g/L，RPD为2.076 6，对外部验证样本进行预测的平均预测差比为3.006%。

在之前的大部分研究中，PLSR方法被广泛地应用于β-乳球蛋白预测模型的建立，其R_P^2范围为0.34～0.64，$RMSE_P$范围为0.05～2.70 g/L，RPD的范围为0.80～1.66[17,24]。可以看出，本研究建立的牛奶中β-乳球蛋白含量预测模型的R_P^2和RPD更高，但RMSE也略高于其他学者的研究[14-15]。R^2衡量了预测值对于真实值的拟合程度，R^2的值越接近1，则模型的预测值越接近真实值，拟合效果越好。RMSE可以用于衡量模型预测值与真实值之间的差异，可以更直观地表达模型预测误差的大小[27-28]。Christophe等[29]研究结果表明，比较模型效果时使用RPD更有意义，且RPD＞2时表明预测模型可实际应用。外部验证集中的样本真实值与预测值的预测偏差小，预测差比均＜5%，说明所建模型的预测精确性较高[28-29]。因此，本研究建立的牛奶中β-乳球蛋白含量预测模型优于其他研究结果[30-31]，有实际应用的潜力。预测效果有差异可能是由于样本来源、真实值的测量方法、预处理方法和特征选择波段等有差异导致[32-33]。之前的研究中很少对MIRS进行预处理或仅做简单的求导处理，同时在建模之前仅去除以高噪声为特征的水吸收区域甚至使用全波段，而本研究使用的二次预处理方法和手工选择16段特征波段可能更大程度上减少了非信息信号的干扰，准确性更高[24]。

本研究比较了不同预处理方法对牛奶中β-乳球蛋白含量预测的准确性，在国内率先利用牛奶的MIRS建立牛奶中β-乳球蛋白预测模型。结果表明基于MIRS的牛奶中β-乳球蛋白模型预测准确性高，能在试验数据范围内预测β-乳球蛋白含量，且建模之前应用SS＋D1的预处理方法和手动选择特征波段是提高预测模型精度的有效方法。本研究建立的模型性能虽有一定的潜在应用价值，但其主要对中国荷斯坦奶牛一个品种样本进行建立，故还需要增加不同地区和不同品种牛奶样品以增加建模样品的多样性，提高模型的准确性、稳健性和通用性；同时，使用不同建模技术和策略对β-乳球蛋白含量的预测模型进一步完善和优化，为我国奶牛生产性能DHI测定指标的拓展和提高牛奶品质等提供技术支撑。

参考文献 References

[1] SINGHAL S, BAKER R D, BAKER S S. A comparison of the nutritional value of cow′s milk and nondairy beverages[J]. Journal of pediatric gastroenterology and nutrition, 2017, 64(5): 799-805.

[2] ZHANG L N, BOEREN S, HAGEMAN J A, et al. Perspective on calf and mammary gland development through changes in the bovine milk proteome over a complete lactation[J]. Journal of dairy science, 2015, 98(8): 5362-5373.

[3] RIJNKELS M. Multispecies comparison of the casein gene loci and evolution of casein gene family[J]. Journal of mammary gland biology and neoplasia, 2002, 7(3): 327-345.

[4] YANG M C, GUAN H H, LIU M Y, et al. Crystal structure of a secondary vitamin D_3 binding site of milk beta-lactoglobulin[J]. Proteins, 2008, 71(3): 1197-1210.

[5] SAKAI K, SAKURAI K, SAKAI M, et al. Conformation and stability of thiol-modified bovine beta-lactoglobulin[J]. Protein science, 2000, 9(9): 1719-1729.

[6] GHALANDARI B, DIVSALAR A, ESLAMI-MOGHADAM M, et al. Probing of the interaction between β-lactoglobulin and the anticancer drug oxaliplatin[J]. Applied biochemistry and biotechnology, 2015, 175(2): 974-987.

[7] BOBE G, BEITZ D, FREEMAN A, et al. Separation and quantification of bovine milk proteins by reversed-phase high-performance liquid chromatography[J]. Journal of agricultural and food chemistry, 1998, 46(2): 458-463.

[8] HE S F, LI X, GAO J Y, et al. Development of sandwich ELISA for testing bovine β-lactoglobulin allergenic residues by specific polyclonal antibody against human IgE binding epitopes[J]. Food chemistry, 2017, 227: 33-40.

[9] HECK J M L, OLIEMAN C, SCHENNINK A, et al. Estimation of variation in concentration, phosphorylation and genetic polymorphism of milk proteins using capillary zone electrophoresis[J]. International dairy journal, 2008, 18(5): 548-555.

