第32卷，第11期
2012年11月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 32，No. 11，pp2992-2996
November，2012

基于舌近红外反射光谱的人体血清多种蛋白含量无创测量

李家星[1,2]，林 凌[1]，李 哲[1]，李 刚[1]，宋 维[3*]

1. 天津大学精密测试技术及仪器国家重点实验室，天津 300072
2. 天津科技大学海洋科学与工程学院，天津 300457
3. 天津师范大学物理与电子信息学院，天津 300387

摘 要 探讨一种基于近红外反射光谱的人体血清白蛋白、球蛋白和总蛋白三种生化指标的无创检测方法。采集58例志愿者舌尖处近红外反射光谱，考虑这些光谱数据与血清蛋白浓度间因个体差异等存在非线性映射关系，在计算归一化光谱反射率及分析样本蛋白含量统计分布上，采用支持向量机分别建立三种蛋白成分近红外光谱定量回归模型，并与传统的偏最小二乘法进行比较。实验结果表明，支持向量机校正模型的预测效果较好且明显优于偏最小二乘法校正模型，对白蛋白、球蛋白和总蛋白的预测相关系数分别达到0.894，0.931和0.863，预测的均方误差为2.19，1.93和4.38。因此，支持向量机可有效抵抗活体检测定量分析中存在的非线性因素，提高模型的鲁棒性。同时也表明舌的近红外光谱信息能够较客观的反映人体理化指标的变化，用于血清蛋白含量的快速无创检测具有较高的可行性。

关键词 近红外反射光谱；血清蛋白；无创测量；支持向量机

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2012)11-2992-05

引 言

人体血清白蛋白(albumin，ALB)、球蛋白(globulin，GLB)及总蛋白(total protein，TP)含量是血液检测中重要的生化指标，其中白蛋白与球蛋白互为缓冲对，维持着血浆渗透压的平衡，两者总和为总蛋白[1]。这些蛋白含量的变化既能反映人体的健康及营养状况，又是肝、肾等疾病临床诊治及判断预后的敏感指标[2,3]，其准确测定具有重要临床意义。目前临床检测手段多通过采血及生化分析，既给患者带来痛苦及感染的风险，又具有较高的测量成本，同时也存在难以实时、连续监测等缺点。相比较之下，有效的蛋白生化指标无创检测方法将带来更高的临床应用价值。

光学检测技术以其高灵敏度、高速、高精度、操作性强等特点，成为时下主流的无创检测手段[4]，广泛应用于血液成分检测[5,6]及疾病诊断等[7,8]生物医学领域。尤其是近红外光谱(near infrared spectroscopy，780～2 526 nm)，其光的吸光度和反射率等参量源于物质内部基团分子间振动的倍频及合频等信息，可反映组织的化学成分及分子结构等信息的微小差异。此外人体组织和血液对近红外波长区的吸收量较小，能够获得相对强的光信号[9]，因此，近红外光谱成为血液成分无创检测中最有前景的波段之一。已有研究采集人体不同部位近红外光谱用于血液成分的无创检测。Wolfgang等[10]提出用一束透过角膜经过前房房水反射回来的近红外光来检测房水中葡萄糖的含量。Li等[11]在指腹处采集近红外反射光谱实现人体血糖的无创检测。Kraitl等[12]从手指的可见-近红外光谱段选择5个波长用于血红蛋白的无创检测。李刚等[13]采集舌部近红外和可见光反射光谱用于红细胞数的定量分析。

然而皮肤及脂肪等组织对光的探测有不同程度衰减以致产生某些不确定因素，因此选择有效的采集部位及定量分析方法是十分重要的。舌体上布满血管，血液流变参数的变化通常能反映到舌象，且舌部没有脂肪组织，比起指尖以及手臂等部位对光信号的衰减相对弱，可更准确的反映人体内部微循环信息。Burmeister等[14]在人体血糖近红外光谱无创检测中指出舌部采集的光信号具有较高信噪比，是较佳的无创测量点。Janelle等[15]选用舌下静脉采集可见光谱并进行血红蛋白无创定量分析，取得较好的效果。但是不同人的舌组

