Application of wavelet transform in infrared spectrometry: Spectral compression and library search

ARTICLE *in* CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS · SEPTEMBER 1998

Impact Factor: 2.32 · DOI: 10.1016/S0169-7439(98)00084-7

4 AUTHORS, INCLUDING:

Foo Tim Chau
The Hong Kong Polytechnic University
212 PUBLICATIONS **3,140** CITATIONS

ELSEVIER

Chemometrics and Intelligent Laboratory Systems 43 (1998) 69–88

Chemometrics and intelligent laboratory systems

Application of wavelet transform in infrared spectrometry: spectral compression and library search

Alexander Kai-man Leung [a], Foo-tim Chau [a,*], Jun-bin Gao [b], Tsi-min Shih [c]

[a] *Union Laboratory of Asymmetric Synthesis and Department of Applied Biology and Chemical Technology, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China*

[b] *Department of Mathematics, Huazhong University of Science and Technology, Wuhan 430074, China*

[c] *Department of Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China*

Received 10 October 1997; accepted 20 May 1998

Abstract

In recent years, a new mathematical technique called wavelet transform (WT) has been developed and adopted for signal processing in analytical chemistry owing to its efficiency, more number of basis functions available, and higher speed in data treatment compared to fast Fourier transform (FFT). In this paper, the fast wavelet transform (FWT) and its derivative, wavelet packet transform (WPT), were applied to compress infrared (IR) spectrum for storage and spectral searching. In WT treatment, the number of data to be processed has to be 2^P with P being any integer. In this work, we proposed the coefficient position retaining (CPR) method to handle data with length of odd number. The performance of the two proposed WT methods in data compression and spectral library searching are compared with that of the FFT method. The results indicated that our proposed WT methods works better than FFT in compression of IR spectra and spectral library searching.

Keywords: Coefficient position retaining method; Fast wavelet transform; Wavelet packet transform; Compression; Spectral library search; Infrared spectrum

1. Introduction

The development of information technology in chemistry is very important because the type and volume of chemical data increases dramatically in recent years. According to the world's largest indexing system, *The Eleventh Collective Index*, about 28×10^6 chemical substances were known by the end of 1989 [1]. Besides, most chemical instruments are now computerized owing to the rapid development of microelectronics technology. Nowadays, a microcomputer is usually connected to an instrument for control of the device, data acquisition, signal processing, interpretation and reporting [2]. Recently, there is a growing trend in combining different chemical devices together to give hyphenated instruments for multi-dimensional studies. This approach greatly

* Corresponding author. Tel.: +852-2766-5603; Fax: +852-2364-9932; E-mail: bcftchau@polyu.edu.hk

0169-7439/98/$ - see front matter
PII: S0169-7439(98)00084-7

enhances information acquisition and even allows experimental work not possible to be achieved before. However, advancement in these multi-dimensional techniques and the huge number of existing chemical substances leads to some serious problems. One of them is that piles of experimental data are generated and are required to be archived for future usage. One possible way to reduce the storage space and processing time is through signal compression. In chemical analysis, signal compression is very important especially in setting up digitized spectral library [3] to diminish the size of the original database and to reduce the time for spectral searching.

Rapid development in computer technology leads to a lot of electronic spectral libraries and database available in the market [4]. Most of them cover compounds fewer than 100,000 out of 10 million known chemical compounds in different formats such as digitized IR, NMR and mass spectra. In order to identify the spectrum of an unknown compound from a reference library, spectral library search techniques are required. The output of a search usually comprises spectra that are most similar to that of the compound under study, together with an indication of the degree of similarity in each case [3]. So, a fast and highly accurate library search algorithm is desirable. Common algorithms of this kind adopted in chemistry include peak position searching method, direct matching method, spectral simulation method as well as artificial intelligence and pattern recognition techniques such as neural network, fuzzy theory and expert system [3].