[10] BATRA B, NARWAL V, KALRA V, et al. Folic acid biosensors: a review[J]. Process biochemistry, 2020, 92: 343-354.

[11] DADOUSIS C, ABLONDI M, CIPOLAT-GOTET C, et al. Genomic inbreeding coefficients using imputed genotypes: assessing different estimators in Holstein-Friesian dairy cows[J]. Journal of dairy science, 2022, 105(7): 5926-5945.

[12] DANIEL J B, FRIGGENS N C, CHAPOUTOT P, et al. Milk yield and milk composition responses to change in predicted net energy and metabolizable protein: a meta-analysis[J]. Animal, 2016, 10(12): 1975-1985.

[13] BENEDET A, FRANZOI M, PENASA M, et al. Prediction of blood metabolites from milk mid-infrared spectra in early-

lactation cows[J]. Journal of dairy science, 2019, 102(12): 11298-11307.

[14] DE MARCHI M, BONFATTI V, CECCHINATO A, et al. Prediction of protein composition of individual cow milk using mid-infrared spectroscopy[J]. Italian journal of animal science, 2009, 8(sup2): 399-401.

[15] ESKILDSEN C E, SKOV T, HANSEN M S, et al. Quantification of bovine milk protein composition and coagulation properties using infrared spectroscopy and chemometrics: a result of collinearity among reference variables[J]. Journal of dairy science, 2016, 99(10): 8178-8186.

[16] MCDERMOTT A, VISENTIN G, DE MARCHI M, et al. Prediction of individual milk proteins including free amino acids in bovine milk using mid-infrared spectroscopy and their correlations with milk processing characteristics[J]. Journal of dairy science, 2016, 99(4): 3171-3182.

[17] NIERO G, PENASA M, GOTTARDO P, et al. Selecting the most informative mid-infrared spectra wavenumbers to improve the accuracy of prediction models for detailed milk protein content[J]. Journal of dairy science, 2016, 99(3): 1853-1858.

[18] 胡敏,张彩云,李志君,等.高效液相色谱法同时测定巴氏杀菌乳中α-乳白蛋白和β-乳球蛋白[J].乳业科学与技术,2021,44(6):15-19. HU M, ZHANG C Y, LI Z J, et al. Simultaneous determination of α-lactalbumin and β-lactoglobulin in pasteurized milk by HPLC[J]. Journal of dairy science and technology, 2021, 44(6): 15-19(in Chinese with English abstract).

[19] 中华人民共和国农业农村部.中国荷斯坦牛生产性能测定技术规范:第4部分 生产性能测定内容和要求:NY/T 1450—2007[S].北京:中国标准出版社,2007. Ministry of Agriculture and Rural Affairs of the People's Republic of China. Technical specification of Chinese holstein cattle performance test: part 4 contents and requirements of production performance measurement: NY/T 1450—2007[S]. Beijing: China Standards Press, 2007(in Chinese).

[20] RINNAN Å, VAN DEN BERG F, ENGELSEN S B. Review of the most common pre-processing techniques for near-infrared spectra[J]. Trends in analytical chemistry, 2009, 28(10): 1201-1222.

[21] ARRONDO J L R, MUGA A, CASTRESANA J, et al. Quantitative studies of the structure of proteins in solution by Fourier-transform infrared spectroscopy[J]. Progress in biophysics and molecular biology, 1993, 59(1): 23-56.

[22] SOYEURT H, DARDENNE P, DEHARENG F, et al. Genetic parameters of saturated and monounsaturated fatty acid content and the ratio of saturated to unsaturated fatty acids in bovine milk[J]. Journal of dairy science, 2008, 91(9): 3611-3626.

[23] RUTTEN M J M, BOVENHUIS H, HETTINGA K A, et al. Predicting bovine milk fat composition using infrared spectroscopy based on milk samples collected in winter and summer[J]. Journal of dairy science, 2009, 92(12): 6202-6209.

[24] BRESOLIN T, DÓREA J R R. Infrared spectrometry as a high-throughput phenotyping technology to predict complex traits in livestock systems[J/OL]. Frontiers in genetics, 2020, 11: 923[2024-05-08]. https://doi.org/10.3389/fgene.2020.00923.

[25] ZOU X B, ZHAO J W, POVEY M J W, et al. Variables selection methods in near-infrared spectroscopy[J]. Analytica chimica acta, 2010, 667(1/2): 14-32.

