

血凝素蛋白及抗体相互作用的太赫兹光谱主成分分析

孙怡雯 钟俊兰 左剑 张存林 但果

Principal component analysis of terahertz spectrum on hemagglutinin protein and its antibody

Sun Yi-Wen Zhong Jun-Lan Zuo Jian Zhang Cun-Lin Dan Guo

引用信息 Citation: *Acta Physica Sinica*, 64, 168701 (2015) DOI: 10.7498/aps.64.168701

在线阅读 View online: <http://dx.doi.org/10.7498/aps.64.168701>

当期内容 View table of contents: <http://wulixb.iphy.ac.cn/CN/Y2015/V64/I16>

您可能感兴趣的其他文章

Articles you may be interested in

相位补偿算法对提高太赫兹雷达距离像分辨率的研究

Improvement in the range resolution of THz radar using phase compensation algorithm

物理学报.2014, 63(14): 148701 <http://dx.doi.org/10.7498/aps.63.148701>

太赫兹双空芯光纤定向耦合器

Terahertz dual air core fiber directional coupler

物理学报.2013, 62(2): 028702 <http://dx.doi.org/10.7498/aps.62.028702>

太赫兹双芯光子带隙光纤定向耦合器

Terahertz dual-core photonic band-gap fiber directional coupler

物理学报.2012, 61(10): 108701 <http://dx.doi.org/10.7498/aps.61.108701>

基于太赫兹量子级联激光器的无线信号传输的实现

Experimental realization of wireless transmission based on terahertz quantumcascade laser

物理学报.2012, 61(9): 098701 <http://dx.doi.org/10.7498/aps.61.098701>

低损宽频高双折射太赫兹光子带隙光纤

A broadband low loss and high birefringence terahertz photonic bandgap photonic crystal fiber

物理学报.2011, 60(9): 098702 <http://dx.doi.org/10.7498/aps.60.098702>

血凝素蛋白及抗体相互作用的太赫兹光谱主成分分析*

孙怡雯^{1)†} 钟俊兰¹⁾ 左剑²⁾ 张存林²⁾ 但果^{1)‡}

1)(深圳大学医学院生物医学工程系, 医学超声关键技术国家-地方联合工程实验室, 广东省生物医学信息检测与超声成像重点实验室, 深圳 518060)

2)(首都师范大学物理系, 北京 100048)

(2015年1月7日收到; 2015年3月23日收到修改稿)

采用太赫兹时域光谱系统, 测量了7种不同浓度的血凝素蛋白及其与特异性抗体、无关抗体对照组反应的透射光谱, 采用光谱预处理及主成分分析法, 对多个太赫兹光谱参数进行分析. 结果显示, 主成分分析在数据降维的同时, 可以突出数据的主要变化趋势; 在原始变量相关性一致的情况下, 约化吸收截面与血凝素蛋白浓度之间表现出最强的相关性, 而介电损耗角正切值更适用于对血凝素蛋白-抗体复合物的聚类效果进行定性分析. 该研究表明主成分分析法对于太赫兹生物光谱的分析及进一步研究蛋白质的结构和功能具有重要的指导意义.

关键词: 太赫兹光谱, 光谱预处理, 主成分分析, 抗原-抗体

PACS: 87.50.U-, 87.64.K-, 87.80.-y, 87.14.E-

DOI: 10.7498/aps.64.168701

1 引言

太赫兹 (THz) 波是指频率介于 0.1—10 THz (波长为 30—3000 μm) 范围内的电磁波. 在整个电磁波谱中, 该区域是最后一个被探索和开发的频段, 与其两侧成熟的微波和红外技术相比, THz 技术的研究也仅有几十年历史. 尽管如此, 近年来随着对 THz 生物现象的理解及相干检测技术的发展, THz 技术已被证明在研究生物物理特性及功能方面具有明显优势^[1-5]. 研究人员已经证明, 蛋白质的集体振动和转动的频率正好位于 THz 频段, 而蛋白质的集体振动模型对研究蛋白质的结构和功能非常重要, 这使得 THz 技术在生物医学领域的研究更加具有潜力^[6-9], 尤其是在探索例如分子间弱相互作用、蛋白质分子骨架集体振动、偶极子旋转及

低频振动吸收频率等问题上^[10-14].

