文章编号:1001 - 9014(2011)06 - 0522 - 04

# 近红外光谱分析中建模光谱宽度的选择

杨皓旻[1,2], 卢启鹏[1*], 黄富荣[3]

(1. 中国科学院长春光学精密机械与物理研究所,应用光学国家重点实验室,吉林 长春 130033;
2. 中国科学院研究生院,北京 100039;3. 暨南大学 光电工程系,广东 广州 510632)

**摘要**:为选择生化样品近红外光谱分析中最优的建模光谱宽度,结合样品特征波长与光谱学原理开展研究. 收集348 份血清样品,以血清中胆固醇、甘油三酯为例进行分析. 根据光谱学原理及 2 mm 厚度血清的光谱吸光度,选择一级倍频区域作为分析波段,考察该区域内不同宽度光谱范围的模型效果. 优选出胆固醇与甘油三酯的最优光谱宽度分别为 70 nm 和 100 nm,并建立相应分析模型,分析的预测标准差和相对预测误差分别是 0.17 mmol/L,3.0% 和 0.14 mmol/L,10.3%,效果不弱于他人建立模型的最好结果,使用的光谱范围更窄.

**关 键 词**:近红外光谱;血清成分测量;偏最小二乘;建模谱宽选择
**中图分类号**:O657.33 **文献标识码**:A

# Selection of spectral width for prediction modeling in near-infrared spectroscopy analysis

YANG Hao-Min[1,2], LU Qi-Peng[1*], HUANG Fu-Rong[3]

(1. State Key Laboratory of Applied Optics, Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China;
2. Graduate School of Chinese Academy of Sciences, Beijing 100039, China;
3. Department of Optoelectronic Engineering, Jinan University, Guangzhou 510632, China)

**Abstract:** In order to select optimal spectral width for biomedical components analysis by near-infrared (NIR) spectroscopy, starting with the characteristic wavelengths of the samples, a modeling procedure based on spectroscopy principles is presented. 348 human serum samples are collected. Cholesterol and triglyceride in human sera are analyzed as an example. According to spectroscopy principles and absorbance of 2 mm-thick sera, first overtone region is selected for analysis. Models with different spectral width are compared. For cholesterol and triglyceride models, optimal spectral width is 70nm and 100nm respectively. Root mean square error of prediction (RMSEP) and mean percent error of prediction (MPEP) are 0.17 mmol/L, 3.0% and 0.14mmol/L, 10.3%, respectively. The performance is comparable with the best works of other groups while the spectral width is narrower.

**Key words:** near-infrared spectroscopy; examination of serum component; partial least squares (PLS); selection of spectral width
**PACS:** 33.20.Ea

## 引言

近红外光谱分析是一种可以无损、原位分析物质成分的方法. 它源于物质中含氢基团振动吸收的合频与倍频[1]. 由于其对样品的非侵入和快捷的分析,符合现场测试与过程控制的特殊要求,在石油化工、烟草品质、食品分析、制药工业、土壤监测等场合均有应用[2-4],也是人体生化指标无创检测最有希望获得成功的方法之一,国内外已经开展了多年研究[5-6].

预测模型是近红外光谱分析具体应用的核心. 模型所用的光谱范围直接影响到模型的精度和适用性. 早期分析的样品体系较为简单,通常是人工挑选特征波长. 随着分析样品越来越复杂,样品成分的特

收稿日期:2011 -03 -04,修回日期:2011 -06 -16 Received date: 2011 -03 -04, revised date: 2011 -06 -16
基金项目:国家自然科学基金(60878052,60938002);应用光学国家重点实验室开放基金(O9Q13FQ090);广东省自然科学基金(8151063201000017)
作者简介:杨皓旻(1985-),男,江西南昌人,博士研究生,主要从事近红外光谱分析工作,E-mail: yanghaomin_3@163.com.
* 通讯作者:E-mail: luqipeng@126.com.

征波长逐渐模糊,人工挑选越来越困难. 化学计量学发展之后,涌现出许多选择建模光谱范围的算法,目的都是选择包含分析物有用的吸收信息光谱,排除干扰信息. 合适的建模光谱宽度是决定预测模型效果的基础.

