基于经验模态分解的两种混合氨基酸太赫兹光谱分析研究

刘 婧[1]，刘海顺[2*]，左 剑[2]，张存林[1,2*]，赵跃进[1]，梁美彦[3]

1. 北京理工大学，北京 100081
2. 首都师范大学，北京 100048
3. 山西大学，山西 太原 030013

摘 要 L-苯丙氨酸和L-酪氨酸在合成神经递质和激素的过程中起到了重要的作用。这两种氨基酸具有极为相似的分子结构，但在生物功能上却具有明显区别。前人的研究表明，这两种氨基酸在低频振动上存在显著差异。近年来，太赫兹(THz)光谱学技术作为研究生物分子低频动力学的有效手段被广泛应用，通过太赫兹光谱对氨基酸进行研究，对进一步了解蛋白质和相关生物活性具有重要意义。多变量校准方法已成功应用于太赫兹多组分光谱数据定量分析研究中。然而，传统校准技术由于仅在光谱和目标之间建立单个模型预测未知样品，其预测性能有时仍不尽人意。因此，具有更好精度的集成建模方法(ensemble modeling method)应运而生。集成建模的基本概念是组合多个单独模型的优势以产生更好的预测结果。由黄锷博士提出的经验模态分解(EMD)的方法，可以将信号自适应地分解为一系列的本征模式函数(IMF)，成功地应用于信号和光谱处理中。基于该方法的信号分析也已在太赫兹波段开始使用。然而，在对物质进行定量分析的过程中，目前还没有报道基于EMD方法的太赫兹光谱偏最小二乘(PLS)回归的相关工作。提出了一种基于PLS的EMD分析，并对不同浓度氨基酸混合物的太赫兹光谱进行了定量研究。具体而言，原始的太赫兹时域信号首先通过EMD手段在不同频段被分解为一系列的IMF和一个残差函数。随后，将前几个IMF相加作为一个整体(此处研究了前两、三、四和五个IMF叠加的结果)，对其进行吸收光谱的重建。最后，建立PLS回归模型，用于进一步的物质定量分析。预测结果表明，与原始吸收光谱及其他分解后重组光谱的PLS结果相比，基于前四个IMF信号之和的吸收谱具有更高的 R(0.9961)和最小的RMSEP(0.019 8)。由此可知，EMD-PLS法可以在太赫兹波段对两种氨基酸混合物进行有效地定量分析，实现较为理想的预测精度。

关键词 太赫兹；经验模态分析；偏最小二乘法回归；氨基酸

中图分类号：O434.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2020)10-3061-05

引 言

氨基酸是构建生物细胞和组织的基本成分。L-苯丙氨酸和L-酪氨酸在合成神经递质和激素的过程中起重要作用，这些神经递质和激素参与了人体的糖和脂肪的代谢过程。这两种氨基酸具有相似的分子结构，不同在于L-酪氨酸多了一个羟基，这却导致了两种氨基酸在功能上具有明显区别。前人的研究表明，这两种氨基酸在低频振动上存在显著差异。近年来，太赫兹(THz)光谱学技术作为研究生物分子低频动力学的有效手段被广泛应用[1-2]，因此通过太赫兹光谱对氨基酸进行研究，对进一步了解蛋白质和相关生物活性具有重要意义。2005年和2010年，Yamamoto等[3-4]利用太赫兹手段对氨基酸及其多肽的低频谱进行了研究。2013年，Yu等[5]在太赫兹波段通过主成分分析(PCA)手段处理与吸收线形函数(ALF)方法，对两种氨基酸混合物进行了识别研究。

多变量校准方法[如偏最小二乘法(partial least squares，PLS)]已成功应用于太赫兹多组分光谱数据定量分析研究中。陈涛等[6]将太赫兹光谱技术与PLS回归手段结合，研究多组分药物混合物的实际浓度与预测浓度之间的一致性。Lu等[7]通过PLS和基于太赫兹吸收光谱的区间偏最小二乘(iPLS)回归对L-谷氨酸和L-谷氨酰胺的二元混合物进行了