甲型流感病毒表面主要存在两种蛋白质, 血凝素 (hemagglutinin, HA) 蛋白和神经氨酸酶 (neuraminidase, NA)^[15], 其中 HA 蛋白占病毒粒子表面膜蛋白的 80% 以上, 在病毒导入宿主细胞时参与引发免疫系统反应, 是设计广谱流感疫苗的潜在靶蛋白^[16]. F10 是近年来才陆续有报道的一种广谱中和性抗体, 可以特异性识别 HA 蛋白分子茎杆区的空间表位, 可为广谱流感疫苗的研发提供证据和新思路. 在本实验中, 我们采用 THz-TDS 系统, 测量了不同浓度的甲型流感 H9N2 血凝素蛋白溶液, 及其与抗体 F10 (特异性结合) 和无关抗体 irmAb (阴性对照组) 反应后的 THz 透射光谱, 通过光谱预处理及主成分分析 (principal component analysis, PCA) 法, 用以筛选在抗原、抗原-抗体反应体系中最敏感的 THz 光学参数.

* 国家自然科学基金 (批准号: 61205092, 11204190)、广东省自然科学基金 (批准号: S2012040007668) 和广东省教育厅高校优秀青年创新人才培养计划 (批准号: 2012LYM_0116) 资助的课题.

† 通信作者. E-mail: ywsun@szu.edu.cn

‡ 通信作者. E-mail: danguo@szu.edu.cn

2 方法及实验

2.1 实验装置

在实验中,我们利用 Z3TM 太赫兹时域光谱系统 (Zomega, THz Corp., Troy, USA), 获得样品的 THz 透射光谱. 系统主要由超快脉冲激光器、时间延迟控制系统、THz 脉冲发射装置和相应的探测装置组成^[17]. 钛蓝宝石激光器激发出脉宽为 120 fs, 中心波长为 800 nm 的激励波 (重复频率约为 80 MHz, 平均功率约为 1 W). 激发脉冲垂直入射样品表面, 获得时间相关的 THz 电场强度波形, 利用快速傅里叶变换, 得到相应的频谱信息. 由于液态样品对 THz 波的吸收, 探测有效频率为 0.1—1.4 THz. 频域 THz 光谱经过数据预处理和典型计算, 即可获得被测样品吸收系数 $a(\omega)$ 、折射率 $n(\omega)$ 等光学参数. 光谱处理程序采用 MATLAB 7 自行编写.

2.2 样品制备

实验中,将重组甲型流感 H9N2 (A/Hong Kong/1073/1999) 血凝素蛋白 (Sino Biological Inc., Cat no.11229-V08H) 分别用 pH 值为 7.4 的磷酸盐缓冲液 (PBS) 配制成浓度梯度为 7.5, 15, 29, 56, 113, 225 和 430 $\mu\text{g}/\text{ml}$ 的样品溶液待用. 230 $\mu\text{g}/\text{ml}$ 全人单克隆抗体 F10 (~160 kD, NBGen) 与 HA 蛋白以 1 : 1 的体积比混合. 全人单克隆抗体 anti-CXCR4 (~160 kD, NBGen) 作为阴性对照组 (irmAb) 的反应抗体, 按照与 F10 相同浓度和体积比的数值关系与 HA 蛋白质溶液混合. 样品蛋白质的浓度由 NanoDrop 分光光度计 (Thermo Scientific) 测定. 所有抗原-抗体反应均在室温孵育 30 min^[18], 使其均匀混合并充分特异性结合.

3 实验结果与讨论

3.1 光谱预处理及主成分分析

特征吸收峰是传统红外光谱分析技术中重要的参数之一,但由于生物大分子在 THz 频段的总响应谱带范围较宽,吸收带重叠,因此对于没有明显吸收峰的 (液相) 样品,用传统归属峰位置及强度分

析的方法,显然比较困难. PCA 则是解决这一问题的有效方法. 由于 THz 频段下不同样品的吸收特性存在明显的差别,因此我们首先以 THz 吸收谱为例,阐述 PCA 及光谱预处理对数据结果的影响.

PCA 是一种有效的数据压缩和特征提取方法^[19],最早由卡尔·皮尔逊于 1901 年创建,随后在人口统计、数量地理学、分子动力学模拟等学科中应用十分广泛^[20]. PCA 的原理是将原变量进行转换,使数目较少的新变量成为原变量的线性组合,而且新变量能最大程度的表征原变量的数据结构特征. 其目的是使数据降维,以消除众多共存信息中相互重叠的部分^[21].

