Оптический журнал

Физическая оптика

УДК 535.8, 617.7

Определение генетически модифицированных продуктов на основе анализа терагерцовых спектров многовесовой нейронной сетью

© 2021 г. Jianjun Liu, Tiejun Li, Senquan Yang, Lanlan Fan Fan, Fan Ding

Генетически модифицированные продукты всегда являются объектами пристального внимания с точки зрения безопасности продуктов питания. Для детектирования генетически модифицированных материалов предложена бионическая модель распознавания на основе многовесовой нейронной сети, анализирующей спектры поглощения образцов в терагерцовой области спектра. Для каждого класса обучающих примеров случайным образом выбраны 50 образцов в качестве обучающей выборки для многовесовой нейронной сети, которые использованы как первый тест для подтверждения качества распознавания. Другие образцы использованы для второго теста, определяющего качество различения. Экспериментальные результаты показали, что предложенная модель обеспечивает эффективную идентификацию генетически модифицированных продуктов со сходными спектроскопическими характеристиками и перспективна для обнаружения и контроля генетически модифицированных организмов.

Ключевые слова: *терагерцовая спектроскопия, метод главных компонентов, многовесовая нейронная сеть*

Detection of genetically modified substances based on terahertz and multi-weight vector neural network

© 2021 г. J. Liu*, PhD; T. Li**, PhD; S. Yang*, PhD;
L. Fan Fan*, postgraduate student; F. Ding, PhD*; X. Pan***, MD

**School of Intelligent Engineering, Shaoguan University, Shaoguan Guangdong, China*
***Scholl of Information Engineering Jimei University, Xiamen, Fujian, China*
****School of Electronics and Information Engineering, Hunan University of Science and Engineering, Yongzhou Hunan, China*

E-mail: liujianjun5518@qq.com

Submitted: 14.10.2020

DOI:10.17586/1023-5086-2021-88-07-03-11

Genetically modified food has always been a hot issue in the field of food safety. Genetically modified food has always been a hot issue in the field of food safety. In order to realize the detection of genetically modified materials, a bionic recognition model of multi-weight vector neural network is proposed by combining multi-weight vector neural network with terahertz time domain spectroscopy. In this paper, for each class of samples, 50 samples are randomly selected as the training set, a multi-weight vector neural network bionic recognition model is established, and 50 samples are selected as the first test set to verify the recognition rate. Other dissimilar samples are used as the second test

set to verify their misjudgment rate. The experimental results show that the model can effectively identify transgenic materials with similar spectral characteristics. The model proposed in this paper provides a new method for the detection and identification of genetically modified organisms.

Keyword: *terahertz spectroscopy, principal component, multi weight vector neuron.*

OCIS codes: *300.0300, 040.2235, 010.1030*

INTRODUCTION

In recent years, with the development of transgenic technology and the promotion of genetically modified crops, genetically modified food has entered people's daily life. The impact of genetically modified crops on human beings, especially in the field of safety, has been controversial, and more attention has been paid to safety evaluation methods [1–3]. At present, the main detection methods of genetically modified food are polymerase chain reaction (PCR) and visible and far infrared spectroscopy. In view of the necessity of a PCR method for the detection of international standard samples, there are some problems in the determination of visible and far infrared spectra, such as difficult determination of parameters, large amount of calculation and so on. Therefore, it is necessary to develop a rapid and simple method for the detection of genetically modified crops [4–9].

The terahertz (THz) range of the spectrum usually refers to electromagnetic waves with frequencies of 0.1–10 THz, or wavelengths from 0.03 to 3 mm, which is located between microwave and infrared light and belongs to the far infrared band [10–11]. Theoretical studies show that the vibrational and rotational frequencies of many biomolecules such as deoxyribonucleic acid and proteins are in the THz frequency range. Therefore, the detection of biological samples by THz time domain spectroscopy will produce resonance absorption peaks, which make it possible to identify biological samples by THz spectroscopy. At present, the identification of genetically modified food by THz technology is still in its infancy in China, and there are few reports on the detection of genetically modified substances. The characteristics of transgenic cotton seed were reported by THz time domain spectroscopy in Ref. [12]. The application of THz technique in the detection of transgenic soybean was reported in Ref. [13]. Therefore, the identification of transgenic materials by THz spectroscopy has important theoretical and practical significance. This paper uses rapid identification method of genetically modified sugar beet based on THz spectrum and multi weight vector neural network [14].