作为无创生化检测的前期验证和无试剂检测的手段,许多研究小组利用近红外光谱技术测量血清、血浆中生化成分[7-14]. Heise 等对血浆中多种生化成分进行了分析[7]; Arnold 等分析了模拟溶液和血清中的生化成分,对比了合频与一级倍频的效果以及最佳光程[8-9]; 陈华才等建立了血清中多种成分模型[10]. 以上研究都是利用通用光谱仪器,也有一些小组研制了专用的光谱仪并开展了生化分析研究. 在早期使用通用光谱仪研究时,经常使用全谱建立模型,自行搭建系统时,需要考虑系统所用光谱范围与宽度. Saptari 等搭建了一套滤光片系统并测量模拟溶液中的葡萄糖浓度,使用合频区域内约 130 nm 宽的光谱范围,对葡萄糖含量的预测精度达到临床可接受的要求[15]. Olesberg 等搭建了调谐激光二极管的系统,调谐范围是 110 nm,但他们认为葡萄糖在体测量需要至少 250 nm 的光谱调谐宽度,400 nm 则更为理想[16]. 本文结合光谱学原理,以测量血清中胆固醇与甘油三酯含量为例,分析优选出最佳的建模光谱宽度,建立了二者的预测模型. 本文的选择建模谱宽的方法可为设计专用近红外光谱分析仪器光谱范围提供依据.

## 1　实验过程

血清样品来自医院健康检查,14 天内共收集 348 份. 样品生化成分的含量由日立 7600 全自动生化分析仪分析得到,使用福斯 XDS 型光栅光谱仪测量血清的透射光谱. 用 2 mm 样品池,光谱范围 1100 ~2500 nm,仪器带宽 2 nm.

人体血清中胆固醇与甘油三酯的正常浓度分别是 3.1 ~5.7 mmol/L 和 0.3 ~1.7 mmol/L. 收集的样品中,共有 69 份胆固醇浓度高于 5.7 mmol/L,最高 12.13 mmol/L; 有 98 份甘油三酯浓度高于 1.7 mmol/L,最高 12.33 mmol/L. 根据实验目的,浓度分布稀疏的样品予以舍去,最终得到 340 份样品. 227 份作为定标集,113 份作为预测集,具体浓度分布如表 1 所示.

表 1　筛选后样品的胆固醇与甘油三酯浓度情况

Table 1　Cholesterol and triglyceride concentration of sifted samples

| | 样品集 | 最小值/mmol/L | 最大值/mmol/L | 平均值/mmol/L | 标准差/mmol/L |
|---|---|---|---|---|---|
| 胆固醇 | 定标集 | 1.90 | 7.98 | 4.67 | 1.18 |
| | 预测集 | 1.91 | 7.82 | 4.65 | 1.17 |
| 甘油三酯 | 定标集 | 0.31 | 4.38 | 1.38 | 0.75 |
| | 预测集 | 0.34 | 4.1 | 1.37 | 0.72 |

## 2　分析与讨论

由于含氢基团合频与各级倍频吸收强度存在量级上的差异,物质的吸光度在不同光谱区域差别较大,相比之下,样品之间浓度差异引起的吸光度差异较小,如果使用全谱建模,浓度引起的吸光度差异很容易湮没在大的差异中,从而降低模型的预测精度. 实际测量中,光谱还受到非线性因素如仪器响应及散射的影响. 使用较窄的光谱范围有以下优点:

(1) 吸光度变化较平缓,不容易湮没浓度差异,光谱信噪比较高;

(2) 相近波长受到散射等非线性因素的干扰效果接近,在模型中较易排除;

(3) 较窄光谱范围在诸如探测器以及分光方式的选择等方面更为灵活,还可使用 LED 或调谐 LD 等高效的窄带光源,仪器设计更为方便.

因而在保证模型精度的情况下,建模光谱范围越窄越好.

血清中 90% 以上是水. 图 1(a) 是 2 mm 厚的水与几份血清样品的光谱. 从图中可以看到,血清与水的光谱形状极为接近,在 1450 nm 与 1940 nm 附近,水有两处吸收非常强烈,对血清光谱也有影响. 根据光谱学与误差原理可知,在仪器条件稳定的情况下,使用吸光度接近 0.43AU 的光谱运算,得到浓度的误差受仪器噪声的影响最小. 建模时应参考吸光度数值选取合适的波段.

通常情况下,有机物在合频(2000 ~2500 nm)、一级倍频(1400 ~1800 nm)、二级倍频(900 ~1200 nm) 区域均有吸收信息. 对 2 mm 血清,一级倍频处光谱的吸光度最接近于 0.43AU,受到仪器噪声的影响最小. 文献[6]中也有类似的结论.