如图 1(a) 所示,7 个不同浓度的 HA 蛋白溶液,在 0.1—1.4 THz 频率范围内的吸收光谱均无特征吸收峰,吸收系数随频率增加而逐渐增大,且由于蛋白分子周围水合层 (hydration shell) 的影响^[22],吸收与浓度间并未依据 Beer-Lambert 定律呈现良好的线性关系. 对原始吸收光谱全波段做主成分分析,得到主成分得分图 (图 1(b)). 由图 1 可见,第一、二主成分 (PC1 和 PC2) 的累计贡献率已达 99.57%. 因此,只需两个主成分就能很好地概括这组数据信息. 此外,由于原始光谱数据中光谱噪声、干扰组分背景光谱及样品本身复杂性等因素影响,图 1(b) 并没有明显表现出 PC1, PC2 与 HA 蛋白浓度之间的关系.

光谱噪声的主要来源有随机噪声、基线漂移、信号本底、光散射及环境背景等. 在未经处理的 THz 透射光谱中,除了样品自身的信息外,还包含了无关噪声和各种外界干扰. 为了提高分析的准确度,我们尝试采用中心和标准化、求导、平滑、中值滤波和多元散射校正 (MSC) 5 种光谱预处理方法,以获得最优的主成分得分累积贡献率,并以具有最好的聚类效果为标准. 经过多次计算及反复匹配,最终确定以 MSC+平滑+S-G 二阶导数+中值滤波作为最优的光谱预处理方式,用以消除多余信息,增加样品之间的差异,使 PCA 结果更清晰明了.

图 1(c) 和 (d) 为预处理后的吸收光谱及主成分分析得分图. 由图 1(d) 可见,主成分的累积贡献率达到 96.42% 以上,第一主成分得分虽然略有下降,但样品浓度与 PC1 之间表现出良好的相关性,即随着 HA 蛋白浓度的增大,PC1 的得分随之减小,