收稿日期：2011-12-02，**修订日期**：2012-02-25

基金项目：国家自然科学基金项目(30973964)，天津市应用基础及前沿技术研究计划项目(11JCZDJC17100)，天津市科技计划项目：科技型中小企业创新基金项目(10ZXCXSY10400)资助

作者简介：李家星，女，1979年生，天津大学精密仪器与光电子工程学院在读博士 e-mail：lillyjiaxing@126.com

*通讯联系人 e-mail：thinksw@163.com

织的异质性造成光的散射及吸收等参数存在差异，且个体间存在的不同病机及环境因素也加剧了随机干扰，这都使测量光谱与人体血液成分浓度间存在复杂的非线性映射。已有报道指出支持向量机(support vector machine，SVM)在存有某些不确定性非线性因素的光谱定量分析中表现出良好的泛化能力。Mello 等[16]在未经任何化学预处理下采用最小二乘支持向量机(least square support vector machine，LS_SVM)建立血清近红外光谱与甲状腺激素间的非线性定量模型，完美预测效果表明 LS_SVM 模型具有优良的鲁棒性。Barman 等[17]在建立人体血糖与拉曼光谱校正模型中，指出 SVM 较偏最小二乘(partial least squares，PLS)更能胜任因个体差异导致非线性定量分析所面临的挑战。

综上，在探讨人体血清 ALB，GLB 和 TP 三种蛋白指标的无创检测方法中，本工作采集 58 例志愿者舌尖部位的近红外反射光谱，在进行一系列预处理后，采用 SVM 和 PLS 两种方法分别建立血清三种蛋白成分的定量预测模型。实验结果显示这种基于近红外光谱的血清蛋白含量检测方法具有一定可行性，且 SVM 更适合建立这种非线性映射关系。

1　实验部分

1.1　装置与采集

选用 Ocean 公司生产的 NIR512 近红外光谱仪，并用配套的 spectralsuit 软件进行采样，光源和光纤采用定制的 GY-30 光纤耦合溴钨灯及其配套的 Y 型光纤。将光源发出的光通过光纤探头垂直照射于舌尖并接收表面的反射光。

58 例志愿者来自河南省黄县第一人民医院住院部多个科室。年龄在 30～60 岁之间，采集时志愿者先漱口，调整坐姿并自然将舌伸出口外。光源预热 5 min 后，将光纤探头距离舌尖表面 5 mm 处垂直照射，光谱仪积分时间定为 35 ms，共采集 50 次，获得 853.59～1 737.26 nm 间 512 个波段的光强信息，鉴于光谱仪特性上的限制，去掉零点(853.59 nm 处)，最终有效波长数据数为 511。

Fig. 1　Experimental device

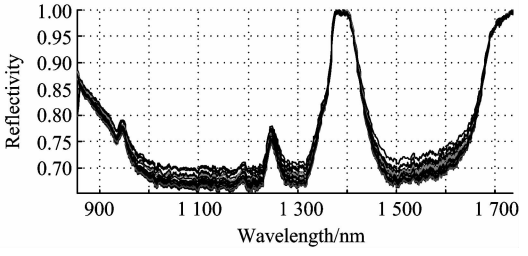

Fig. 2　NIR normalized reflectivity of 58 samples

1.2　光谱数据预处理

物体的光谱反射率属于其自身的物理特性，不随外界照明条件的改变而改变。因此计算舌尖的光谱反射率能够克服光源参数及背景噪声等因素所产生的干扰。为确保反射率计算准确，通过多次测量光源光谱并求均值的方式对光源定标。为进一步消除测量误差以获取到更精准的信息，对所计算的光谱反射率进行归一化处理，58 组光谱反射率归一化曲线如图 2 所示。