综合以上考虑,在一级倍频区域内建立预测模型. 图 1(b) 是一级倍频处 2 mm 厚血清与水的局部光谱,区域中 1550 ~1850 nm 光谱范围内吸光度均小于 1,水的吸收在 1670 nm 与 1810 nm 两处有局部极小点,两点之间光谱受水的干扰较小,更容易反映其它成分的信息. 此外,根据通常的吸收光谱原理,1740 nm 是 C-H 键一级倍频吸收的近似峰值[2],可以采用该波长为基点考察一级倍频内不同谱宽下的

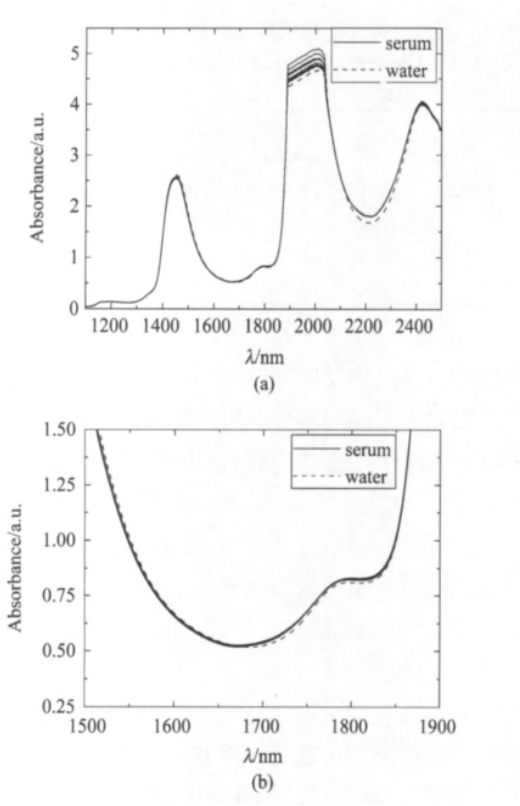


图1 2 mm厚度水与血清的光谱(a)全谱光谱(b)一级倍频段光谱

Fig. 1 Spectra of 2 mm thick water and serum (a) full spectra (b) spectra in first overtone region

建模预测效果.

用于评价模型效果的指标有:交叉验证均方根误差(RMSECV)、预测均方根误差(RMSEP)和相对预测误差平均值(MPEP).式(1)~(3)是RMSECV、RMSEP及MPEP计算公式.

$$\mathrm{RMSECV} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_{ic})^2}{n-1}}, \quad (1)$$

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{m}(y_i - y_{ip})^2}{m-1}}, \quad (2)$$

$$\mathrm{MPEP} = \frac{\sum_{i=1}^{m}(|y_i - y_{ip}|/y_i)}{m} \times 100\%, \quad (3)$$

式中$y_i$表示第$i$号样品经参考方法测得的浓度,$y_{ic}$与$y_{ip}$是第$i$号样品在定标集或预测集中的计算值,$i$为样品序号,$n$是定标集的总样品数,$m$是预测集的总样品数.

考察不同谱宽模型效果的过程为:以1740 nm为中心,向两侧逐步扩展光谱范围.为保证足够的光谱信息,最窄的建模谱宽选取50 nm,为验证对水吸收峰影响的推断,波长范围最大扩展至1610~1870 nm,谱宽260 nm.对每种建模谱宽都使用偏最小二乘算法(PLS)建模,模型因子数在1~10之间变化,预处理方法为中心化.图2是不同建模谱宽对应的最小RMSECV.

由图2可见,胆固醇与甘油三酯模型的RMSECV并未随建模谱宽增大而单调增加或减小.建模谱宽超过240 nm时,由于包含了水在1940 nm附近的吸收峰,模型效果迅速恶化.胆固醇模型的RMSECV在建模谱宽70~100 nm之内较小;甘油三酯模型的RMSECV在建模谱宽约100 nm和200 nm时较小.根据图2,为胆固醇与甘油三酯的建模谱宽选取几个代表点,胆固醇模型选取70 nm、100 nm和260 nm,甘油三酯模型选取100 nm、200 nm和260 nm.为研究仅使用一级倍频内光谱建立的胆固醇与甘油三酯预测模型效果,使用去除水峰的全谱建立模型,三种代表性建模谱宽与全谱模型的效果对比见表2.

由表2可见,使用全谱预测的效果比其它模型差.其余三种建模谱宽对应的胆固醇与甘油三酯模型中,260 nm模型最差,剩余两种模型的效果都在误差范围之内.使用70 nm胆固醇模型和100 nm甘油三酯模型足以代表该区域内模型效果.表3是所建模型与文献[7,9]中模型效果比较.由于实验条件不同,只选取文献中最佳效果,比较模型的预测精度.所建立的胆固醇模型与文献[7]的结果是相同的,均优于文献[9]的结果;甘油三酯模型的RMSEP与MPEP比文献[7,9]的数值大,在误差范围之内,效果基本等同.