To carry out any search, the spectral library must be constructed first from a set of reference spectra and accompanied by additional information such as structure, name, connection data, and molecular mass of individual compound [5]. In order to reduce the storage space of spectral data, different compression techniques have been developed. Data compression methods that are commonly used in chemistry include binary encoding [6–8], spline [9–11], Fourier transform [12–15] and factor analysis [16–18]. Currently, Fourier transform (FT) and the inverse FT are the major tools adopted to process experimental data in chemical studies [19,20]. Since 1989, a new mathematical technique called wavelet transform (WT) has been proposed for signal processing in various fields of analytical chemistry owing to its efficiency, more number of basis functions available, and speed in data treatment when it is compared with FT. About 80 papers have been published by applying WT in analytical chemistry [21]. These include work on flow-injection analysis [22], high performance liquid chromatography [23–25], infrared spectrometry [26–30], mass spectrometry [31–33], nuclear magnetic resonance spectrometry [34], potentiometric titration [35], ultra-violet visible spectrometry [36,37] and voltammetry [38–42] for data preprocessing. Besides, WT is also employed for general signal processing in chemistry [43–49] and quantum chemistry [50–55]. In very recent years, Chau et al. have also successfully applied WT to smooth and compress UV–VIS and IR spectra [56–58].

Different methods have been reported in the literature to facilitate searching of IR spectral database. Some of them are based on peak width and intensity [59,60], principal component analysis [16] and Fourier transform and interferogram [61–63]. In this paper, new data compression procedures have been developed to manipulate IR spectra by utilizing the fast wavelet transform (FWT) and wavelet packet decomposition (WPT) techniques. In this approach, the IR spectrum is converted into the wavelet domain through FWT and WPT treatment. In order to minimize the storage space of the spectrum, an absolute cutoff method was utilized to select the coefficients to be archived. Besides, the Shannon–Weaver entropy calculation was also employed to choose the best basis in WPT computation. Then, individual compressed IR spectra were employed to set up a small scale spectral library for testing the efficiency of our approach in improving the spectral library search. The results thus obtained are compared with those from the fast Fourier transform (FFT) treatment.

2. Method of investigation

In this work, two WT techniques, namely FWT and WPT, were employed to process a selected set of IR spectra. After the FWT or WPT treatment, the coefficients obtained were compressed by using appropriate methods to reduce the storage size. Selected coefficients were utilized to construct a wavelet compressed spectral library for future use. The library consists of compressed IR data in the wavelet domain, compound names

of the compressed spectra, and selected parameters such as the types of wavelet function used, the original data length and the assigned resolution level J for spectral reconstruction. The scale coefficients of each compound obtained at level J in the spectral library were employed for the preliminary library search which was performed by comparing the scale coefficients of the unknown spectrum with the reference spectra at a particular resolution level via the direct matching method. A small group of IR spectra were thus extracted from the library. Then, a detail searching was performed by comparing the unknown spectrum and reference spectra at resolution level 0. This search may not be required for IR spectra with very different spectral structures. However, compounds with similar molecular structures as that adopted in this work gives IR spectra with minor differences which are not easy to distinguish using the scale coefficients obtained at higher J levels. Therefore, a second search in the original domain was also carried out based on spectral information obtained at resolution level 0. In order to solve the side-lobe problem in data reconstruction, the translation–rotation transformation (TRT) method, as used in our previous works [15,56], was also applied. Besides, the coefficient position retaining (CPR) method has also been developed to handle spectrum with data length in odd number. Details of the above mentioned methods will be discussed in the following sub-sections.

2.1. Fast wavelet transform

Wavelet transform is a tool that can be utilized to convert data, functions or operators into different frequency components. Then each component is studied with a resolution matched to its scale [64]. There are different types of wavelet functions available in the literature such as Haar wavelet, Meyer's wavelet, Mexican hat wavelet, spline wavelet, and Daubechies wavelets [65]. The later one is now the industry standard for signal compression especially in chemical studies [66]. The fast wavelet transform algorithm was developed by Daubechies [67,68]. She adopted the multi-resolution signal decomposition (MRSD) algorithm, which was developed by Mallat [69,70], to construct families of compact supported wavelets and coupled them to quadrature mirror filtering (QMF). Details of the wavelet theory will not be stated here and can be found in Refs. 67 to 70.