EXPERIMENTAL EQUIPMENT AND SAMPLES

The samples used in the experiment were purchased from the National Standard material Research Center with a purity of 99.8%. In order to reduce the absorption of THz wave by water in the sample, the solid sample was dried before pressing (the solid sample was baked in a drying box for more than half an hour). The temperature of the dryer is set at about 50 °C) to remove the absorption of moisture to THz waves and minimize the effect of moisture. Then, the dry

Table 1. Experimental sample information

Samples*	Type**	Diameter, cm	Shape	Thickness, mm	Sample number
A gene	+	1.2	Circular sheet	1.2	100
A gene parent	/	1.2	Circular sheet	1.2	100
B gene	+	1.2	Circular sheet	1.2	100
B gene parent	/	1.2	Circular sheet	1.2	100

Note. * In order to protect the intellectual property rights of the experimental samples, A and B were used to replace the hidden transgenic name, ** "+" as transgenic, "/" indicating non-transgenic.

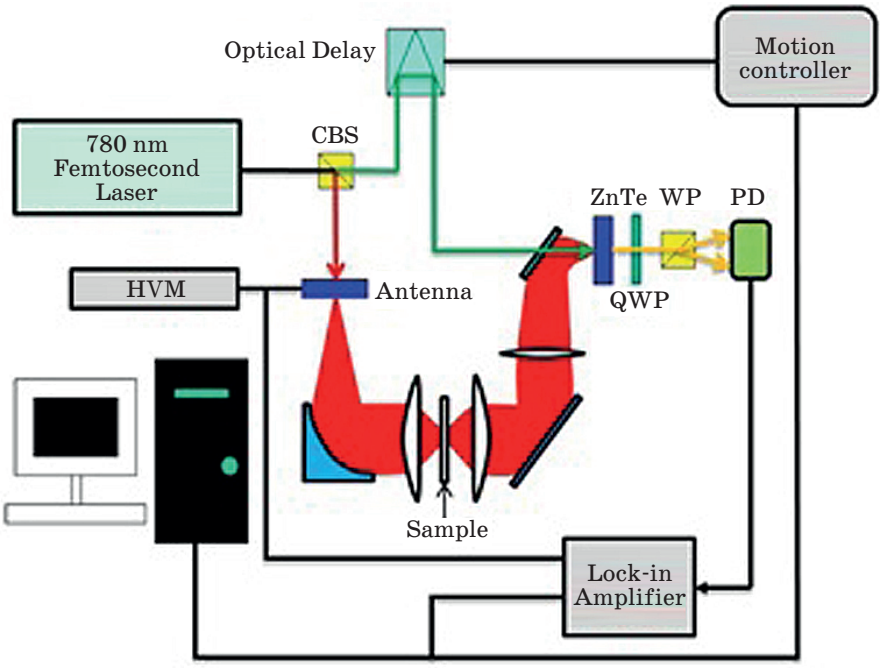

Fig. 1. The schematic diagram of THz time domain spectroscopy system.

solid sample and the appropriate amount of polyethylene powder are put into an agate mortar for careful grinding until they are mixed evenly. The appropriate amount of sample is put into the die of the press machine to press the sample into a circular sheet with a thickness of about 1.2 mm and a diameter of 1.2 cm left as the sample used in the experiment (the pressure of the press used in the experiment is 12 MPa). The details of the sample are shown in Table 1.

The THz time domain spectroscopy system used in this paper is a transmission THz time domain spectroscopy system as showed in Fig. 1 [15]. The central wavelength of the laser is 780 nm. In order to ensure the accuracy of the experiment, when the relative humidity in the system reaches 2%, the system is filled with dry air. The indoor relative humidity is 25% and the constant temperature is 292 K.

IDENTIFICATION MODEL OF MULTI-WEIGHT VECTOR NEURAL NETWORK

In order to solve the limitation of single weight neural network in practical application (as shown in Fig. 2a), Shoujue Wang proposed a more general neural network model in 2003 [16] (as shown in Fig. 2b).