收稿日期：2019-08-16，**修订日期**：2019-12-28

基金项目：北京成像理论与技术高精尖创新中心项目(19530012003)，国家重大仪器专项基金项目(2012YQ140005-09-01)，国家自然科学基金青年科学基金项目(11804209)资助

作者简介：刘 婧，女，1987年生，北京理工大学博士研究生 e-mail：newone_kaka@163.com

*通讯联系人 e-mail：cunlin_zhang@cnu.edu.cn；phscdream@163.com

定性和定量的分析研究。

然而，传统校准技术由于仅在光谱和目标之间建立单个模型预测未知样品，其预测性能有时仍不尽人意。因此，具有更好精度的集成建模方法(ensemble modeling method)应运而生[8]。集成建模的基本概念是组合多个单独模型的优势以产生更好的预测结果。

1998 年，Huang[9]提出了经验模态分解(empirical mode decomposition，EMD)的方法。该方法可以将信号自适应地分解为一组本征模式函数(intrinsic mode functions，IMF)，成功地广泛应用于信号和光谱处理中[10-11]。基于 EMD 方法的信号分析也已在太赫兹波段开始使用[12-15]。然而，在对物质进行定量分析的过程中，目前还没有报道基于 EMD 方法的太赫兹光谱 PLS 回归的相关工作。本文提出了一种基于 EMD 的 PLS 方法，用于定量分析研究不同浓度氨基酸混合物的太赫兹吸收光谱。该方法提取了基于前几个 IMF 的吸收光谱，用于建立 PLS 回归模型，比较了其结果与原始吸收谱的 PLS 建模结果。

1 实验部分

1.1 数据处理

EMD 的主要思想是将信号 $f(t)$分解为一系列本征模式函数(IMFs)。每个 IMF 应满足两个基本标准：(1)极值和零交叉数量必须相等或在整个数据集中最多有一个差异；(2)由局部最小值和最大值定义的包络，其平均值应为零[9]。该信号可写为

$$f(t) = \sum_{k=1}^{N} x_k(t) + r_N(t) \quad (1)$$

其中 $x_k(t)$是第 k 个 IMF 分量，$r_N(t)$是残差函数。

信号 $f(t)$的分解过程可归纳如下：

(1)找出 $f(t)$的所有极值(最大值或最小值)；

(2)使用三次样条曲线将所有局部最大值或最小值连接为上限或下限；

(3)计算包络 $m_1(t)$的平均值；

(4)提取新的数据序列 $h_1(t)=f(t)-m_1(t)$；

(5)迭代 $h_1(t)$，直到 $h_1(t)$满足 IMF 的上述两个标准，以此来找到第一个 IMF 分量 $x_1(t)$；

(6)对信号 $r_1(t)=f(t)-x_1(t)$重复上述步骤，并获取其余的 IMF。

当残差函数 $r_N(t)$变为单调函数或常数时，该过程即可停止。由此可见，信号 $f(t)$可以分解为一组 IMF 和残差函数。这里，IMF 由不同的振荡模式组成，并且更高阶的 IMF 对应于较低频率的信息。

PLS 是一种较为成熟的线性回归方法[6-7]。该模型的性能主要通过相关系数(R)，校正均方根误差(RMSEC)和预测均方根误差(RMSEP)这几个参数来评估。当一个模型具有更高 R，更小的 RMSEC 和 RMSEP 时，该模型被认为是较理想的模型。

EMD-PLS 方法的流程示意图如图 1 所示。原始的太赫兹时域信号首先通过 EMD 手段，分解为一系列 IMF 和一个残差函数，然后前几个 IMF 相加作为一个整体，随后对其吸收光谱进行重建。最后，建立 PLS 模型用于进一步的物质定量分析。