2　结果与讨论

2.1　校正集与预测集的分配

校正集与预测集中样本浓度值的分布对定量分析模型的鲁棒性影响较大。为确保后期所建立模型具有良好的泛化能力，所选择的校正集与预测集样本浓度取值范围及相对比例应尽量保持一致。58 例受试者的血清中白蛋白(ALB)、球蛋白(GLB)、总蛋白(TP)生化检测含量统计分析结果如表 1 所示，可见采集样本的 3 项生化指标较生物参考区间缺少一定的高值分布，这必定影响模型在高值区的预测能力。本文分别对各个蛋白的含量范围进行合理划分，从每个含量子区间按约为 4∶1 的比例随机抽取，分配为校正集和预测集。

Table 1　Statistic data of 3 kinds of protein contents for 58 samples

生化项目	含量范围	均值	标准差	参考区间
ALB($g\cdot L^{-1}$)	20.8～47.1	36.7	5.5	35～55
GLB($g\cdot L^{-1}$)	15.1～37.1	26.6	4.7	20～45
TP($g\cdot L^{-1}$)	35.9～79.5	63.3	8.1	60～80

从表 2 校正集与预测集各蛋白生化含量的统计结果中可见两者分布情况大体相同，且与原始集统计信息接近，应能较全面地反映原始样本空间信息。由于样本数量较少，且浓度值分布不均匀，因此所分配的三种蛋白成分校正集与预测集的分布存在不一致性。

Table 2　Statistic data of 3 kinds of protein contents for calibration samples and prediction samples

生化项目	校正集均值	校正集标准差	预测集均值	预测集标准差	校正/预测集样本比例
ALB($g\cdot L^{-1}$)	36.4	5.7	3..0	5.0	44/14
GLB($g\cdot L^{-1}$)	27.0	4.6	26.1	4.5	42/16
TP($g\cdot L^{-1}$)	63.7	8.2	63.6	8.7	43/15

2.2　建模与预测结果分析

(1)SVM 回归建模

面对高维的光谱数据建模时，当样本数远小于光谱维数时常导致 Hughes 现象[18]，致使建模效果变差。SVM 是专门针对小样本的学习算法，对光谱数据的高维特性并不敏感，一定程度上可克服 Hughes 现象。这个观点也在实验中得到验证：采用主成分分析(PCA)对光谱数据进行降维，对 SVM

的建模效果并无实质性改善。

当进行 SVM 运算时，需选择合适的核函数及确定最佳的模型参数。在众多核函数中，RBF 核函数作为非线性函数能够减少训练过程中计算的复杂性。搜索最优的惩罚因子 c 和选择 RBF 核函数参数 g 是确保模型具有最佳的泛化能力的关键。实验中采用网格搜索法(grid searching)和留一剔除交叉验证(leave one out cross validation)相结合的方式，逐步聚焦到最小均方误差条件下的最佳参数点。c 和 g 的值都在 $(2^{-10}, 2^{10})$区域内进行搜索。如图 3 所示，ALB，GLB 和 TP 三种蛋白成分最佳 SVM 回归模型参数分别为(16，1.414 2)，(11.3 137，2)及(32，2)。从值的分布来看所选 3 组参数相对适中，其中 g 的值决定回归量的大小，过大会导致过拟合现象，而过小则影响预测精度。而 c 过大则导致模型结构风险加大。