The proposed model makes the description ability of affine pattern recognition based on neuron more flexible, and its mathematical expression is as follows:

$$\mathbf{Y} = f(\Phi(\mathbf{W}_1, \mathbf{W}_2, ..., \mathbf{W}_n, \mathbf{X})),$$

where f is a nonlinear mapping function, Φ represents a function of a scalar value, that is, a multi-input vector is converted to a scalar, $\mathbf{W}_1$, $\mathbf{W}_2$, ..., $\mathbf{W}_n$ represent n weight vectors, $\mathbf{X}$ and $\mathbf{Y}$ represent the input and output vectors of neurons respectively[1].

In the case of sample recognition, the result of answering the question is "*yes*" or "*no*", so the step function can be used, that is

$$f(x) = \begin{cases} 1 \text{ if } x \le k \\ 0 \text{ if } x \ge k \end{cases}.$$

Let the eigenspace be the n-dimensional real number space R^n, vector function equation be

$$\Phi(W_1, W_2, ..., W_n, X) = k,$$

where, k is a constant, which can be regarded as a trajectory of vector $\mathbf{X}$ in the feature space

[1] Editor remark. As the authors apply below the scalar formulae for calculating the values of $\mathbf{Y}$, mentioning and writing $\mathbf{Y}$ as scalar seems to be more correct.

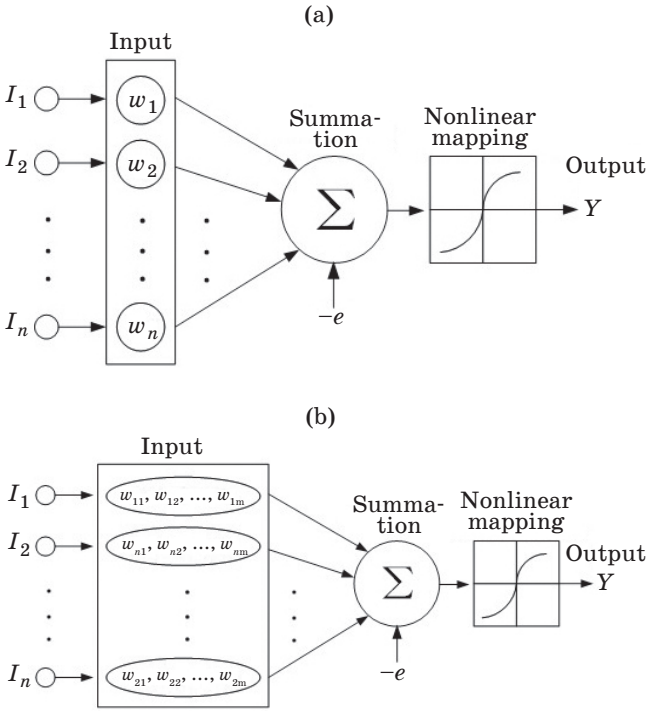

Fig. 2. Single weighted neural network (a) and multi-weighted vector neural network (b).

R^n determined by W_1, W_2, ..., W_n. This locus is a $(n-1)$-dimensional hypersurface (or hyperplane) of R^n, which divides R^n into two parts. On the one side, there are $\Phi \geq k$, $Y = 1$ and on the other side, there are $\Phi \leq k$, $Y = 0$. If $\Phi(W_1, W_2, ..., W_n, X) = k$ is a closed hypersurface, then a finite coverage area will be formed in the feature space. If an input vector $\mathbf{X}$ is in this coverage area, then $Y = 1$, otherwise $Y = 0$.

Let the feature vector mapped to the feature space of the collected sample be S_1, S_2, ..., S_n, and the weight vectors W_1, W_2, ..., W_n are determined by S_1, S_2, ..., S_n. That is

$$W_1 = w_1(S_1, S_2, ..., S_n),$$

$$W_2 = w_2(S_1, S_2, ..., S_n),$$

...

$$W_n = w_n(S_1, S_2, ..., S_n).$$

Therefore, $\Phi(W_1, W_2, ..., W_n, X)$ must be rewritten as $\Psi(S_1, S_2, ..., S_n, X)$.

By using the principle of bionic pattern recognition and the method of geometric analysis in high-dimensional space, a suitable function is found to make $X = S_1$ or $X = S_2$, ..., or $X = S_n$, so the above formula can be established. In this way, the feature space condition of the sample coverage is determined by the feature vectors S_1, S_2, ..., S_n, so the feature vector plays a role in the fusion of all sample information.