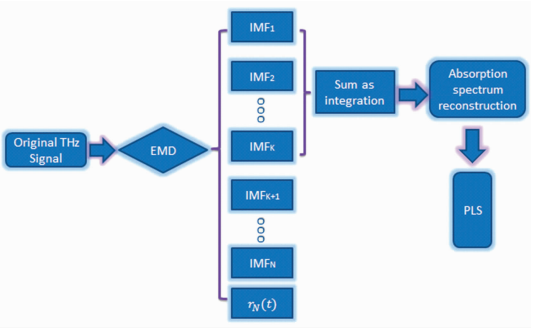

图 1 EMD-PLS 建模流程图

Fig. 1 Flowchart of EMD-PLS modeling

1.2 方法

氨基酸样品(L-苯丙氨酸和 L-酪氨酸)与聚乙烯粉末充分混合(L-苯丙氨酸质量占比分别为 0%，15%，25%，40%，50%，55%，60%，61%，64%，70%，75%，85%，95%，100%)，然后研磨成均匀的颗粒。并在 5 t 压力下被压成圆片。样品均购于 sigma-aldrich 公司。圆片样品的厚度约为 0.6 mm，直径为 13 mm。每个浓度的样品数量为 2，共有 28 个圆片样品。使用太赫兹时域光谱(THz-TDS)系统对样品进

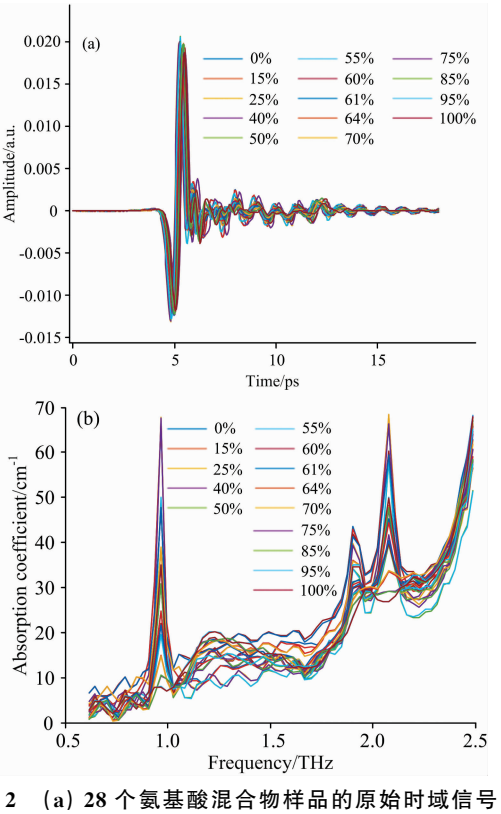

图 2 (a) 28 个氨基酸混合物样品的原始时域信号；(b) 28 个氨基酸混合物样品的原始吸收信号

Fig. 2 (a) 28 original temporal signals and (b) absorption spectra of 28 original amino acids samples

行测试，样品被放置在两个抛物面镜的焦点之间。所有测量均在 21 ℃下进行，相对湿度小于 4%。

2　结果与讨论

图 2(a)和(b)为 28 个原始 THz 时域信号及其在 0.7～2.5 THz 波段的吸收光谱。由图可知，该氨基酸混合物的三个吸收峰分别位于 0.97，1.9 和 2.08 THz。可以看出，随着 L-苯丙氨酸含量从 100%降至 0%，混合光谱吸收峰的幅值逐渐增加。所以，L-苯丙氨酸没有明显的特征峰，三个峰均来自于 L-酪氨酸。前人的理论模拟结果表明，L-酪氨酸的吸收峰主要由分子的振动和扭转引起，即分子的不同振动模式和强度产生了不同的吸收峰[16]。

此处仅对 L-苯丙氨酸浓度为 0%样品的时域信号进行举例分析，以说明信号分解的过程，其余 27 个时域信号均按此方法进行处理。图 3 为该样品的分解结果，它可以分解为 8 个 IMF 和 1 个残差函数。很明显，一阶 IMF(IMF1)信号具有最多的信号能量，而其余 IMF 信号能量随着阶数增加而减少。可以看出，低阶模式更接近原始时间信号。