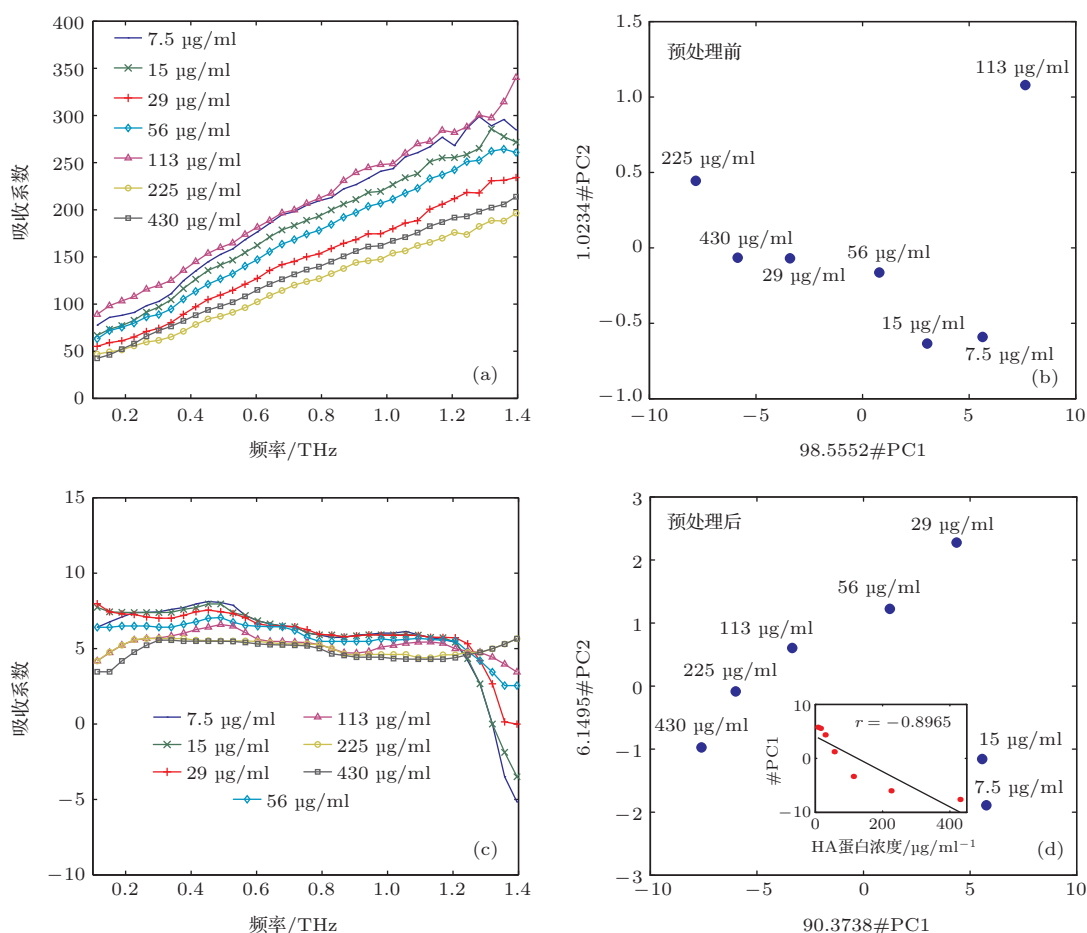


图1 (网刊彩色) 不同浓度HA蛋白THz吸收光谱和主成分分析得分图 (a) 原始光谱; (b) 原始光谱PCA得分图; (c) 预处理后的光谱; (d) 预处理后光谱PCA得分图
 Fig. 1. (color online) THz absorption spectra of HA proteins and the principal component scores plots (a) original spectra; (b) principal component scores plot of the original spectrum; (c) preprocessed data; (d) principal component scores plot of the preprocessed data.

经计算, 相关系数 $r = -0.8965$ (图1(d)中插图). 由此可见, 经过光谱预处理的主成分分析可在数据降维的同时, 更加突出数据的主要变化趋势.

3.2 THz光学参数及相关性分析

蛋白质是具有偶极子特性的典型介电材料, 蛋白质分子结构变化所引起的极化特性, 可由在THz场中的复介电常数来描述. 在复介电常数 $\epsilon^*(\omega) = \epsilon'(\omega) - i\epsilon''(\omega)$ 中, 实部 $\epsilon'(\omega)$ 由材料内部的各种位移极化引起, 代表材料的储能项; 虚部 $\epsilon''(\omega)$ 由电场弛豫极化所致, 代表着材料的损耗项[18].

介电损耗角 δ 或介电损耗角正切值 $\tan \delta$ 通常被定义为介电常数的虚部和实部的比值[22]. 介电损耗角正切代表电磁能量转化为其他能量(如热能)时在其内部引起的能量损耗, 因此对蛋白质分

子侧链动态变化异常敏感.

约化吸收截面(reduced absorption cross-section, RACS)首次由Kawaguchi[23]提出, 并采用THz-TDS系统对氨基酸及相应多肽的结构进行分析. 其研究表明, 在THz频段利用RACS参数, 可以明显区分生物分子单体和多聚物.

我们分别对不同浓度HA蛋白样品的约化吸收截面、介电损耗角正切值和折射率进行分析, 光谱预处理后的主成分得分如图2所示, 三个参数的主成分累积贡献率分别为93.11%, 92.81%和89.61%. 我们对相关系数的取值范围进行分类: $|r| \geq 0.8$ 为极强相关, $|r| \leq 0.2$ 为弱(无)相关, $0.6 < |r| < 0.8$ 强相关, $0.4 < |r| < 0.6$ 中等程度相关, $0.2 < |r| < 0.4$ 弱相关, $0.3 < |r| < 0.5$ 为中等程度相关. 由得分图的相关性分析可知, HA