In FWT, a signal at resolution level 0, $P_0 f(\lambda)$, is represented by a sum of the data obtained at different resolution levels of the original spectrum through the following formula [67,70]:

$$P_0 f(\lambda) = \sum_k c_k^{(J)} \sqrt{2^{(J)}}\, \phi_{J,k}(\lambda) + \sum_{j=1}^{J} \sum_k d_k^{(j)} \sqrt{2^j}\, \Psi_{j,k}(\lambda) \tag{1}$$

with $\phi_{j,k}(\lambda)$ and $\Psi_{j,k}(\lambda)$ representing the scaling and wavelet function respectively, and J the highest resolution level assigned in the WT calculation. In the above expression, $C^{(j)}(=c_k^{(j)})$ and $D^{(j)}(=d_k^{(j)})$ are the scale and wavelet coefficients at the jth resolution level, respectively, and $C^{(0)}$ represents the original signal in the discrete form. A signal is usually transformed by a high-pass filter G and a low-pass filter H and is represented, respectively, by a series of scale $C^{(j)}$ and wavelet $D^{(j)}$ coefficients at the jth resolution level. The scale coefficients $C^{(j)}$ denote the approximation of the raw signal $C^{(0)}$ with a resolution of one point per every 2^J point of the original one. The wavelet coefficients $D^{(j)}$ represent details of the original signal at different resolution levels [71]. These quantities can be deduced through the following formulae [20].

$$c_k^{(j)} = \sqrt{2} \sum_n c_n^{(j-1)} h_{n-2k} \tag{2}$$

and

$$d_k^{(j)} = \sqrt{2} \sum_n c_n^{(j-1)} g_{n-2k} \tag{3}$$

with $n = -\infty$ to $+\infty$ and k being a running index which has a variable length depending on the type of wavelet filter and the data length used. The variables h_k and g_k in these equations denote the coefficients of the low-pass and high-pass filters, respectively, with the following properties:

$$g_k = (-1)^k h_{1-k} \tag{4}$$

with

$$\sum_k h_k = 1 \text{ and } \sum_k g_k = 0. \tag{5}$$

In this work, coefficients for filters G and H were derived from the Daubechies wavelet, D_{2m}, with m being any positive integer. Details of the WT calculation can be found in references by Borde [66] and Daubechies [64].

In the general case, FWT is applied to a signal with a length of $N(=2^P)$ where P equals to any positive integer [70]. Suppose a spectrum is expressed in the digital form as $C^{(0)} = \{c_0^{(0)}, c_1^{(0)}, \ldots, c_{N-1}^{(0)}\}$. In FWT treatment, $C^{(0)}$ is first extended periodically on both sides. Then, the scale and wavelet coefficients at the $j = 1$ resolution level are determined via Eqs. (2) and (3) respectively. The numbers of elements of $C^{(1)}$ and $D^{(1)}$ are the same and are equal to $N/2$. Then, the same decomposition process as mentioned above is applied to $C^{(1)}$ again to obtain the required coefficients at the next resolution level. The process is repeated until the desired Jth resolution level is reached. Finally, the original spectrum is expressed as a collection of the scale and wavelet coeffi-