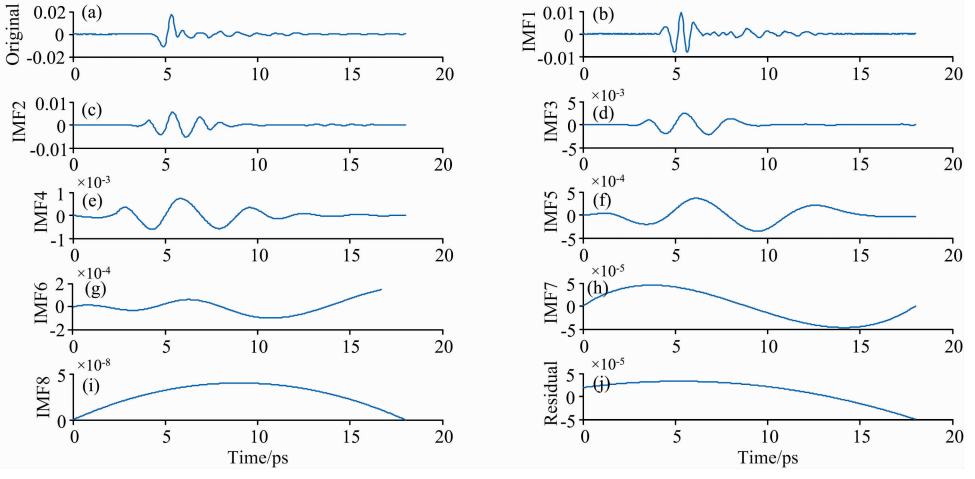

图 3　L-苯丙氨酸浓度为 0%样品 EMD 分解后的 IMF 和残差函数

Fig. 3　EMD decomposed IMFs and residual function of concentration＝0% sample

图 4 描述了 L-苯丙氨酸浓度为 0%样品的时域信号经过 EMD 分解后，第一个 IMF(IMF1)，前两个 IMF 叠加(IMF1＋IMF2)，前三个 IMF 叠加(IMF1＋IMF2＋IMF3)，前四个 IMF 叠加(IMF1＋IMF2＋IMF3＋IMF4)和前五个 IMF 叠加(IMF1＋IMF2＋IMF3＋IMF4＋IMF5)相应的吸收光谱。可以看出，由于低频信息不完整，IMF1 的吸收光谱明显不同于其他吸收光谱。因此进一步建模中，我们不考虑 IMF1。随后，使用 PLS 对剩余数据集与目标值之间建立了定量分析模型。此处，采用 Kennard-Stone 方法将数据集划分为校正和预测集。实验数据集中，18 个样本作为校正集，并将剩余的 10 个样本作为预测集。这五组 THz 吸收光谱(原始与分解后)的 PLS 统计分析结果列于表 1 中。与原始结果相比，前

表 1　对两种氨基酸混合物的 PLS 校正与预测效果

Table 1　PLS calibration and prediction performance statistics for binary mixtures

Signal components	R_{cal}	RMSEC /%	R_p	RMSEP /%
Original signal	0.994 6	0.021 8	0.990 0	0.025 1
IMF1＋IMF2	0.993 2	0.021 7	0.991 7	0.029 0
IMF1＋IMF2＋IMF3	0.995 2	0.019 2	0.993 7	0.022 4
IMF1＋IMF2＋IMF3＋IMF4	0.999 4	0.006 8	0.996 1	0.019 8
IMF1＋IMF2＋IMF3＋IMF4＋IMF5	1.000 0	0.000 5	0.990 9	0.033 0

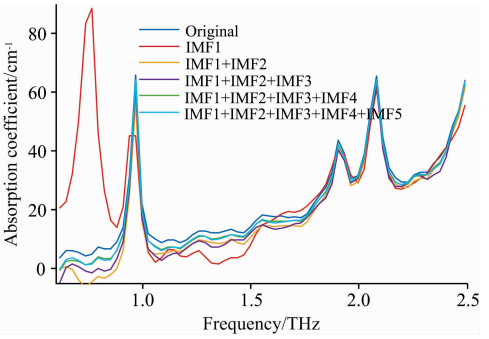

图 4　L-苯丙氨酸浓度为 0%的样品 EMD 分解后，第一个 IMF(IMF1)，前两个 IMF 叠加(IMF1＋IMF2)，前三个 IMF 叠加(IMF1＋IMF2＋IMF3)，前四个 IMF 叠加(IMF1＋IMF2＋IMF3＋IMF4))和前五个 IMF 叠加(IMF1＋IMF2＋IMF3＋IMF4＋IMF5)后相对应的吸收光谱

