Spectroscopy 22 (2008) 397–404
DOI 10.3233/SPE-2008-0352
IOS Press

A study on the early detection of colon cancer using the methods of wavelet feature extraction and SVM classifications of FTIR

Cun-Gui Cheng [a,*], Yu-Mei Tian [a] and Wen-Ying Jin [b]
[a] *Zhejiang Key Laboratory for Reactive Chemistry on Solid Surfaces, Department of Chemistry, Zhejiang Normal University, Jinhua 321004, P. R. China*
[b] *Department of Computer Science and Engineering, Yiwu Industrial and Commercial College, Yiwu 322000, P. R. China*

Abstract. This paper introduces a new method for the early detection of colon cancer using a combination of feature extraction based on wavelets for Fourier Transform Infrared Spectroscopy (FTIR) and classification using the Support Vector Machine (SVM). The FTIR data collected from 36 normal SD rats, 60 1,2-DMH-induced SD rats, and 44 second generation rats of those induced rats was first preprocessed. Then, 12 feature variants were extracted using continuous wavelet analysis. The extracted feature variants were then inputted into the SVM for classification of normal, dysplasia, early carcinoma, and advanced carcinoma. Among the kernel functions the SVM used, the Poly and RBF kernels had the highest accuracy rates. The accuracy of the Poly kernel in normal, dysplasia, early carcinoma, and advanced carcinoma were 100, 97.5, 95% and 100% respectively. The accuracy of RBF kernel in normal, dysplasia, early carcinoma, and advanced carcinoma was 100, 95, 95% and 100% respectively. The results indicated that this method could effectively and easily diagnose colon cancer in its early stages.

Keywords: FTIR, wavelet feature extraction, SVM, colonic earlier stage cancer

1. Introduction

Colon cancer is a potent disease that is one of the major causes of mortality in both men and women. Curing this disease depends on early diagnosis and the prompt treatment that follows [1–6]. If colorectal cancer is detected at an early stage, then the 5-year relative survival rate is 90%; however only 37% of colorectal cancers are diagnosed at early stages [7]. Apart from conventional methods of cancer diagnosis, there is a need to develop new approaches which are simple, objective, accurate, quick, convenient, and inexpensive. Fourier transform mid-infrared spectroscopy has been used to detect the carcinoma of several types of organs [8–12] because of its sensitivity in detecting molecular composition and structure. Recently, more improved methods of FTIR to detect cancer have been reported [13–15]. A method for identifying malignant colon cell by means of FTIR microspectroscopy and artificial neural network (ANN) is presented in [7]. In this article, thin sections of human normal, polp, and malignant tissues were used from the biopsies of 24 patients. However, FTIR data obtaining directly from sample's spectroscopy is usually affected by man-induced factor, and the number of tissue samples available was low. Therefore, the computational method they used had a low accuracy rating.

* Corresponding author: Cun-Gui Cheng, Department of Chemistry, Zhejiang Normal University, Zhejiang, Jinhua 321004, P. R. China. Tel.: +86 0579 83126962; Fax: +86 0579 82282489; E-mail: ccg@zjnu.cn.

Wavelet transform plays an important role in signal analysis and feature extraction. It can be used to detect the singularity of a signal and identify a small difference between two signals. The wavelet transform can be divided into two categories: discrete wavelet transforms (DWT) and continuous wavelet transforms (CWT). Bai et al. [16] reported a feature extracting method based on the dyadic wavelet transform (DWT) for FTIR cancer data analysis. The cost of DWT computation is less than the cost of CWT computation. However, the CWT is better in detecting the singularity of a signal.

Support vector machine (SVM) is a state-of-the-art classification technique which has a good theoretical foundation in statistical learning theory. SVM fixes the classification decision function based on structural risk minimum mistake instead of the minimum mistake of the misclassification on the training set to avoid over-fitting problem. It performs binary classification problem by finding maximal margin hyperplanes in terms of a subset of the input data (support vectors) between different classes. If the input data is not linearly separable, SVM first maps the data into a high dimensional feature space and then classifies the data by the maximal margin hyperplanes [17].