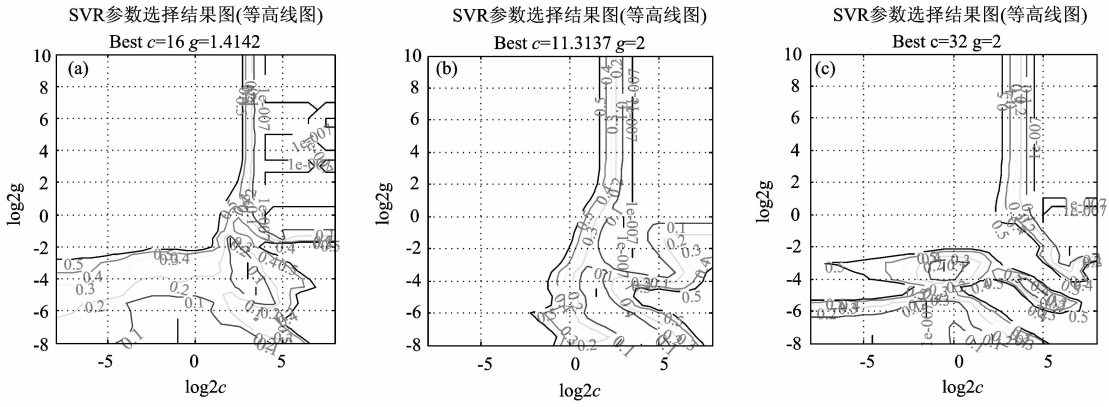

Fig. 3 SVM model parameter distributions of ALB, GLB and TP by grid searching

(a): ALB; (b): GLB; (c): TP

(2)PLS 回归建模

PLS 是目前近红外光谱化学计量分析中最有效的方法之一，可有效解决近红外光谱间的多重共线性问题，明显改善数据结果的可靠性和准确度[19]。PLS 在模型建立前，先针对预测变量对高维光谱进行正交变换提取主因子，以达到降维目的。为防止产生过拟合，本研究依据留一剔除交叉验证计算及交叉验证均方残差和(root mean square errors of cross-validation，RMSECV)最小的规则确定最佳主因子数。ALB，GLB 和 TP 三种蛋白成分 PLS 最佳主因子数分别为 14，19 和 15，如图 4 所示。

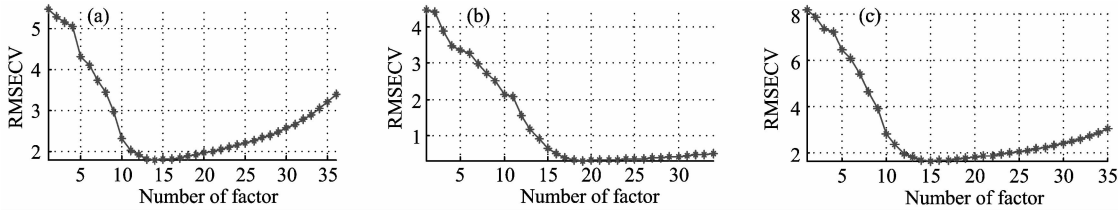

Fig. 4 Influence of main of factors on RMSECV for determination of ALB, ALB and TP

(a): ALB; (b): GLB; (c): TP

Table 3 Evaluation indicators of all models

生化项目	SVR 模型预测评价				PLS 回归模型预测评价			
	R_c	R_p	RMSEP/(g·L^{-1})	最大相对误差/%	R_c	R_p	RMSEP/(g·L^{-1})	最大相对误差/%
ALB	0.954	0.894	2.19	12.3	0.965	0.814	3.49	16.4
GLB	0.995	0.931	1.93	16.8	0.998	0.751	4.13	37.4
TP	0.955	0.863	4.38	18.1	0.986	0.766	6.16	20.1

(3)预测结果分析

应用 SVM 和 PLS 方法建立各蛋白成分的回归模型分别对 ALB 预测集的 14 例样本、GLB 预测集的 16 例样本及 TP 预测集的 15 例样本进行预测，表 3 显示对各模型的评价指标，预测效果分别如图 5 和图 6 所示。从各组模型预测相关系数(R_p)上看，利用人体舌尖的近红外光谱计算的血清蛋白组分预测值与其测定标准值之间存在较为显著的相关关系(R_p>0.75)，可见这种无创检测方法具有一定可行性。虽然三种蛋白成分的 PLS 模型对校正集预测的相关系数(R_c)都达到了 0.96 以上，但对个别非建模样本的预测存在较大偏差(如图 6)，且从计算的预测相关系数(R_p)和预测标准误差(RMSEP)上看，SVM 模型的预测效果明显优于 PLS 模型，存在较好的泛化能力，能够在一定程度上克服个体差异等引起的非线性因素。然而 SVM 模型对个别样本的蛋白含量预测相对偏差仍较大(见图 5)，而各种蛋白存在较大预测偏差的样本大都向校正集的均值偏移，因此说明建模样本分布不