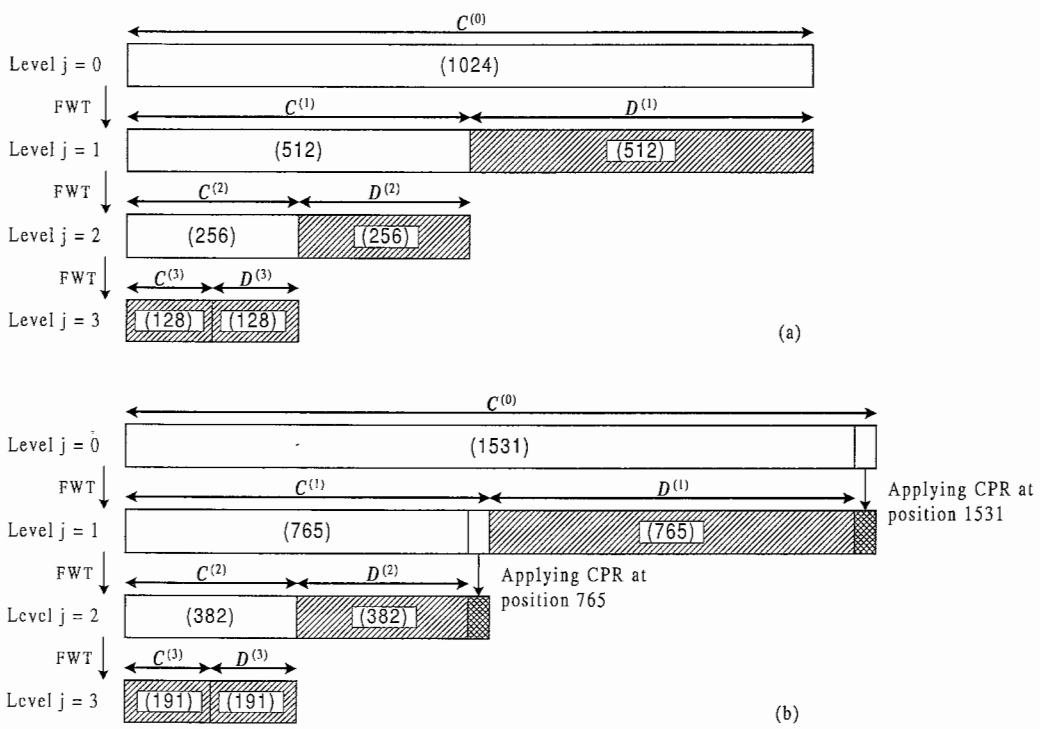

Fig. 1. A schematic diagram showing the operation of the FWT method with data lengths of (a) $N = 1024$ and (b) $N = 1531$ coupled with CPR treatment. The slanting line represents coefficients to be stored and the cross line shows the position of the coefficient(s) to be archived in using the CPR method.

cients in the form of $\{C^{(J)}, D^{(J)}, D^{(J-1)}, \ldots, D^{(1)}\}$. The total length or number of coefficients must be equal to the length of the original spectrum. Fig. 1a shows a schematic diagram for the FWT treatment with data length equals to 1024 ($=2^{10}$).

2.2. Wavelet packet transform

Wavelet packet transform is a derivative of WT that was developed by Coifman et al. [71]. The discrete WT is generalized in the WPT treatment to provide a more flexible tool for data analysis [72]. In FWT, a partial multi-resolution analysis is performed. Only $C^{(j)}$ is employed to deduce both scale and wavelet coefficients at the next resolution level. However, WPT allows a full multi-resolution analysis. $D^{(j)}$ is also involved to produce its scale and wavelet coefficients at the same time. As a result, a library of orthonormal bases is obtained. Fig. 2a shows a schematic diagram for the WPT operation with a data length of 2^P. In this figure, the original spectrum can be expressed as a suitable combination of bases to form a wavelet packet table. For examples, one possible combination of the bases subset to represent the original spectrum is $\{C^{(3,1)}, D^{(3,1)}, D^{(2,1)}, C^{(2,2)}, C^{(3,4)}, D^{(3,4)}\}$. Another possible combination of the bases subset is $\{C^{(3,1)}, D^{(3,1)}, D^{(2,1)}, C^{(3,3)}, D^{(3,3)}, D^{(2,2)}\}$. Again, the total number of all these coefficients must be equal to that of the original spectrum.