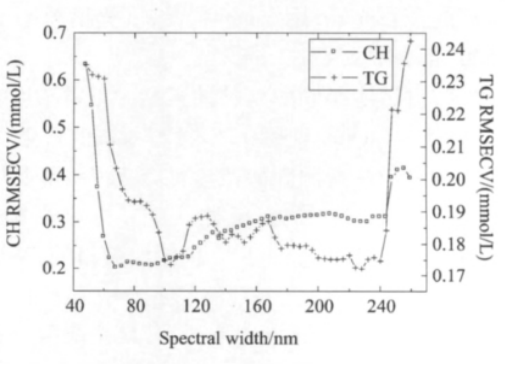


图2 不同建模谱宽下胆固醇(CH)与甘油三酯(TG)模型的RMSECV

Fig. 2 RMSECV of cholesterol (CH) and triglyceride (TG) models in different spectral width

表 2　不同建模谱宽下的模型

**Table 2　Models with different spectral width**

| | 建模谱宽/nm | 光谱范围/nm | $R_c$ | RMSECV/mmol/L | $R_p$ | RMSEP/mmol/L | PLS 因子数 | MPEP/% |
|---|---|---|---|---|---|---|---|---|
| 胆固醇 | 70 | 1700 ~ 1770 | 0.98 | 0.20 | 0.99 | 0.17 | 8 | 3.0 |
| | 100 | 1690 ~ 1790 | 0.98 | 0.21 | 0.99 | 0.18 | 10 | 3.4 |
| | 260 | 1610 ~ 1870 | 0.94 | 0.39 | 0.97 | 0.27 | 10 | 5.0 |
| | 1250 | 1100 ~ 2350 | 0.88 | 0.58 | 0.90 | 0.51 | 10 | 9.6 |
| 甘油三酯 | 100 | 1690 ~ 1790 | 0.97 | 0.17 | 0.98 | 0.14 | 8 | 10.3 |
| | 200 | 1640 ~ 1840 | 0.97 | 0.18 | 0.99 | 0.12 | 10 | 8.5 |
| | 260 | 1610 ~ 1870 | 0.94 | 0.24 | 0.96 | 0.21 | 10 | 14.3 |
| | 1250 | 1100 ~ 2350 | 0.90 | 0.32 | 0.87 | 0.35 | 10 | 23.1 |

表 3　与其他文献模型最优效果比较

**Table 3　Comparing with the best results of other papers**

| | 模型来源 | 光谱范围/nm | $R_p$ | RMSEP/mmol/L | PLS 因子数 | MPEP/% |
|---|---|---|---|---|---|---|
| 胆固醇 | 表 2 | 1700 ~ 1770 | 0.99 | 0.17 | 8 | 3.0 |
| | 文献[7] | 1660 ~ 1820, 2210 ~ 2370 | 0.98 | 0.18 | 21 | / |
| | 文献[9] | 2060 ~ 2350 | / | 0.31 | 13 | 4.9 |
| 甘油三酯 | 表 2 | 1690 ~ 1790 | 0.98 | 0.14 | 8 | 10.3 |
| | 文献[7] | 1660 ~ 1820, 2210 ~ 2370 | 0.99 | 0.13 | 20 | / |
| | 文献[9] | 2060 ~ 2350 | / | 0.11 | 13 | 5.4 |

## 3　结论

利用光栅光谱仪采集 2 mm 厚血清的近红外光谱，选择一级倍频区域建立胆固醇与甘油三酯的预测模型. 经过计算，在一级倍频区域内，使用 70 nm 与 100 nm 宽的光谱建立模型，得到胆固醇的RMSEP和MPEP 分别是 0.17 mmol/L 与 3.0%，甘油三酯的 RMSEP 和 MPEP 分别是 0.14 mmol/L 与 10.3%，优于全谱建模的结果. 根据实验及计算结果，对于 2 mm厚的血清，使用一级倍频内宽度不大于 100 nm 的光谱建模，与已有报道中较优秀模型结果相比，文中建模谱宽小，胆固醇模型效果更好，甘油三酯模型效果相近. 较窄的光谱范围可以减少样品光谱非线性对于模型的影响，便于选用效率更高的窄带光源，具有更高的信噪比，还可避免引入其他干扰信息. 文中对光谱宽度的选择方法，同样适用于采用近红外光谱分析其他样品体系与分析物，同时可以作为专用近红外分析仪设计时选取光谱宽度的依据.