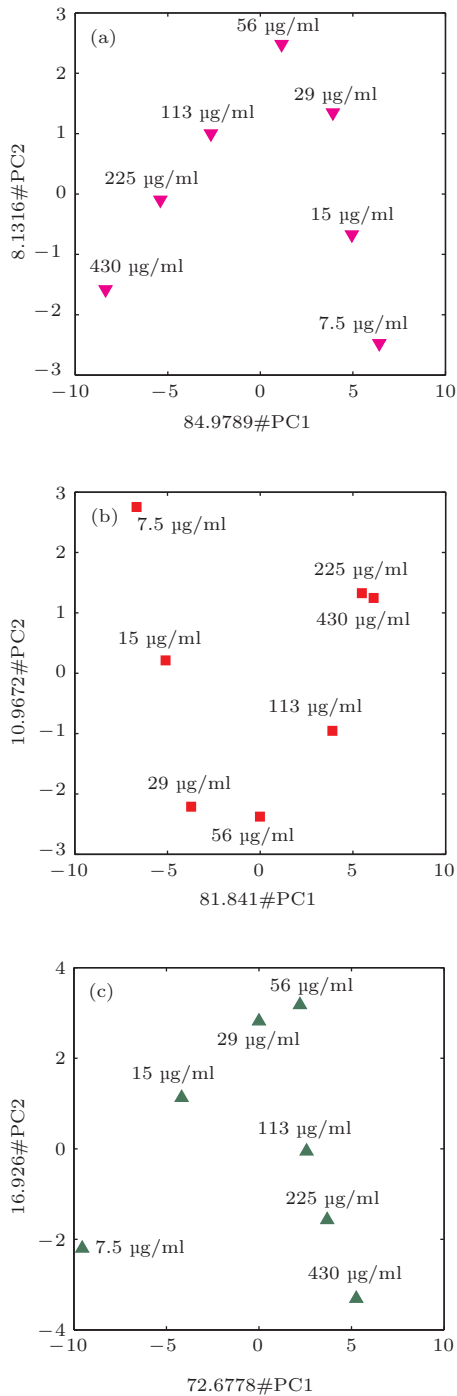


图2 (网刊彩色) HA 蛋白溶液各参数经平滑、二阶导数处理和中值滤波后的光谱曲线及其 PCA 得分图 (a) 约化吸收截面; (b) 介电损耗角正切值 $\tan \delta$; (c) 折射率
 Fig. 2. (color online) Spectra and principal component scores plots of different parameters for HA proteins after Smooth + Second derivative + Median filter pre-processing: (a) RACS; (b) dielectric loss angle tangent $\tan \delta$; (c) refractive index.

蛋白溶液浓度分别与约化吸收截面、介电损耗角正切值和吸收系数的 PC1 呈现极强相关; 与吸收系数、约化吸收截面为负相关, 相关系数分别为

$r = -0.8965$ 和 $r = -0.9265$ (见表 1); 与介电损耗角正切值为正相关, 相关系数为 $r = 0.8328$; HA 蛋白浓度与折射率第一、二主成分都呈现正强相关, 而与吸收系数、约化吸收截面和介电损耗角正切值的第二主成分则呈现弱相关. 由分析可知, 在原始变量相关性一致的情况下, 约化吸收截面的第一主成分与 HA 蛋白浓度之间表现出最强的相关性特征, 预示着该参数可作为 THz 谱分析中研究单一蛋白质体系浓度的敏感参数.

表 1 HA 蛋白溶液浓度与 THz 光学参数各主成分之间的相关系数

Table 1. Correlation coefficient of HA's concentration in different parameter with PC1 and PC2.

	n	a	RACS	$\tan \delta$
PC1	0.7014	-0.8965	-0.9265	0.8328
PC2	-0.6561	-0.1957	0.2738	0.3093

3.3 抗原-抗体反应的 THz 谱主成分分析

在抗原-抗体反应实验中, 我们将上述 7 种浓度的 HA 蛋白溶液分别加入单克隆抗体 F10 和无相关抗体 irmAb. HA-F10 复合物和阴性对照组 HA/irmAb 的折射率、吸收系数、约化吸收截面及介电损耗角正切值的主成分得分图如图 3 所示. 由计算可知, PC1 和 PC2 的累计贡献率均可达 98% 以上.

主成分得分图的优点是可以将样品的信息进行聚类, 但从图 3 中可见, 直接用原始数据并不能将 HA 蛋白、HA-F10 及阴性对照组清晰地区分开来, 因此我们仍采用光谱预处理方法, 对主成分分析结果进行重新计算.

图 4 为经预处理后的主成分得分图. 从图 4(a)–(c) 中我们可以看出, HA 蛋白三个样本组根据成分的差异被不同程度的区分开, 但由于抗原-抗体反应体系的复杂性, 区分效果有改善但并不明显. 但介电损耗角正切值的聚类信息最大程度存在重叠, PC1 的累计贡献率也远高于其他参数. 因此, 我们认为介电损耗角正切值数据最适合对实验中 HA-F10 反应的聚类效果进行定性分析. 由此可得, 介电损耗角由于代表了介观系统的能量损耗, 对于发生在蛋白质分子侧链的抗原-抗体特异性结合反应的敏感性优于其他参数.