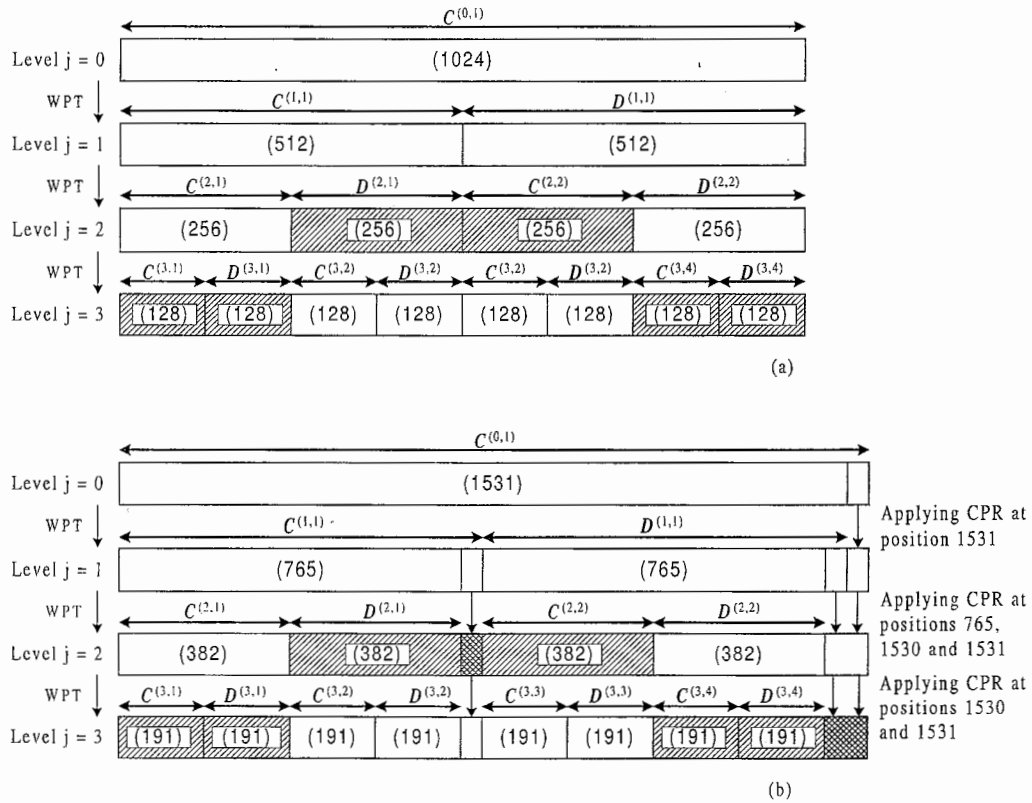

Fig. 2. A schematic diagram depicting the operation of the WPT method with data lengths of (a) $N=1024$ and (b) $N=1531$ coupled with CPR treatment. The slanting line represents coefficients to be stored and the cross line indicates the position of the coefficient(s) to be archived in using the CPR method.

In order to choose the best basis subset that represents the original data in the most effective way from a huge number of possible combinations of bases, the Shannon–Weaver entropy method was employed in this work to search the best basis [19,73]. The Shannon–Weaver entropy of a sequence $x = \{x_j\}$ can be expressed as:

$$H_{SW}(x) = -\sum_j q_j \log q_j \tag{6}$$

where $q_j = |x_j|^2 / \|x_j\|^2$. The symbols $|x_j|$ and $\|x_j\|$ represent the absolute value and root-mean-square norm of x_j respectively. Since Eq. (6) does not obey the theorem of additive measure of information [71], Eq. (7) instead of Eq. (6) is utilized for entropy calculation..

$$\lambda_{SW}(x) = -\sum_j |x_j|^2 \log |x_j|^2. \tag{7}$$

Once the wavelet packet table is set up, the entropy of each basis is determined by Eq. (7). Then, a comparison of the entropy values between two adjacent levels is performed in the following manner for the selection of the best basis. For example, in Fig. 2a, if the total sum of entropy of $\lambda_{SW}(C^{(3,1)})$ and $\lambda_{SW}(D^{(3,1)})$ is less than their parent $\lambda_{SW}(C^{(2,1)})$, then both $C^{(3,1)}$ and $D^{(3,1)}$ are chosen as part of the best basis. On the other hand, if the total sum of entropy of $\lambda_{SW}(C^{(3,2)})$ and $\lambda_{SW}(D^{(3,2)})$ is greater than their parent $\lambda_{SW}(D^{(2,1)})$, then $D^{(2,1)}$ is selected as another part of the best basis. This comparison process is repeated from resolution level $(J-1)$ to 1. This selection process is called the best basis method [74].