In our research, FTIR spectroscopy, using a single bounce HATR accessory, was used to detect normal and different stage colonic cancer tissues of rates. We focused primarily on how to identify normal, dysplasia, early cancerous, and advanced cancerous tissues of rates more efficiently. Classifying normal, dysplasia, and early cancerous tissues was difficult because their Fourier infrared spectrums were very similar. In order to improve classification accuracy, some important features needed to be extracted in the CWT domain. These features were inputted into the SVM to identify the normal, dysplasia, early carcinoma, and advanced carcinoma tissues of rates. The results showed that the new method was more efficient and much better in identification than traditional methods solely based on Fourier infrared spectrum.

2. Materials and methods

2.1. Preparation of samples

120 normal SD (Sprague-Dawley) rats, 6–8 weeks old, about 180 g, were used in the experiment. All rats were provided by the Department of Animals of Jinhua Institute for Drug Control (P. R. China). The induced reagent of colon cancer was syn-dimethylhydrazine dihydrochloride (DMH · 2HCl). An aqueous solution of 1,2-DMH · 2HCl was injected into the SD rats' abdominal cavity weekly for 22 weeks. The dosage was 25 mg · kg^{-1} and the concentration was 3%. The sample tissues were taken from 60 1,2-DMH-induced SD colon cancer models, 44 second generation rats of induced rats (no injection of 1,2-DMH), and 36 normal rats with a single-side cutter. The collected sample tissues were washed in physiology salt water and then put into the Fourier Transform Infrared Spectrometer to collect their FTIR data. The specimens were taken at the same time for pathological analysis. The sample tissues included 100 normal tissues, 120 dysplastic tissues, 100 early staged cancerous tissues and 110 advanced cancerous tissues.

2.2. Fourier transform infrared measurements

Fourier transform infrared spectra were measured using a Nicolet NEXUS 670 FTIR spectrometer (ThermoFisher, Madison, Wisconsin) equipped with a deuterated triglycine sulfate (DTGS) detector in the region 4000–650 cm^{-1}. All spectra were recorded as 64 scans with 4 cm^{-1} resolution.

The sampling accessory was equipped with a horizontal attenuated total reflectance (HATR) accessory with a single bounce reflection in the FTIR spectrometer. The FTIR spectrum background was recorded before collecting the sample's FTIR spectrum, and samples were collected directly by FTIR spectrometer with HATR. Three spectra were collected for each sample and the average was used for further analysis.

2.3. Continuous wavelet transform (CWT) [17]

The continuous wavelet transform is the sum of the signal multiplied by the scaled and shifted mother wavelet function. The wavelet coefficients are a function of scale and position. Normally wavelet decomposition consists of calculating the "resemblance coefficients" between the signal and the wavelet located at position b of scale a. If the coefficients are large, the resemblance is strong. Otherwise it is weak. The coefficients $C(a, b)$ are calculated as below:

$$C(a,b) = \int_R s(t)\frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right)\mathrm{d}t, \tag{1}$$

$$a \in R^+ - \{0\},\ b \in R, \tag{2}$$

ψ: wavelets, s: signal, a: scale, b: position.

2.4. Basic theory of support vector machines (SVMs) [18]

This classification procedure is based on the statistical learning theory proposed by Vapnik and Chervonenkis. The SVM uses structural risk minimization, rather than a non-convex, unconstrained minimization problem, as in standard neural network training technique using empirical risk minimization. Assume that the training data with k number of samples is represented by $\{x_i, y_i\}, i = 1, 2, \ldots, k$, where $x \in R^n$ is an n dimensional vector and $y \in \{-1, +1\}$ is the class label. These training patterns are said to be linearly separable if a vector ω and a scalar β can be defined so that inequalities (3) and (4) are satisfied:

$$\omega \cdot x_i + \beta \geqslant +1, \quad \text{for all } y = +1, \tag{3}$$

$$\omega \cdot x_i + \beta \leqslant -1, \quad \text{for all } y = -1. \tag{4}$$

The aim is to find a hyperplane that divides the data so that all the points with the same label are on the same side of the hyperplane. These amounts to finding ω and β such that:

$$y_i(\omega \cdot x_{i+}\beta) > 0. \tag{5}$$

If a hyperplane exists that satisfies (5), the two classes is said to be linearly separable. In this case, it is always possible to rescale ω and β so that $\min_{1 \leqslant i \leqslant k} y_i(\omega \cdot x_i + \beta) \geqslant 1$. That is, the distance from the closest point to the hyperplane is $1/\|\omega\|$. Then (6) can be written as