均匀性对模型泛化能力存在较大影响。因此，进一步补充训练样本增大校正集的动态范围及改善其分布均匀性可有效提高预测精度。

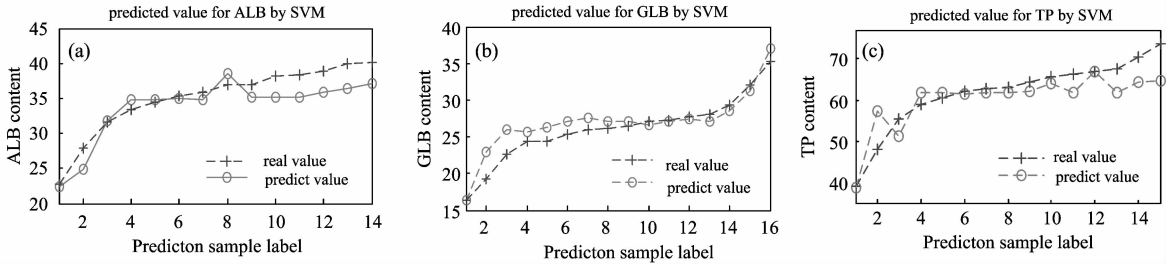

Fig. 5　Predicted results by SVM regression model of ALB, GLB and TP

(a): ALB; (b): GLB; (c): TP

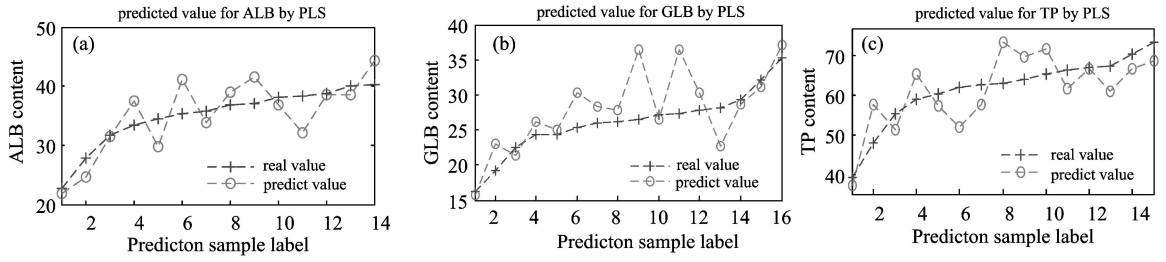

Fig. 6　Predicted results by PLS regression model of ALB, GLB and TP

(a): ALB; (b): GLB; (c): TP

3　结　论

探讨了一种基于近红外光谱的人体血清 ALB、GLB 和 TP 三种蛋白指标的无创检测方法。选用舌尖作为光谱信息的采集部位，并通过光源定标、计算反射率及归一化等方式，提高信噪比。运用 SVM 和 PLS 分别建立了近红外光谱反射率与三种蛋白成分的定量回归模型，实验结果表明，SVM 可改善个体差异等带来的非线性因素，显著提高了模型的鲁棒性。同时也表明舌的光谱信息能够较客观的反映人体生理生化指标的变化，而进一步提高光谱灵敏度、拓展波段范围及增强模型的非线性处理能力，将有望为人体血液蛋白成分的无创检测提供一种新途径。