2.3. Translation–rotation transformation method

In order to apply MRSD to FWT and WPT calculations, the spectral data vector $C^{(0)}$ needs to be extended periodically at the two extremes. However, if $c_0^{(0)}$ and $c_{N-1}^{(0)}$ at the two extremes do not have the same value, a small delay which is due to discontinuity of the spectral data at the boundary will be observed at both ends of the reconstructed IR spectrum. Such phenomenon is known as side-lobe problem and deteriorates the quality of the reconstructed data [56]. To solve such a problem, the translation–rotation transformation (TRT) method was adopted [75]. The TRT algorithm involves subtraction of the data vector $C^{(0)}$ by selected quantities B to give the rotated array by

$$c_{k,\mathrm{TRT}}^{(0)} = c_k^{(0)} - b_k \tag{8}$$

with

$$b_k = c_0^{(0)} + \frac{\left(c_{N-1}^{(0)} - c_0^{(0)}\right)k}{N-1}. \tag{9}$$

2.4. Coefficient position retaining method

In WT treatment, the data length for a basis to be processed at resolution level j must be an even number. If an odd number data set is encountered at a particular resolution level, the WT calculation will be stopped automatically and cannot be processed to the next higher resolution level. It is because the data length of the scale and wavelet coefficients must be the same after WT treatment. Hence, the number of spectral data must equal to 2^P where P equals to any positive integer. In Fig. 1a, if the original spectral data length is equal to $1024 (= 2^{10})$, the data length of all bases at each resolution level j can be guaranteed to be an even number. In practice, it is not easy for a chemical instrument to generate 2^P data exactly. To cope with this problem, a series of zeros is usually appended to one end or both ends of the original data set in order to bring the total length to the next power of 2. This method is called zero padding method and is widely used in fast Fourier transform calculation [76]. Besides, truncation of data at the end or both ends of the original data to the previous power of 2 can also be adopted in some cases.

A new method, called Coefficient Position Retaining (CPR) method, was developed in this work to process spectrum with odd number of data in the FWT and WPT calculations. In this approach, if the data length $N_{c,j}$ of

a scale coefficient $C^{(j)}$ is an even number, FWT is applied as usual by using Eqs. (5) and (6). The scale and wavelet coefficients obtained at resolution level $(j+1)$ will have the same number of coefficients $N_{c,j+1}(= N_{c,j}/2)$ and $N_{d,j+1}(= N_{c,j}/2)$ respectively. On the other hand, if $N_{c,j}$ is an odd number, FWT is adopted without using the last coefficient of $C^{(j)}$ in the calculation. This coefficient is retained and transferred downward to the same position at the next resolution level. Then, it becomes the last coefficient of $D^{(j+1)}$ at the next resolution level. As a result, the scale and wavelet coefficients will have $N_{j+1,c}(= (N_{j,c}-1)/2)$ and $N_{j+1,d}(= ((N_{j,c}-1)/2)+1)$ elements respectively. Fig. 1b shows a schematic diagram for applying FWT to a spectrum with 1531 data with the use of CPR.

In WPT calculation, the scale coefficient $C^{(j)}$ with odd number of data length is handled in the same way as that for FWT as described above. However, for the wavelet coefficient $D^{(j)}$, the treatment is slightly different. Again, the last coefficient in $D^{(j)}$ is retained and transferred downward to all resolution levels that is below the present one at the same position. All these coefficients selected are not involved in FWT calculation. In WPT, the wavelet coefficient $D^{(j)}$ may consist more than one coefficients rather than only one coefficient retained as in FWT treatment. Fig. 2b shows a schematic diagram for the WPT calculation on a spectrum with 1531 data with CPR treatment. In the first cycle of WPT computation, the last coefficient in $C^{(0,1)}$ is transferred to position 1531 from levels 2 to 4. This coefficient is not involved in the WPT computation for $D^{(1,1)}$ and $D^{(2,2)}$. In the second cycle, since the data length of $D^{(1,1)}$ is an odd number, the last coefficient of $D^{(1,1)}$ at position 1530 is retained and transferred to position 1530 from levels 3 to 4. As a result, $D^{(2,2)}$ contains two retaining coefficients. Again, these coefficients are not involved in the WPT calculation for $D^{(2,2)}$. In general, the retaining coefficients are not considered in WT calculation.