$$y_i(\omega \cdot x_i + \beta) \geqslant 1. \tag{6}$$

The hyperplane for which the distance to the closest point is maximal is called the optimal separating hyperplane (OSH). As the distance to the closest point is $1/\|\omega\|$, the OSH can be found by minimiz-

ing $\|\omega\|^2$ under constraint (6). The minimization procedure uses Lagrange multipliers and quadratic programming (QP) optimization methods. If $\alpha_i \geqslant 0$, $i = 1, \ldots, k$, are the non-negative Lagrange multipliers associated with constraint (6), the optimization problem becomes one of maximizing:

$$L(\alpha) = \sum_i \alpha_{\mathrm{i}} - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (x_i x_j). \tag{7}$$

Under constrains: $\alpha_i \geqslant 0$, $i = 1, \ldots, k$. If $\alpha^m = (\alpha_1^m, \ldots, \alpha_k^m)$ is an optimal solution of the maximization problem (7) then the optimal separating hyperplane can be expressed as

$$\omega_m = \sum_i y_i \alpha_i^m x_i. \tag{8}$$

If the data are not linearly separable then a slack variable ξ_i, $i = 1, \ldots, k$, can be introduced with $\xi_i = 0$ such that (12) can be written as Eq. (9), and the solution to find a generalized OSH, also called a soft margin hyperplane, can be obtained using the conditions, Eqs (10)–(12):

$$y_i(\omega \cdot x_i + \beta) - 1 + \xi_i \geqslant 0, \tag{9}$$

$$\min_{\alpha\beta\xi_1,\ldots,\xi_k} \left[\frac{1}{2}\right] \|\omega\|^2 + C \sum_{i=1}^{k} \xi_i, \tag{10}$$

$$y_i(\omega \cdot x_i + \beta) - 1 + \xi_i \geqslant 0, \tag{11}$$

$$\xi_i \geqslant 0,\ i = 1, \ldots, k. \tag{12}$$

The first term in (10) is the same as in the linearly separable case to control the learning capacity. The second term controls the number of misclassified points, and parameter C is chosen by the user. A larger value of C assigns a higher penalty to errors.

In situations in which it is not possible to have a hyperplane defined by linear equations on the training data, the techniques discussed for linearly separable data can be extended to allow for non-linear decision surfaces. A technique introduced by Vapnik maps input data into a high dimensional feature space through some non-linear mapping. The transformation to a higher dimensional space spreads the data out in a way that facilitates the finding of linear hyperplane. After replacing x by its mapping in the feature space $\Phi(x)$, Eq. (7) can be written as

$$L(\alpha) = \sum_i \alpha_{\mathrm{i}} - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \big(\Psi(x_i)\Psi(x_j)\big). \tag{13}$$

It is convenient to introduce the concept of the *kernel function* K in order to make the computation easier in feature space, such as Eq. (14):

$$K(x_i, x_j) = \Psi(x_i)\Psi(x_j). \tag{14}$$

It is the kernel function that performs the non-linear mapping. Thus, to solve Eq. (7), only the kernel function is computed rather than $\Phi(x)$, which could be computationally expensive. Equation (15) can be used to classification function:

$$Y = \text{sign}\left\{\sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j) + \beta\right\}. \tag{15}$$

In brief, SVM first maps the data which are not linearly separable into a high-dimensional feature space. It then classifies the data by the maximal margin hyper-planes.

2.5. Data analysis

Absorbance from different wave bands was collected based on the characteristics of absorption values. A chart with the PCA case scores from the main composition analysis was then created. With Morlet wavelet as the original function, one-dimension scale continuous wavelets from different samples were transformed to collect the sample's character variables. The Matlab 6.1 software was then used to process the sample's estimated continuous wavelet. The training sample number was 100. The experiment samples included 20 normal tissues, 140 early staged cancerous tissues and 160 advanced cancerous tissue. The selected character variables were used for SVM training and verification.

3. Results and discussion

3.1. FTIR analysis

Figure 1 is the HATR-FTIR spectra of normal, dysplasia, early carcinoma and advanced carcinoma tissues of rates.