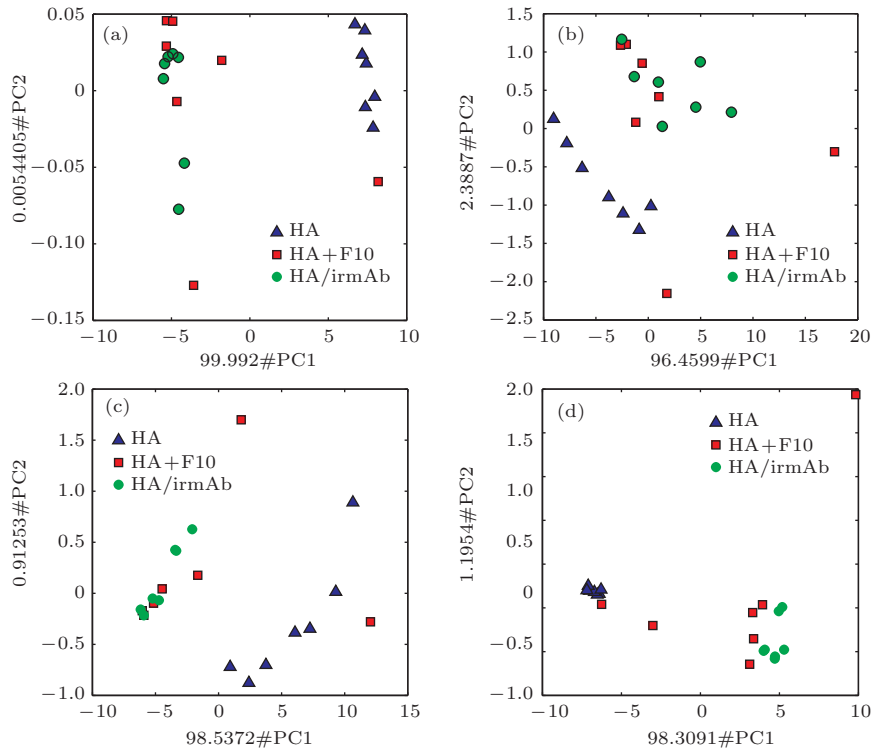


图3 (网刊彩色) HA, HA-F10 和 HA/irmab 各参数主成分得分图 (a) 折射率; (b) 吸收系数; (c) 约化吸收截面; (d) 介电损耗角正切值

Fig. 3. (color online) Principal component scores plots of HA, HA-F10 and HA/irmAb base on different parameter: (a) refractive index; (b) absorption; (c) RACS; (d) tangent of the dielectric loss angle.