For pattern recognition, the multi-weight vector neural network (MWNN) is different from the single-weighted neurons such as BP, RBF and DBF, and f is usually a step function. Therefore, MWNN represents a complex geometric shape determined by multiple weights w_1, w_2, ..., w_n[1]. In fact, it can be found that BP, RBF and DBF are all special forms of MWNN.

Because the f is a step function, so a multi-weight neuron can construct a geometric shape with very complex structure, that is MWNN.

Therefore, the MWVNN is very suitable for constructing approximate coverage of a certain kind of sample distribution in bionic pattern recognition. According to the principle of homologous continuity of bionic pattern recognition, we must be able to find a suitable way to connect the adjacent two similar sample points in the high dimensional feature space, so that it can become a gradual order, and many low-dimensional popular topologies can be used to approximate this connection.

SPECTRAL FEATURE EXTRACTION

Terahertz absorption spectra of samples

When the experimental samples are prepared, the experimental samples are put into the THz time domain spectroscopy system one by one, and THz information in the 0.2–1.5 THz is extracted. To ensure the accuracy of the experimental data, each sample was scanned many times at different time and different position, and then the THz time domain spectral information of the experimental sample was obtained on average, and the THz optical parame-

[1] Editor remark. Though the authors use the notations w_1, w_2, ..., w_n as transformations resulting in vector values, it seems to be more understandable to replace these notations with the resulting vectors $\mathbf{W}_1$, $\mathbf{W}_2$, ..., $\mathbf{W}_n$ in the sentence above.

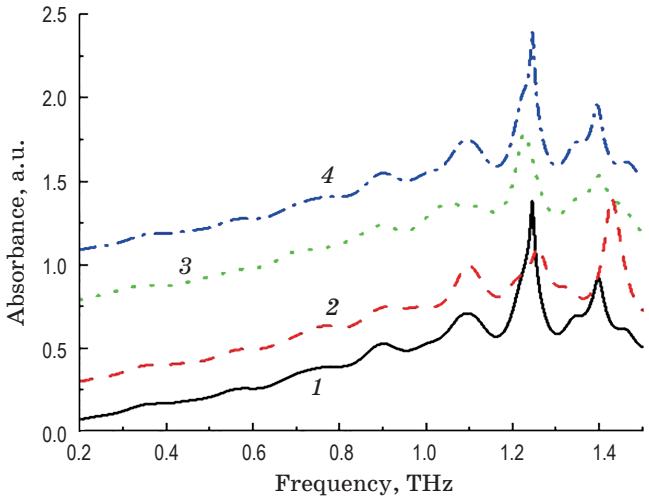

Fig. 3. Terahertz characteristic absorption peaks of four experimental samples. *1* — A gene, *2* — A gene parent, *3* — B gene, *4* — B gene parent.

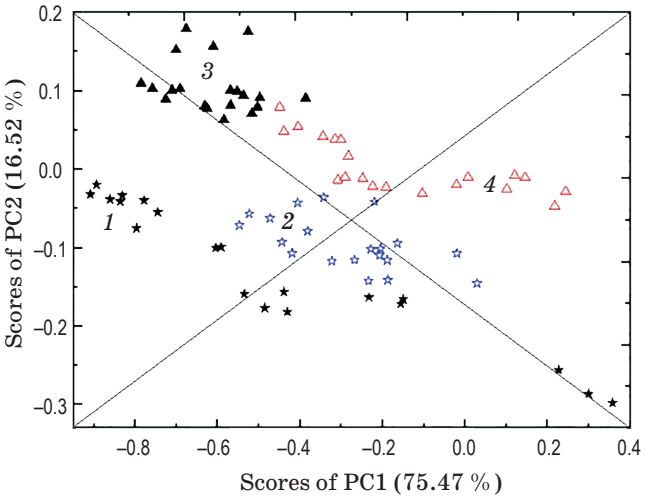

Fig. 4. Principal component score diagram of experimental sample. *1* — A gene, *2* — A gene parent, *3* — B gene, *4* — B gene parent.

ters were extracted according to the THz optical parameters. The THz characteristic absorption peak of the sample is obtained. Figure 3 shows the characteristic absorption peaks of different transgenic materials in 0.2–1.5 THz range. It can be seen from the Fig. 3 that the THz spectra of A gene and B gene are very similar to those of their respective parents. Therefore, it is difficult to distinguish all types of genetically modified substances from their characteristic peaks.