As shown in Fig. 1, the location, intensity, and shape of the absorbance peak changes with the different sample tissues. The Hydroxyl absorption peak from protein, nucleic acid and grease is located at 3400 cm^{-1} with a similar intensity. The Carbonyl group absorption peak from protein is located at 1648.96 cm^{-1}, and its intensity decreases as the cancerous tissue progresses. Absorption peak from amide II band is located at 1538.11 cm^{-1}, and its intensity also decreases as the cancerous tissue progresses. The other spectrum bands, such as symmetrical flexing vibration and unsymmetrical flexing vibration of di-phosphate ester from nucleic acid, had absorption peaks located at 1083.60 cm^{-1} and 1242.49 cm^{-1} respectively. The absorption peak of collagen protein is located at 1342.26 cm^{-1}.

The higher information quantities exist in 3600–2800 cm^{-1} and 1800–650 cm^{-1}. Because the former includes the absorbability of O–H and N–H stretching bands, and the character is not obvious. The later includes fingerprint region which contains more molecule structure information. Thus we select the region (1800–650 cm^{-1}) to extract the feature of spectra.

3.2. Continuous wavelet transform (CWT)

The proper wavelet base and decomposition level was decided by analyzing the signal spectra property and comparing the decomposition results with different wavelet bases and decomposition levels. We chose three feature peaks in the CWT domain to extract the features of FTIR. Morlet wavelet was selected as the analysis wavelet. The wavelet decomposition level was set as 18.

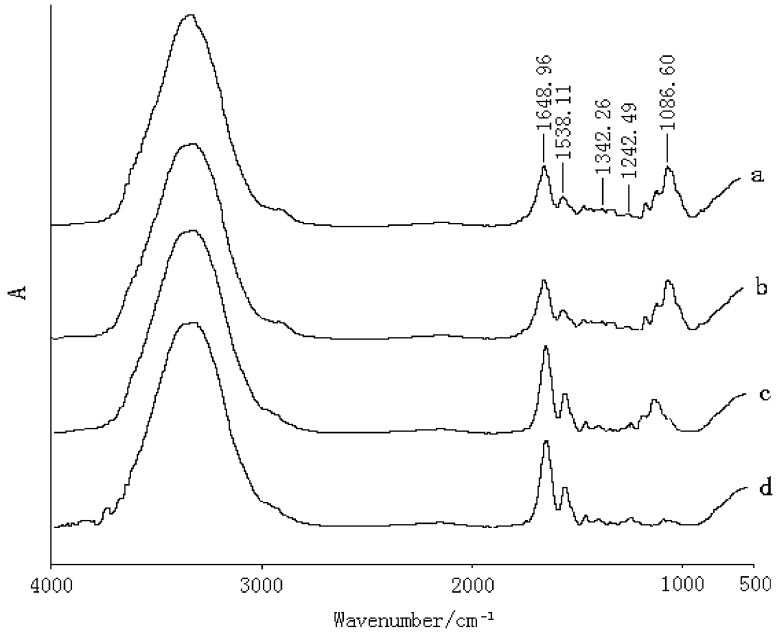

Fig. 1. FTIR spectra of normal (a), dysplasia (b), early carcinoma (c) and advanced colon carcinoma (d) tissues of rates.

The pre-processed Fourier infrared spectra of cancerous colon tissue and its CWT coefficients are shown in Fig. 2; L1–L15 contains the detailed information after decomposition (fine-to-coarse).

From Fig. 2, we can see that the detailed signal L1 is a high frequency signal and that the signals L5, L10, and L15 are more sensitive to the changes of the spectra. Since L5, L10 and L15 had a strong response to the three feature peaks of the original signal, they were used as feature vector spaces. Figure 3 shows a diagram of a divided feature space. Each detailed signal selected four feature bands, which corresponds with the four feature peaks. The feature vectors were defined as the average value of energy at each feature band. Thus, twelve feature variances were generated from the three detailed signals.

3.3. Data standardization

Standardization of the input sample data is required for the special kernel function because it would improve the parameters of the Hessian matrix for optimization and ensure the consistency of the training method. The data standardization process is of the following:

$$\hat{x}^l = \frac{x^l - \text{Min}\{x^l\}}{\text{Max}\{x^l\} - \text{Min}\{x^l\}}, \tag{16}$$

x^l is the feature parameter of the input data. $\text{Max}\{x^l\}$ and $\text{Min}\{x^l\}$ are the maximum and minimum values of this parameter.