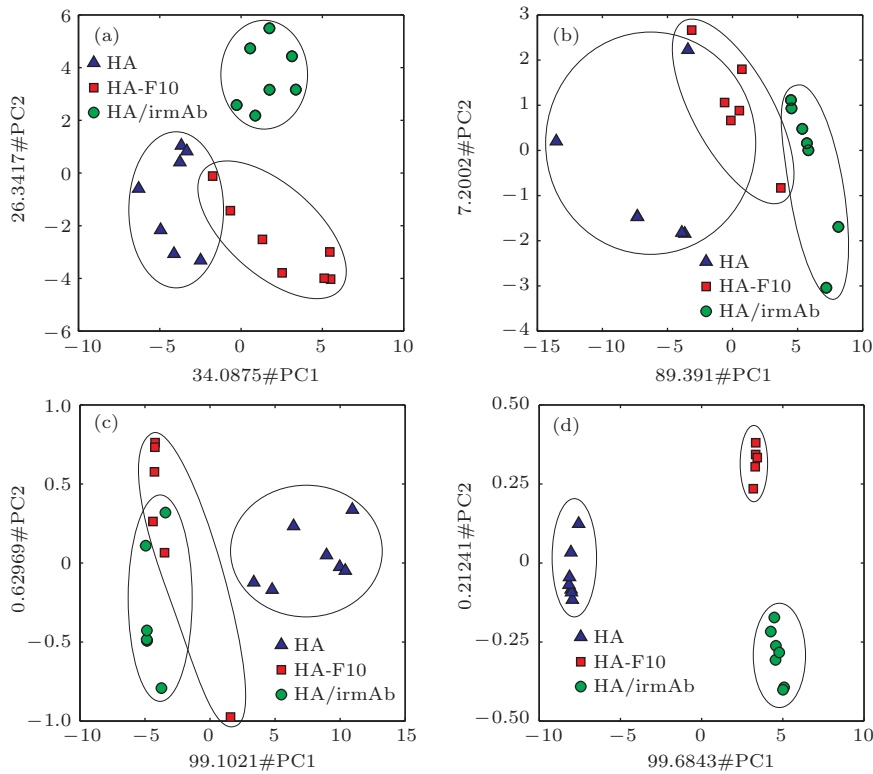


图4 (网刊彩色) HA, HA-F10 和 HA/irmab 光谱预处理后各参数主成分得分图 (a) 折射率; (b) 吸收系数; (c) 约化吸收截面; (d) 介电损耗角正切值

Fig. 4. (color online) Principal component scores plots of HA, HA-F10 and HA/irmAb in different physical parameter after preprocessing: (a) refractive index; (b) absorption; (c) RACS; (d) dielectric loss angle tangent.

4 结 论

本文主要针对甲型流感病毒表面HA蛋白及其与抗体反应的THz透射光谱,分析了不同样本间THz光学参数在主成分分析中的特点.实验发现,根据反应体系的不同,不同的THz光学参数会产生不同的分类效果.因此,合理地选择样品物理参数以及进行适当的预处理是解决问题的关键.本文研究同时表明,在原始变量相关性一致的条件下,约化吸收截面与HA蛋白浓度之间表现出最强的相关性特征,预示着该参数可作为THz谱分析中研究单一蛋白质体系浓度的敏感参数;而根据聚类效果的最优比较,介电损耗角正切值参数更适合于HA-F10反应聚类的定性分析.该研究表明主成分分析方法对于THz生物光谱的解读分析及进一步研究蛋白质的结构和功能具有重要的指导意义.

参考文献

- [1] Liu S G, Zhong R B 2009 *JESTC* **38** 481 (in Chinese)
[刘盛纲, 钟任斌 2009 电子科技大学学报 **38** 481]
- [2] Brucherseifer M, Nagel M, Haring B P, Kurz H, Bosserhoff A, Büttner R 2007 *Appl. Phys. Lett.* **77** 4049
- [3] Mickan S P, Menikh A, Liu H, Mannella C A, MacColl R, Abbott D, Munch J, Zhang X C 2002 *Phys. Med. Biol.* **147** 3789
- [4] Fischer B M, Walther M, Jepsen P U 2002 *Phys. Med. Biol.* **47** 3807
- [5] Kindt J T, Shmittenmaer C A 1996 *J. Phys. Chem.* **100** 10373
- [6] Tian L, Zhou Q L, Zhao, Shi Y L, Zhao D M, Zhao S Q, Zhao H, Bao R M, Zhu S M, Miao Q, Zhang C L 2011 *Chin. Phys. B* **20** 010703
- [7] Dong C, Bao R M, Zhao K, Xu C H, Jin W J Zhong S X 2014 *Chin. Phys. B* **23** 127802
- [8] Shi Y L, Zhou Q L, Zhang C L 2009 *Chin. Phys. B* **18** 4515
- [9] Leng W X, Ge L N, Xu S S, Zhan H L, Zhao K 2014 *Chin. Phys. B* **23** 107804
- [10] Zhao H W, Ge M, Wang W F 2005 *Chemistry* **68** 87 (in Chinese) [赵红卫, 葛敏, 王文峰 2005 化学通报 **68** 87]
- [11] Markelz A G, Whitmire S, Hillebrecht J 2002 *Phys. Med. Biol.* **47** 3797
- [12] Markel A G, Roitberg A, Heilweil E J 2000 *Chem. Phys. Lett.* **320** 42
- [13] Nagel M, Bolivar P H, Brucherseifer M, Kurz H, Bosserhoff A, Büttner R 2002 *Appl. Phys. Lett.* **80** 154
- [14] Xu J, Plaxco K W, Allen S J 2006 *Protein Sci.* **15** 1175
- [15] Jianhua S, William C H, Sandra P, Wei G, Daniel A 2009 *Nat. Struct. Mol. Biol.* **16** 265
- [16] White J M, Hoffman L R, Arevalo J H, Wilson I A 1997 *In Structural Biology of Viruses* (New York: Oxford University Press) p80
- [17] Pickwell E, Cole B E, Fitzgerald A J, Wallace V P, Pepper M 2004 *Appl. Phys. Lett.* **84** 2190
- [18] Sun Y W, Zhang Y T, Pickwell-MacPherson E 2011 *Biophys. J.* **100** 225
- [19] Smith L I 2002 *A tutorial on Principal Components Analysis* (Vol.51) (USA: Cornell University), p52
- [20] Pearson K 1901 *Philos. Mag.* **2** 559
- [21] Cao B H, Fan M B, Jing S Y 2010 *Spectrosc. Spect. Anal.* **30** 1748 (in Chinese) [曹丙花, 范孟豹, 荆胜羽 2010 光谱学与光谱分析 **30** 1748]
- [22] Sun Y W, Zhu Z, Chen S, Balakrishnan J, Abbott D 2012 *PLoS ONE* **7** e50306
- [23] Kawaguchi S, Kambara O, Shibata M, Kandori H, Tomi-inaga K 2010 *Phys. Chem. Chem. Phys.* **12** 10255

