Accepted Manuscript

Classification and identification of molecules through factor analysis method based on terahertz spectroscopy

Jianglou Huang, Jinsong Liu, Kejia Wang, Zhengang Yang, Xiaming Liu

PII:	S1386-1425(18)30213-0
DOI:	doi:10.1016/j.saa.2018.03.017
Reference:	SAA 15889
To appear in:	*Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*
Received date:	21 June 2017
Revised date:	2 March 2018
Accepted date:	8 March 2018

Please cite this article as: Jianglou Huang, Jinsong Liu, Kejia Wang, Zhengang Yang, Xiaming Liu , Classification and identification of molecules through factor analysis method based on terahertz spectroscopy. The address for the corresponding author was captured as affiliation for all authors. Please check if appropriate. Saa(2017), doi:10.1016/j.saa.2018.03.017

Classification and identification of molecules through factor analysis method based on terahertz spectroscopy

Jianglou Huang[a], Jinsong Liu[a*], Kejia Wang[a], Zhengang Yang[a], and Xiaming Liu[b]

[a]*Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China.*
[b]*Department of urology, Tongji Hospital, Tongli Medical College, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China.*
**jsliu4508@vip.sina.com*

Abstract: By means of factor analysis approach, a method of molecule classification is built based on the measured terahertz absorption spectra of the molecules. A data matrix can be obtained by sampling the absorption spectra at different frequency points. The data matrix is then decomposed into the product of two matrices: A weight matrix and a characteristic matrix. By using the K-means clustering to deal with the weight matrix, these molecules can be classified. A group of samples (spirobenzopyran, indole, styrene derivatives and inorganic salts) has been prepared, and measured via a terahertz time-domain spectrometer. These samples are classified with 75% accuracy compared to that directly classified via their molecular formulas.

Keywords: terahertz spectroscopy, organic macromolecules, factor analysis, classification

Introduction

During years of researches, terahertz (THz) wave has found itself an important role in chemical and biological fields(1–4) due to its non-destructive and non-invasive way of detecting samples. It has low sample damage or photoionization(5), with photon energy 1 million times weaker than X-ray. THz absorption spectrum has been measured for many types of molecules, including amino acids, saccharides, proteins, nucleotides and DNA(6–13) . However, most organic or biological molecules present very complex and obscure THz spectral lines, making very hard to resolve(14).

The classification of molecules is necessary and important in some fields, such as chemical, material, biological and medical treatment. In conventional infrared or visible light region, molecular spectra have very narrow spectral peaks and large peak separation, so the characteristic absorption peaks can be picked and assigned to distinguish molecular structures(15). For THz spectra, on the contrary, molecular absorption bands overlap each other heavily, making hard to use the THz spectra to classify molecules. In order to solve this problem, some methods have been developed to perform the classification of molecules by analyzing the THz absorption spectrum of the molecules, such as support vector machine (SVM)(16), principal component analysis (PCA) (17, 18) and multivariate data analysis (MDA)(18). In these methods, a $n \times m$ order data matrix is obtained by sampling the measured THz absorption curves of a group of samples at different frequency points, in which n is the sample number in the group and m is the number of the frequency points. The classification can be performed by treating the dada matrix. A quite satisfying accuracy rate can be obtained by SVM(19), but high algorithmic complexity and extensive memory requirement are unavoidable, although PCA and MDA can overcome these weaknesses in a certain extent.

In this paper, inspired by a multivariate analysis statistical method, named as factor