3.4. Kernel function selection and results

Support vector machine transforms the non-linear classification problem to a linear problem in a high-dimension feature space by applying a kernel function. Some common kernel functions include Poly-

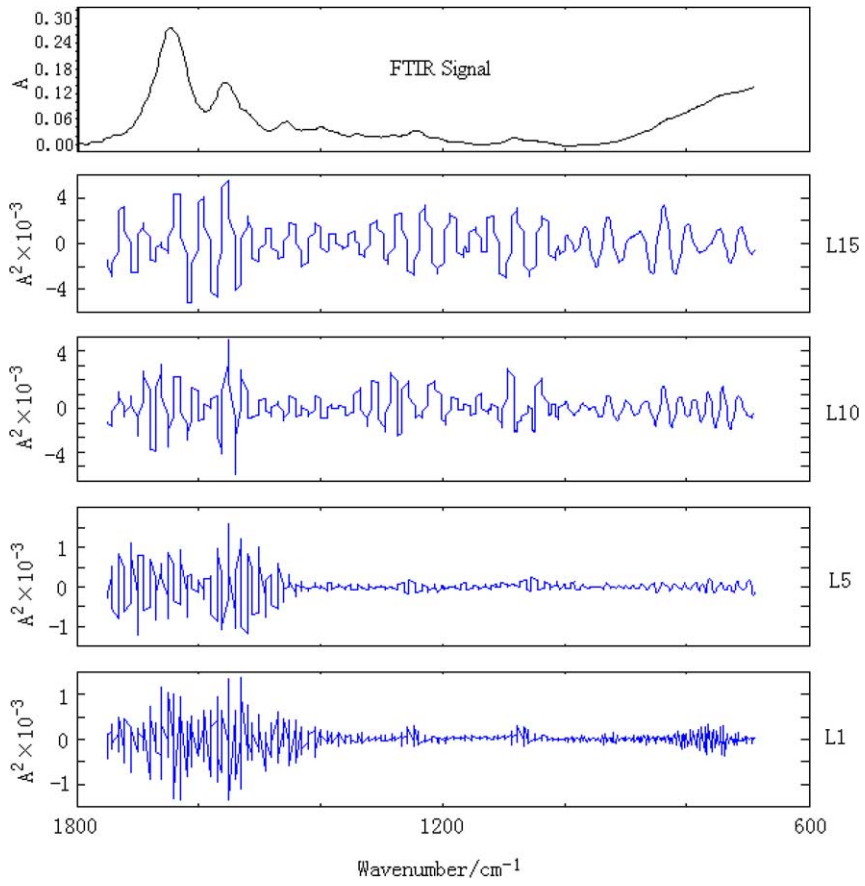

Fig. 2. The result of pre-processed spectra with CWT.

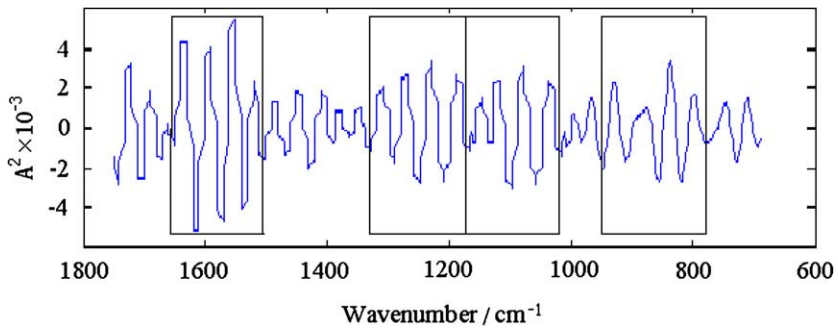

Fig. 3. Division of three feature regions of detail signal in the CWT domain.

nomial function, Radial Basis function, and Sigmoid function. Table 1 shows the classification results of sample tissues for normal, dysplasia, early carcinoma, and advanced carcinoma with different kernel functions and a linear function.