2.5. Criteria for data compression

After FWT or WPT processing, the total data length of all C and D coefficients does not change. To reduce storage space, some of these coefficients must be discarded. Several methods have been proposed to select which coefficients in an optimal basis are negligible. Those adopted in this work are the absolute cutoff, relative energy and entropy criterion methods [71] and are described as follows. The absolute cutoff method is the simplest one. A cutoff value ε with magnitude greater than zero is assigned. Then, the absolute value of any coefficient c or d with a value less than is rejected. In the relative energy method, a cutoff value is defined in the range $0 < \varepsilon \leq 1$. Any coefficient in the absolute-square form ($|c|^2$ or $|d|^2$) that is less than $\varepsilon\|x\|^2$ is discarded. In the entropy criterion method, an average energy of a significant coefficient is defined as $\exp(-\lambda_{SW}(x)/\|x\|^2)$. Also, the cutoff value is defined in the range of $0 < \varepsilon \leq 1$. Any coefficient in the absolute-square form that has value less than that of $\varepsilon \exp(-\lambda_{SW}(x)/\|x\|^2)$ is neglected.

In both FWT and WPT treatment, the transformed IR spectra are represented by the scale coefficients $C^{(j)}$ in FWT and $C^{(j,1)}$ in WPT at the required resolution level j. As usual, one cannot remove part of these coefficients for the compression purpose because this will cause loss of useful information and will lead to serious error during signal reconstruction. So no compression were performed on the scale coefficients. As a result, the proposed compression method was only applied to the wavelet coefficients.

2.6. Direct matching method for library search

Direct matching method is the simplest way for spectral library search. It employs a point-by-point comparison calculation to match sequentially an unknown spectrum against the reference spectra available in a library [73]. Both the unknown and reference spectra are represented as points in a multidimensional space in a fixed range of wavelength. The degree of similarity between two spectra is measured by the sum of separations between data points. Six different search routines have been proposed for the purpose. They include absolute difference D_{absdiff} [5], absolute derivative D_{absder}, square difference D_{sqrdiff}, square derivative D_{sqrder}, correlation

coefficient D_{corr} and Euclidean distance D_{edist}. Mathematical expressions of these routines are given as follows [60,73,77]:

$$D_{absdiff} = \sum_{i=1}^{N} |x_i - y_i| \tag{10}$$

$$D_{absder} = \sum_{i=1}^{N-1} |(x_i - x_{i+1}) - (y_i - y_{i+1})| \tag{11}$$

$$D_{sqrdiff} = \sum_{i=1}^{N} (x_i - y_i)^2 \tag{12}$$

$$D_{sqrder} = \sum_{i=1}^{N-1} [(x_i - x_{i+1}) - (y_i - y_{i+1})]^2 \tag{13}$$

$$D_{corr} = \frac{N\sum_{i=1}^{N} x_i y_i - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)}{\sqrt{\left[N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2\right]\left[N\sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2\right]}} \tag{14}$$

and

$$D_{Edist} = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2}\,. \tag{15}$$

In these expressions, x_i and y_i denote, respectively, the data points in the reference and unknown spectra in a particular domain. N represents the total number of data points in the compressed or uncompressed spectrum. A perfect match of any two spectra would give one by using correlation coefficient method (Eq. (14)) and a zero value by using the other methods. A lower D_{corr} value or a higher D value for the other methods indicates higher dissimilarity between the spectra. Both the absolute derivative and square derivative routines were applied when the base-line offset was observed. These two routines can discriminate the correct spectrum from the other spectra with similar D value in the library. However, if the unknown spectrum is not available in the library, they may give poor match values for spectra that have similar spectral features. In this study, D_{corr}, $D_{absdiff}$ and D_{absder} were employed in the spectral search.

Some workers did not carry out the direct matching method in the original domain because it involves large number of calculation. They suggested to transform the IR spectrum by Fourier transform before spectrum matching [78]. In this investigation, both FWT and WPT were applied to treat IR spectra for setting up the library. These two methods provide a way to minimize the search time for direct matching of spectra because it involves only minimal number of coefficients represented in the wavelet domain instead of using the whole spectrum.