Fig. 4　The corresponding absorption spectra of concentration ＝0% sample for the first one IMFs (IMF1), two IMFs superposition (IMF1 ＋ IMF2), three IMFs superposition (IMF1 ＋ IMF2 ＋ IMF3), four IMFs superposition (IMF1＋IMF2＋IMF3＋IMF4) and five IMFs superposition (IMF1＋IMF2＋IMF3＋IMF4＋IMF5)

两个 IMF 叠加的预测效果不好，说明前两个 IMF 叠加删除冗余信息的同时丢失了某些有用信息。虽然前五个 IMF 叠加在校正集中有较好结果，但是其 RMSEP 较大，说明其中可能存在噪声导致过度拟合。通过比较可以确定前四个 IMF 叠加具有最佳的预测效果，这证实了 EMD 方法的有效性。

图 5(a)和(b)分别为单独使用 PLS 模型和使用 EMD-PLS 模型(基于前四个 IMF 之和的结果)对不同氨基酸混合物样品实际浓度与预测浓度之间的相关性，可以看出 EMD-PLS 模型可以获得更理想的预测结果。

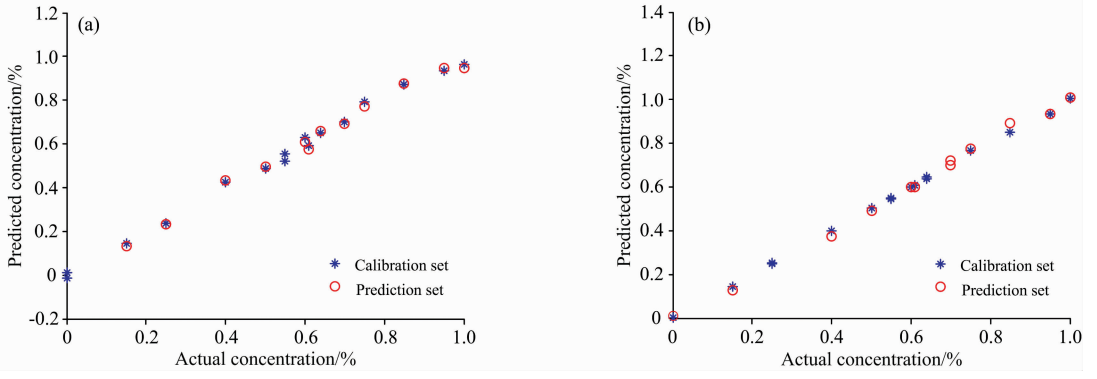

图 5 (a)PLS 模型和(b)EMD-PLS 模型(IMF1＋IMF2＋IMF3＋IMF4)下不同浓度氨基酸混合物样品的实际浓度与预测浓度的关系

Fig. 5 Scatter plots of the actual concentration versus the predicted concentration using (a) PLS model and (b) EMD-PLS model (IMF1＋IMF2＋IMF3＋IMF4) for different concentrations of amino acids mixture samples

3 结 论

提出了一种基于太赫兹光谱技术的多元校正模型(EMD-PLS)，对氨基酸混合物进行了定量分析。该方法首先通过 EMD 方法分解太赫兹时域信号，并将前几个 IMF 信号叠加替代原始信号，然后对原始信号和使用 EMD 处理信号对应的吸收谱进行 PLS 回归分析。定量分析结果表明，与其他吸收谱相比，基于前四个 IMF 叠加的吸收光谱具有更好的预测结果(R_p＝0.996 1 和 RMSEP＝0.019 8)，这说明 EMD 可以作为一种有效的预处理手段。该工作表明了基于 EMD 的太赫兹信号定量分析技术的有效性，证明了 EMD-PLS 模型可以实现较为理想的预测精度。