The higher accuracy results from Polynomial and Radial Basis kernel functions were the same as the pathological results. This means that the method using wavelet feature extraction and SVM classification

Table 1
The results of SVMs using different kernel (%)

Samples		Kernel			
		Linear	Polynomial	RBF	Sigmoid
Train	Normal	98.8	100	100	93.8
	Dysplasia	92.5	97.5	93.8	87.5
	Early carcinoma	95	98.8	98.8	92.5
	Advanced carcinoma	98.8	100	100	98.8
Test	Normal	95	100	100	95
	Dysplasia	90	97.5	95	97.5
	Early carcinoma	90	95	95	85
	Advanced carcinoma	96.7	100	100	96.7

of FTIR is an efficient and accurate method for identifying the normal, dysplasia, early carcinoma, and advanced carcinoma tissues.

4. Conclusion

It is very difficult to distinguish the normal, dysplasia, and early staged cancerous tissues of rates because there are no significant differences among them. But by applying the wavelet feature extraction of the FTIR data and SVM classification, better results can be achieved to distinguish between these tissues. This will facilitate and greatly aid in the early detection of colon cancer tissues of rates.

References

[1] C.G. Cheng, Y.M. Tian and W.Y. Jin, *Acta Chim. Sinica* **65** (2007), 2539–2543.
[2] F.K.F. Michael, S. Mary, E. Pascale, F. Wylam, Z.M. Nadia and T.T.W. Patrick, *Gynecol. Oncol.* **66** (1997), 10–15.
[3] Y.Z. Xu, L.M. Yang, Z. Xu, Y. Zhao, X.F. Ling, Q.B. Li, J.S. Wang, N.W. Zhang, Y.F. Zhang and J.G. Wu, *Chin. Sci. Ser. B* **34** (6) (2004), 465–472.
[4] J. Zhou, L.Q. Zhao, M.M. Xiong, X.Q. Wang, G.R. Yang, Z.L. Qiu, M. Wu and Z.H. Liu, *World J. Gastroenterol.* **9** (2003), 9–15.
[5] P. Czeslawa and W.M. Kwiatek, *J. Mol. Struct*. **565/566** (2001), 329–334.
[6] H.P. Wang, H.C. Wang and Y.J. Huang, *Sci. Total Environment* **204** (1997), 283–287.
[7] S. Argov, J.R.A. Salman, I.S.J. Goldstien, H. Guterman and S. Mordechai, *J. Biomed. Opt*. **7** (2002), 248–254.
[8] A.R. Shaw, Y.S. Low, M. Leroux and H.H. Mantsch, *Clin. Chem.* **47** (2000), 1493–1495.
[9] C.G. Cheng, H.Q. Shi, X.J. Zhu, R.Q. Zhen and S.T. Zhu, *Spectrosc. Spect. Anal*. **24** (2004), 1342–1344.
[10] Z.F. Cheng, L.Y. Cheng, W.Y. Jin and C.G. Cheng, *J. Harbin Eng. Univ*. **27**(Suppl.) (2006), 366–369.
[11] R.K. Sahu and S. Mordechai, *Future Oncol*. **1** (2005), 635–647.
[12] S. Mark, R.K. Sahu, K. Kantarovich, A. Podshyvalov, H. Guterman, J. Jagannathan, S. Argov and S. Mordechai, *J. Biomed. Opt*. **9** (2004), 558–567.
[13] Q.B. Li, Z. Xu, Y.Z. Xu, Y.F. Zhang, N.W. Zhang, L.X. Wang, X.J. Sun, L. Zhang, F. Wang, L.M. Yang, Y. Zhao, Y. Ren, Z. Liu, S.F. Weng, W.J. Zhou and J.G. Wu, *Chem. J. Chin. Univ*. **25** (2004), 2010–2012.
[14] B. Rigas and P. Wong, *Cancer Res*. **52** (1992), 84–88.
[15] R.G. Peter, H.S. Yang, Q.B. Li, X.F. Ling, J.S. Wang, L.M. Yang, Y.Z. Xu, S.F. Wen and J.G. Wu, *Spectroc. Spect. Anal*. **24** (2004), 1025–1027.
[16] L. Bai and Y.H. Liu, *Proc. of SPIE* **6001** (2005), 83–89.
[17] C.G. Cheng, G.L. Sun and C.J. Zhang, *Spectroscopy* **22**(11) (2007), 38–42.
[18] Q.S. Chen, J.W. Zhao, C.H. Fang and D.M. Wang, *Spectrochim. Acta Part A* **66** (2007), 568–574.