[26] 张淑君,张静静,樊懿楷,等.牛奶中α-乳白蛋白的中红外快速批量检测方法:CN202111283550.5[P].2024-02-20. ZHANG S J, ZHANG J J, FAN Y K, et al. Mid-infrared rapid batch detection method for α-lactalbumin in milk: CN202111283550.5[P]. 2024-02-20(in Chinese).

[27] 蒋宏霖,刘会杰,王学杰,等.近红外光谱技术在烟叶化学分析中的应用[J].科技与创新,2019(18):153-155. JIANG H L, LIU H J, WANG X J, et al. Application of near infrared spectroscopy in chemical analysis of tobacco leaves[J]. Journal of technology and innovation, 2019(18): 153-155(in Chinese)

[28] 孙东永,王义民,黄强,等.均方根误差最小准则的水库群典型年选取[J].西安理工大学学报,2011,27(3):275-279. SUN D Y, WANG Y M, HUANG Q, et al. The selection of typical years of reservoir group based on the smallest criterion of root-mean-square error[J]. Journal of Xi'an University of Technology, 2011, 27(3): 275-279(in Chinese with English abstract).

[29] CHRISTOPHE O S, GRELET C, BERTOZZI C, et al. Multiple breeds and countries' predictions of mineral contents in milk from milk mid-infrared spectrometry[J/OL]. Foods, 2021, 10(9): 2235[2024-05-08]. https://doi.org/10.3390/foods10092235.

[30] BONFATTI V, DEGANO L, MENEGOZ A, et al. Mid-infrared spectroscopy prediction of fine milk composition and technological properties in Italian Simmental[J]. Journal of dairy science, 2016, 99(10): 8216-8221.

[31] RUTTEN M J M, BOVENHUIS H, HECK J M L, et al. Predicting bovine milk protein composition based on Fourier transform infrared spectra[J]. Journal of dairy science, 2011, 94(11): 5683-5690.

[32] GRELET C, DARDENNE P, SOYEURT H, et al. Large-scale phenotyping in dairy sector using milk MIR spectra: key factors affecting the quality of predictions[J]. Methods, 2021, 186: 97-111.

[33] XIAO S J, WANG Q H, LI C F, et al. Rapid identification of A_1 and A_2 milk based on the combination of mid-infrared spectroscopy and chemometrics[J/OL]. Food control, 2022, 134: 108659[2024-05-08]. https://doi.org/10.1016/j.foodcont.2021.108659.

A mid-infrared spectroscopy and machine learning algorithm-based method for rapidly detecting content of β-lactoglobulin in milk

ZOU Huiying[1], WANG Dongwei[1], FAN Yikai[1], LIU Weihua[2], YANG Junhua[2],
YU Wenli[3], SABEK Ahmed Abdalla Ahmed Ibrahim[3], ZHANG Shujun[1]

1.*College of Animal Science and Technology, College of Veterinary Medicine, Huazhong Agricultural University, Wuhan 430070, China*;
2.*Ningxia Institute of Veterinary Drug and Feed Supervision, Yinchuan 750000, China*;
3.*Shijiazhuang Tianquan Breeding Dairy Co., Ltd., Shijiazhuang 050061, China*;

Abstract　501 milk samples of healthy Chinese Holstein cows were collected from major milk-producing regions in Northwest, North, and Central China to establish a method that can rapidly, in batch, and efficiently detect the content of β-lactoglobulin in milk from Chinese Holstein cows, high-performance liquid chromatography (HPLC) was used to determine the content of β-lactoglobulin in milk samples, and the mid-infrared spectroscopy (MIRS) data of milk samples were synchronously measured and collected 12 methods of spectra pretreatment were randomly combined twice in a row, and the characteristic bands were manually selected with MIRS as the predictor variable and the content of β-lactoglobulin as the dependent variable. Partial least squares regression (PLSR) was used as a traditional machine learning algorithm to establish an optimal model for the prediction of the content of β-lactoglobulin in milk. The results showed that the R_C^2 and R_P^2 of the cross validation set and test set in the established model was 0.812 9 and 0.768 8, with the root mean square errors, $RMSE_C$ and $RMSE_P$ of 0.476 2 g/L and 0.524 9 g/L, the RPD of 2.076 6, meeting the requirements for measuring the production performance of livestock and poultry. It is indicated that MIRS can be used to establish a model for predicting the content of β-lactoglobulin in milk from Chinese Holstein cows.

Keywords　mid-infrared spectroscopy (MIRS); milk; β-lactoglobulin; machine learning algorithm; spectra pretreatment

（责任编辑：胡　敏）