References

[1] Baxter J B, Guglietta G W. Analytical Chemistry, 2011, 83(12): 4342.
[2] El Haddad J, Bousquet B, Canioni L, et al. TrAC Trends in Analytical Chemistry, 2013, 44: 98.
[3] Yamaguchi M, Miyamaru F, Yamamoto K, et al. Applied Physics Letters, 2005, 86(5): 053903.
[4] Ponseca C S, Kambara O, Kawaguchi S, et al. Journal of Infrared Millimeter & Terahertz Waves, 2010, 31(7): 799.
[5] Yu F, Zuo J, Mu K-j, et al. International Symposium on Photoelectronic Detection and Imaging 2013: Terahertz Technologies and Applications, 2013, 8909.
[6] CHEN Tao, LI Zhi, MO Wei, et al(陈 涛, 李 智, 莫 玮, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2013, 33(5): 1220.
[7] Lu S, Zhang X, Zhang Z, et al. Food Chemistry, 2016, 211: 494.
[8] Bian X, Diwu P, Liu Y, et al. Journal of Chemometrics, 2018, 32(11): e2940.
[9] Huang N E, Zheng S, Long S R, et al. Proceedings Mathematical Physical & Engineering Sciences, 1998, 454(1971): 903.
[10] Liu H, Fan Y-X, Li L, et al. Optics Express, 2018, 26(21): 27279.
[11] Bian X, Zhang C, Liu P, et al. Chemometrics & Intelligent Laboratory Systems, 2017, 170: 29(21): 1880.
[12] Liu H, Fan Y X, Han X, et al. IEEE Photonics Technology Letters, 2017, 29(21): 1880.
[13] Su Y, Zheng X, Deng X. Journal of Infrared Millimeter & Terahertz Waves, 2017, 38(8): 972.
[14] Bian X, Li S, Lin L, et al. Analytica Chimica Acta, 2016, 925: 16.
[15] Zhang R, Wu T, Zhao Y. Optik, 2019, 183: 906.
[16] Yan Z, Hou D, Huang P, et al. Measurement Science and Technology, 2008, 19(1): 015602.

Terahertz Spectrum Analysis for Binary Amino Acids Mixture Based on Empirical Mode Decomposition

LIU Jing[1], LIU Hai-shun[2*], ZUO Jian[2], ZHANG Cun-lin[1,2*], ZHAO Yue-jin[1], LIANG Mei-yan[3]

1. School of Optoelectronics, Beijing Institute of Technology, Beijing 100081, China
2. Key Laboratory of Terahertz Optoelectronics, Ministry of Education, Capital Normal University, Beijing 100048, China
3. Department of Electronic Information Engineering, Shanxi University, Taiyuan 030013, China

Abstract L-Phenylalanine and L-Tyrosineplay essential roles in synthesizing neurotransmitters and hormones. The two amino acids have similar structures which lead to an obviously functional distinction between the two amino acids. Previous studies have shown that there are remarkable differences between the two amino acids on low-frequency vibrations. Recently, terahertz (THz) spectroscopy has been proven to be a useful technique on studying low-frequency dynamic of biologic molecules. Many multivariate calibration methods have been successfully applied to quantitative analysis multi-components spectra data due to the linear behaviors revealed by terahertz absorption spectra. However, the predictive performances of traditional calibration techniques are sometimes unsatisfied as only a single model is built between spectra and targets to predict the unknown samples. Thus, the ensemble modeling method with better accuracy came into being. The empirical mode decomposition (EMD) method, firstly proposed by Dr. Huang in 1998, is used to decompose the signal into a set of intrinsic mode functions (IMF) self-adaptively, which is widely applied in signal and spectra processing. We proposed an empirical mode decomposition (EMD) based partial least squares (PLS) method for terahertz spectra quantitative analysis on amino acids mixture with various concentrations. The terahertz time signals were decomposed into a series of intrinsic mode functions (IMF) with different frequencies by the EMD method. The several top IMFs (from 2 to 5) based absorption spectra were obtained for quantitative analysis by employing PLS. The predicted results indicated that the top four IMFs based absorption spectra acquired higher R (0.996 1) and lowered RMSEP (0.019 8) compared tothe single PLS regression and theother top several IMFs'results. Thus, the successful application with EMD-PLS method manifests the effectiveness in quantitative analysis of binary mixtures within the THz region.

Keywords Terahertz; Empirical mode decomposition; Partial least squares; Amino acids

(Received Aug. 16, 2019; accepted Dec. 28, 2019)

* Corresponding authors