References

[1] Ohwada H, Nakayama T. British Journal of Nutrition, 2008, 100(6): 1291.
[2] Kim J M, Lim Y M P. Journal of Clinical Microbiology, 2005, 43(5): 2452.
[3] WEN Jie, ZHU De-zeng(文　洁, 朱德增). Journal of Lanzhou University · Medical Sciences(兰州大学学报 · 医学版), 2011, 37(1): 70.
[4] Kuang Y, Walt D R. Biotechnology and Bioengineering, 2007, 96(2): 318.
[5] Tighe P J, Elliott C E, Lucas S D, et al. Acta Anaesthesiologica Scandinavica, 2011, 55(10): 1239.
[6] Li L N, Li Q B, Zhang G J. Journal of Infrared Millimeter and Terahertz Waves, 2009, 30(11): 1191.
[7] Jayanthi J L, Subhash N, Stephen M, et al. Journal of Biophotonics, 2011, 4(10): 696.
[8] Bergholt M S, Zheng W, Lin K, et al. International Journal of Cancer, 2011, 128(11): 2673.
[9] Berger A J, Itzkan I, Feld M S, et al. Spectrochimica Acta Part A, Molecular and Biomolecular Spectroscopy, 1997, 53A(2): 287.
[10] Wolfgang S, Petra M, Jurgen P, et al. Journal of Molecular Structure, 2005, 735(SI): 299.
[11] Li Q B, Li L N, Zhang G J. Infrared Physics & Technology, 2010, 3(5): 410.
[12] Kraitl J, Ewald H, Gehring H. Journal of Optics A-Pure and Applied Optics, 2005, 7(6): S318.
[13] LI Gang, ZHAO Jing, LI Jia-xing, et al(李　刚, 赵　静, 李家星, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2011, 31(5): 1328.
[14] Burmeister J J, Arnold M A, Small G W. Diabetes Technology & Therapeutics, 2000, 2(1): 5.
[15] Bender J E, Shang A B, Moretti E W, et al. Optics Express, 2009, 17(26): 23396.
[16] Mello C, Marangoni A, Poppi R. Analytica Chimica Acta, 2011, 696(1—2): 47.
[17] Barman I, Kong C R, Dingari N C, et al. Analytical Chemistry, 2010, 82(23): 9719.

[18] Melgani F, Bruzzone L. IEEE Transactions on Geoscience and Remote Sensing, 2004, 42(8): 1778.
[19] CHU Xiao-li, XU Yu-peng, LU Wan-zhen(褚小立, 许育鹏, 陆婉珍). Chinese Journal of Analytical Chemistry(分析化学), 2008, 36(5): 702.

Noninvasive Measurement of Human Serum Protein Concentration by Near-Infrared Reflection Spectra for Tongue Inspection

LI Jia-xing[1, 2], LIN Ling[1], LI Zhe[1], LI Gang[1], SONG Wei[3]*
1. State Key Laboratory of Precision Measurement Technology and Instruments, Tianjin University, Tianjin 300072, China
2. College of Marine Science and Engineering, Tianjin University of Science & Technology, Tianjin 300457, China
3. College of Physics & Electronic Information, Tianjin Normal University, Tianjin 300387, China

Abstract In the present paper, a kind of noninvasive determination for human serum protein concentration of albumin, globulin and total protein was explored based on the technology of near-infrared reflectance spectra. Reflectance spectra on the tongue tip of 58 volunteers were collected. Because these is an nonlinear mapping relationship induced by the individual differences between these spectra data and serum protein concentration, SVM was used to establish quantitative regression models of 3 kinds of protein concentration respectively after the normalized spectral reflectance was calculated and the protein content statistics distribution of the sample set was analyzed. In addition, results of SVM were compared with that of PLS. The results show that the predictive effect for calibrated model of SVM is obviously better than that of PLS. Using SVM model to predict the prediction set, the correlation coefficients of ALB, GLB and TP are respectively 0.894, 0.931 and 0.863, and root mean square errors are 2.19, 1.93 and 4.38. So SVM can resist the impact of nonlinear factors among *in-vivo* determinations, and enhance the robustness of the models. Meanwhile it was also showed that the near infrared spectral information for tongue can relatively objectively reflect changes in human physiological and biochemical indexes, and this technology for noninvasive determination of serum protein concentration is highly feasible.

Keywords Near infrared reflectance spectroscopy; Serum protein; Noninvasive measurement; Support vector machine

(Received Dec. 2, 2011; accepted Feb. 25, 2012)

* Corresponding author