Spectral feature extraction

Transgenic gene and their parents have similar THz characteristic absorption peaks and high dimension in THz band, so they cannot be distinguished by screening feature information directly from original data. In this paper, the principal component features of the THz original data of four samples are extracted by using the principal component analysis (PCA) method [17–18], and the model is established by using the extracted principal components instead of the original high-dimensional data sets. The THz absorption spectrum data of four reference samples obtained in the experiment were analyzed by principal component analysis and the characteristic database of the samples was established. Figure 4 shows the score diagram of the first two principal components of the sample. It can be seen from the Fig. 4 that the eigenvalue information of the first two components can basically contain most of the information of the original spectral variables, and the main feature information of the sample can be extracted effectively.

Table 2. PCA score tables for four different reference samples

Sample	PC1	PC2
A gene	−0.82867	0.7535
A gene parent	0.33728	−1.35380
B gene	1.26521	0.75798
B gene parent	−0.77382	−0.15772

Among the principal components, the cumulative variance contribution rate of the first two principal components reached 91.99%, of which PC1 was 75.47% and PC2 was 16.52%. Therefore, the first two principal components represent the spectral eigenvalues of the sample. After principal component analysis, the original spectral data is changed from multi-dimensional to 2-dimensional. When the original THz spectral data is large, the dimension reduction feature extraction using PCA method is very useful for the real-time detection and recognition of biomolecules. Table 2 lists the scores of the first two principal components of the experimental samples after principal component analysis.

RESULTS AND DISCUSSION

In order to verify the effectiveness of this method, the Schmitt error balance analysis is carried out on the experimental samples, and the results are shown in Fig. 5. It can be seen from the diagram that each experimental sample can be maintained in an effective error range, and only a very small number of samples have a large error.

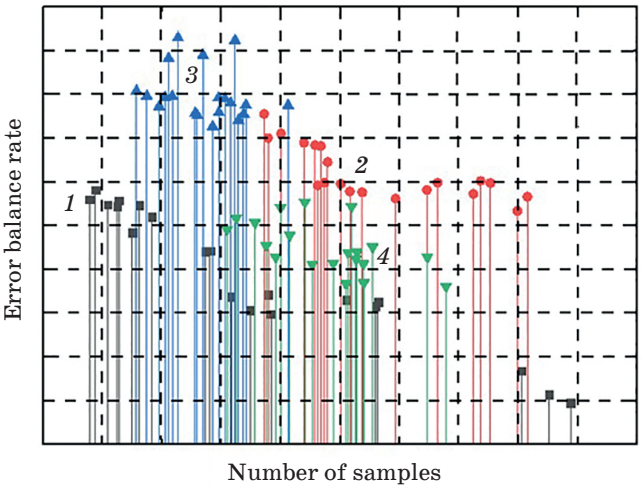

Fig. 5. The error balance analysis of samples. *1* — A gene, *2* — A gene parent, *3* — B gene, *4* — B gene parent.

In addition to verifying the recognition effect of the proposed MWVNN, the recognition effect of the proposed method is compared with that of support vector machine based on grid method (Grid-SVM) and support vector machine based on genetic algorithm (GA-SVM).

The recognition results of 200 test samples by three methods are given in Tables 3 to 5. Because the absorption spectra of transgenic samples are similar to those of their parents, the recognition effects of the three methods on the four kinds of samples are quite different. Among them, the method proposed in this paper can correctly identify transgenes and their parents. The GA-SVM method has the situation of misjudgment in the identification of experimental samples, of which the maximum misjudgment rate is up to 8%. The recognition effect of Grid-SVM on samples was the worst among the three methods, especially in the samples of A gene and its parents. Five A gene samples were misjudged as their parents, one A gene was misjudged

Table 3. The classification results of initial and cross validation by using MWVNN