Principal component analysis of terahertz spectrum on hemagglutinin protein and its antibody*

Sun Yi-Wen^{1)†} Zhong Jun-Lan¹⁾ Zuo Jian²⁾ Zhang Cun-Lin²⁾ Dan Guo^{1)‡}

1) (*National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, Department of Biomedical Engineering, School of Medicine, Shenzhen University, Shenzhen 518060, China*)

2) (*Department of Physics, Capital Normal University, Beijing 100048, China*)

(Received 7 January 2015; revised manuscript received 23 March 2015)

Abstract

One has proved that the collective structural vibrational modes of proteins are in the terahertz (THz) frequency range. These frequencies relate to the polypeptide backbone and are thought to be essential for conformational dynamics necessary for protein function. Hemagglutinin (HA) is the main surface glycoprotein of the influenza A virus. The H9N2 subtype influenza A virus is recognized as the most possible pandemic strain as it crosses the species barrier, infects swine and humans. In this paper we use principal component analysis (PCA) to study the 7 different concentrations dependent terahertz spectra of hemagglutinin proteins, and detect the binding interaction of HA with the broadly neutralizing monoclonal antibody F10 in liquid phase. Spectrum pretreatment and band selection play a vital role in the THz spectroscopic analysis due to the fact that the original spectrum contains a large amount of interference information. In order to compress variables and extract useful information, we use a variety of pretreatment methods, such as second derivative, multiplicative scatter correction (MSC), least square polynomial fitting derivation, standard normalization, smoothing, moving window median filtering before PCA analysis. We even consider MSC + smoothing + SG second derivative + median filtering as the optimized pretreatment method finally. THz spectrum parameters including refractive index, absorption coefficient, reduced absorption cross-section and dielectric loss angle tangent are calculated in a frequency range of 0.1–1.4 THz for comparison. The results indicate that the reduced absorption cross-section presents the highest correlation response to the concentration variation of HA protein, and the dielectric loss angle tangent appears to be more appropriate for qualitative analysis of HA-antibody binding interaction. PCA method provides a feasible and effective way to find the sensitive parameters for further analyzing the function of protein and the antigen-antibody interaction using terahertz spectrum, whereas an appropriate pretreatment method is required.

Keywords: Terahertz spectroscopy, spectrum preprocessing, principal component analysis, antigen-antibody

PACS: 87.50.U–, 87.64.K–, 87.80.–y, 87.14.E–

DOI: 10.7498/aps.64.168701

* Project supported by the National Natural Science Foundation of China (Grant Nos. 61205092, 11204190), the Guangdong Natural Science Foundation, China (Grant No. S2012040007668), and the Foundation for Distinguished Young Talents in Higher Education of Guangdong, China (Grant No. 2012LYM_0116).

† Corresponding author. E-mail: ywsun@szu.edu.cn

‡ Corresponding author. E-mail: danguo@szu.edu.cn