## REFERENCES


[1] YAN Yan-Lu, ZHAO Long-Lian, HAN Dong-Hai, *et al*. *Elements and application of near-infrared spectra analysis* [M]. Beijing: China Light Industry Press(严衍禄，赵龙莲，韩东海，等. 近红外光谱分析基础与应用. 北京: 中国轻工业出版社), 2005.

[2] LU Wan-Zhen. *Modern near-infrared spectroscopy analytical technology* [M]. 2nd ed. Beijing: China Petrochemical Press(陆婉珍. 现代近红外光谱分析技术，第二版. 北京: 中国石化出版社), 2006.

[3] Arnold M A, Small G W. Noninvasive glucose sensing [J]. *Analytical Chemistry*, 2005, **77**(17): 5429-5439.

[4] CHEN Xing-Dan. Possibility of noninvasive clinical biochemical examination by near infrared spectroscopy [J]. *Optics and Precision Engineering*(陈星旦. 近红外光谱无创生化检验的可能性. 光学精密工程), 2008, **16**(5): 759-763.

[5] DING Hai-Quan, LU Qi-Peng, PIAO Ren-Guan, *et al*. Optimum choice of combination wavelengths in near infrared analysis for soil organic matter [J]. *Optics and Precision Engineering*(丁海泉，卢启鹏，朴仁官，等. 土壤有机质近红外光谱分析组合波长的优选. 光学精密工程), 2007, **15**(12): 1946-1951.

[6] GAO Hong-Zhi, LU Qi-Peng, DING Hai-Quan. Choice of characteristic near-infrared wavelengths for soil total nitrogen based on successive projection algorithm [J]. *Spectroscopy and Spectral Analysis*(高洪智，卢启鹏，丁海泉. 基于连续投影算法的土壤总氮近红外特征波长的选取. 光谱学与光谱分析), 2009, **29**(11): 2951-2954.

[7] Heise H M, Marbach R. Multivariate determination of blood substrates in human plasma by FT-NIR spectroscopy [J]. *Proc. SPIE*, 1992, **1575**: 507-508.

[8] Chen J, Arnold M A, Small G W. Comparison of combination and first overtone spectral regions for near-infrared calibration models for glucose and other biomolecules in aqueous solutions [J]. *Analytical Chemistry*, 2004, **76**(18): 5405-5413.

[9] Hazen K H, Arnold M A, Small G W. Measurement of glucose and other analytes in undiluted human serum with near-infrared transmission spectroscopy [J]. *Analytica Chimica Acta*, 1998, **371**(2-3): 255-267.

[10] CHEN Hua-Cai, YANG Zhong-Guo, LI Hui-Ying, *et al*. Feasibility study for rapid determination of cholesterol concentration in human serum by using fourier transform near infrared spectroscopy [J]. *Acta Laser Biology Sinica*(陈华才，杨仲国，李惠英，等. 人血清中胆固醇近红外光谱快速检测初步研究. 激光生物学报), 2004, **13**(6): 429-432.

[11] Peuchant E, Salles C, Jensen R. Determination of serum cholesterol by near-infrared reflectance spectrometry [J]. *Analytical Chemistry*, 1987, **59**(14): 1816-1819.

[12] Haaland D M, Robinson M R, Koepp G W, *et al*. Reagentless near-infrared determination of glucose in whole blood using multivariate calibration [J]. *Applied Spectroscopy*, 1992, **46**(10): 1575-1578.

[13] Hall J W, Pollard A. Near-infrared spectroscopic determination of serum total proteins, albumin, globulins, and urea. clinical biochemistry [J]. *Clinical Biochemistry*, 1993, **26**(6): 483-490.

[14] da Costa Filho P A, Poppi R J. Determination of triglycerides in human plasma using near-infrared spectroscopy and multivariate calibration methods [J]. *Analytica Chimica Acta*, 2001, **446**(1-2): 39-47.

[15] Saptari V, Youcef-Toumi K. Design of a mechanical-tunable filter spectrometer for noninvasive glucose measurement [J]. *Applied Optics*, 2004, **43**(13): 2680-2688.

[16] Olesberg J T, Arnold M A, Mermelstein C, *et al*. Tunable laser diode system for noninvasive blood glucose measurements [J]. *Applied Spectroscopy*, 2005, **59**(12): 1480-1484.