MWVNN	Samples name	Samples name			
		A gene	A gene parent	B gene	B gene parent
Initial group*					
Number	A gene	50	0	0	0
	A gene parent	0	50	0	0
	B gene	0	0	50	0
	B gene parent	0	0	0	50
Percentage	A gene	100%	0	0	0
	A gene parent	0	100%	0	0
	B gene	0	0	100%	
	B gene parent	0	0	0	100%
Cross validation**					
Number	A gene	50	0	0	0
	A gene parent	0	50	0	0
	B gene	0	0	50	0
	B gene parent	0	0	0	50
Percentage	A gene	100%	0	0	0
	A gene parent	0	100%	0	0
	B gene	0	0	100%	0
	B gene parent	0	0	0	100%

Note. * 100% of the initial grouped samples have been correctly classified. ** Only the cases in the analysis are cross-validated. In cross-validation, each sample is classified according to a function derived from all samples other than that sample, and 100% of the cross-validation grouped samples have been correctly classified.

Table 4. The classification results of initial and cross validation by using GA-SVM

GA-SVM	Samples name	Samples name			
		A gene	A gene parent	B gene	B gene parent
Initial group*					
Number	A gene	46	2	2	0
	A gene parent	2	47	1	0
	B gene	0	0	48	2
	B gene parent	1	0	2	47
Percentage	A gene	92%	4%	4%	0
	A gene parent	4%	94%	2%	0
	B gene	0	0	96%	4%
	B gene parent	2%	0	4%	94%
Cross validation**					
Number	A gene	47	2	0	1
	A gene parent	2	47	0	1
	B gene	0	1	48	1
	B gene parent	1	0	1	48
Percentage	A gene	94%	4%	0	2%
	A gene parent	4%	94%	0	2%
	B gene	0	2%	96%	2%
	B gene parent	2%	0	2%	96%

Note. * 92% of the initial grouped samples have been correctly classified. ** Only the cases in the analysis are cross-validated. In cross-validation, each sample is classified according to a function derived from all samples other than that sample, and 94% of the cross-validation grouped samples have been correctly classified.

Table 5. The classification results of initial and cross validation by using Grid-SVM

Grid-SVM	Samples name	Samples name			
		A gene	A gene parent	B gene	B gene parent
Initial group*					
Number	A gene	43	5	1	1
	A gene parent	4	44	1	1
	B gene	1	2	43	4
	B gene parent	1	1	3	45
Percentage	A gene	86%	10%	2%	2%
	A gene parent	8%	88%	2%	2%
	B gene	2%	4%	86%	8%
	B gene parent	2%	2%	6%	90%
Cross validation**					
Number	A gene	44	5	0	1
	A gene parent	4	44	1	1
	B gene	2	0	45	3
	B gene parent	0	1	5	44
Percentage	A gene	88%	10%	0	2%
	A gene parent	8%	88%	2%	2%
	B gene	4%	0	90%	2%
	B gene parent	0	2%	10%	88%

Note. * 86% of the initial grouped samples have been correctly classified. ** Only the cases in the analysis are cross-validated. In cross-validation, each sample is classified according to a function derived from all samples other than that sample, and 88% of the cross-validation grouped samples have been correctly classified.

as B gene, and one A gene was misjudged as B gene parent. In general, the methods proposed in this paper, GA-SVM and Grid-SVM are different in distinguishing transgenes and their parents. In this paper, the recognition result of MWVNN is the best, GA-SVM is the second, and Grid-SVM is the worst.

CONCLUSION

In this paper, a method of transgenic beet recognition based on THz spectrum and multi-weight vector neural network was proposed. In order to establish the recognition model of multi-weight vector neurons in transgenic sugar beet and its parents, principal component analysis was used to extract THz spectral feature information. The results show that the recognition accuracy of the model proposed in this paper is high. This method provides an accurate, rapid, simple and nondestructive method for transgenic detection, and has a certain guiding significance for the detection of other transgenic products.

This work is supported by Support for scientific research projects (scientific research projects in colleges and universities) (No. 2019KTSCX165); supported by Foundation Funded Project of doctoral (No. 99000617); supported in part by research grants from the Science and Technology Program of Shaoguan (No. 2019sn056; 2019sn066), supported in part by the Key platforms and major scientific research projects of Universities in Guangdong (No. 2017KQNCX183), supported in part by the Key Project of Shaoguan University (No. SZ2017KJ08).

REFERENCE

1. *Jianjun Liu*. Recognition of genetically modified product based on affinity propagation clustering and terahertz spectroscopy // Spectrochem. Acta A. 2018. V. 194. P. 14–20.
2. *Jianjun Liu*. High-sensitivity detection method for organochlorine pesticide residues based on loop-shaped absorber // Mater. Chem. Phys. 2020. V. 242. P. 122542.
3. *Moreira Ivanira, Scarminio Ieda Spacino*. Chemometric discrimination of genetically modified Coffea arabica cultivars using spectroscopic and chromatographic fingerprints // Talanta. 2013. V. 107. № 3. P. 416–422.
4. *Vergragt P.J., Brown H.S*. Genetic engineering in agriculture: New approaches for risk management through sustainability reporting // Technol. Forecast. Soc. Change. 2008. V. 75. P. 783–798.
5. *Jianjun Liu, Lili Mao, Jinfeng Ku, et al*. THz spectroscopy detection method for GMOs based on adaptive particle swarm optimization // Opt. and Quantum Electron. 2016. V. 48. № 2. P. 167–173.
6. *Jianjun Liu, Zhi Li, Fangrong Hu, et al*. Identification of transgenic organisms based on terahertz spectroscopy and hyper sausage neuron // J. Appl. Spectrosc. 2015. V. 82. № 1. P. 104–110.
7. *Jianjun Liu, Zhi Li, Fangrong Hu, et al*. Hyper sausage neuron: Recognition of transgenic sugar-beet based on terahertz spectroscopy // Opt. and Spectrosc. 2015. V. 118. № 1. P. 175–180.
8. *Jianjun Liu, Zhi Li*. The terahertz spectrum detection of transgenic food // Optik. 2014. V. 125. № 23. P. 6867–6869.
9. *Nakamura K., Akiyama H., Kawano N., et al*. Evaluation of real-time PCR detection methods for detecting rice products contaminated by rice genetically modified with a CpTI-KDEL-T-nos transgenic construct // Food Chem. 2013. V. 141. P. 2618–2624.
10. *Alcantara G.B., Bsrison A., Santos M.S., et al*. Assessment of genetically modified soybean crops and different cultivars by Fourier transform infrared spectroscopy and chemometric analysis // Orbital. Electron. J. Chem. 2010.V. 2. P. 41–52.
11. *Burnett A.D., Fan W.H., Upadhya P.C., et al*. Broadband terahertz time-domain spectroscopy of drugs-of-abuse and the use of principal component analysis // Analyst. 2009. V. 134. P. 1658–1668.
12. *Jianjun Liu, Zhi Li, Fangrong Hu, et al*. Method for identifying transgenic cottons based on terahertz spectra and WLDA // Internat. J. Light Electron Opt. 2015.V. 126. P. 1872–1877.
13. *Nie Jun-Yang, Zhang Wen-Tao, Xiong Xian-Ming, et al*. Recognition of transgenic soybeans based on terahertz spectroscopy and PCA-BPN network // Acta Photonica Sinica. 2016. V. 45. № 5. P. 1–7.

14. *Jianjun Liu, Zhi Li, Fangrong Hu, et al.* Identification of transgenic organisms based on terahertz spectroscopy and hyper sausage neuron // J. Appl. Spectrosc. 2015. V. 82. № 1. P. 109–114.

15. *Liu J., Yang S.* Research on the terahertz temperature correlation of *L*-tyrosine and *L*-alanine // JOT. 2021. V. 88. P. 1–8.

16. *Wang Shou-Jue, Xu Jian, Wang Xian-Bao, et al.* Multi-camera human-face personal identification system based on the biomimetic pattern recognition // Acta Electronica Sinica. 2003. V. 31. № 1. P. 1–3.

17. *Xiao-Na, Huang Da-Zhuang, Shen Zu-Rui, et al.* Research on artificial neural network method used for insects classification and identification principal componet analysis and mathematical modeling // J. Biomathematics. 2013. V. 28. № 1. P. 23–33.

18. *Cao Fang, Wu Di, He Yong, et al.* Variety discrimination of grapes based on visible-near reflection infrared spectroscopy // Acta Optica Sinica. 2009.V. 29. № 2. P. 537–540.