3. Experimental

Twenty organic compounds of AR grade (see Tables 2 and 3) as utilized in this work were obtained from Aldrich (Gillingham, UK) and were used without further purification. Their IR spectra were recorded by using the Nicolet Magna-IR™ Model 750 Fourier transform infrared spectrometer (Nicolet Instrument, WI, US) in the range of 400 to 4000 cm^{-1} in percentage of transmittance at 4 cm^{-1} resolution with 16 times of scan and Happ–Genzel apodization. In this setting, each IR spectrum contains 1868 data points. ASCII data were ex-

ported from the Nicolet ONMIC FT-IR software which was used to control the FT-IR spectrometer. Solid samples were recorded by using KBr pellet method while liquid samples were recorded by liquid film method using NaCl plates. The organic compounds were chosen purposely with similar structures or isomers of one another and their IR spectra are quite similar. The data set comprises the IR spectra of the mono-, di- and tri-substituted benzene with Cl, Br, I, NO_2, OH, CH_3 and CH_2Cl as substituents. All IR spectra were background corrected and normalized to unit transmittance first (Eq. (16)) before further treatment through the following formula:

$$c_i = \frac{A_i}{A_{\max}} \tag{16}$$

where c_i and A_i are the fraction of total absorbance and the absorbance at wavelength i, respectively, and $A_{\max}$ is the maximum absorbance in the spectrum.

All WT computations were carried out using the Spectral Library Compression and Search (SLCS) Toolbox Version 1.0 as developed in this study. It was coded in MATLAB® for Windows Version 4.2c [79] under Microsoft® Windows™ 95 environment on a PC compatible with a 133MHz Pentium® processor. The toolbox can be employed to optimize the required parameters for FWT and WPT calculation as well as to perform spectral library compression, construction and reconstruction, and spectral library searching.

4. Results and discussion

4.1. Choice of parameters for FWT and WPT calculations

In both FWT and WPT calculations, the Daubechies wavelet function, D_{2m} resolution level J, and cutoff value ε have to be selected to give optimal performance. In this work, the compression ratio R_{comp} (Eq. (17)) measures the compression efficiency of the selected method while D_{corr} (Eq. (14)) measures the similarity between the original and reconstructed IR spectrum.

$$R_{\text{comp}} = \frac{\text{No. of bytes of the original data} - \text{No. of bytes of the compressed data}}{\text{No. of bytes of the original data}} \times 100\% \tag{17}$$

The best performance is considered when R_{comp} and D_{corr} have values close to 100% and 1 respectively. In computing R_{comp} for data obtained from WT and FT methods, the file sizes of the compressed data were used.

Table 1
Results of applying the proposed FWT scheme to compress IR spectrum of benzoic acid with the use of different cutoff values ε and Daubechies wavelets function D_{2m} at resolution levels of $J = 3$, 4 and 5

	$J=3$		$J=4$						$J=5$	
	$\varepsilon=0.2$		$\varepsilon=0.1$		$\varepsilon=0.2$		$\varepsilon=0.3$		$\varepsilon=0.2$	
	R_{comp}	D_{corr}	R_{comp}	D_{corr}	R_{comp}	D_{corr}	R_{comp}	D_{corr}	R_{comp}	D_{corr}
D_2	86.51	0.9927	87.37	0.9968	91.38	0.9905	92.61	0.9844	93.42	0.9892
D_4	86.62	0.9968	90.42	0.9980	91.92	0.9955	92.83	0.9912	94.27	0.9943
D_6	86.51	0.9975	90.58	0.9984	91.81	0.9962	92.88	0.9909	94.00	0.9956
D_8	86.78	0.9969	90.26	0.9986	92.24	0.9956	92.99	0.9917	94.33	0.9950
D_{10}	86.83	0.9972	90.69	0.9984	92.08	0.9958	92.67	0.9931	94.27	0.9949
D_{12}	86.94	0.9964	90.15	0.9986	92.13	0.9955	92.72	0.9929	94.27	0.9945
D_{14}	86.67	0.9975	90.47	0.9985	92.02	0.9962	92.88	0.9921	94.16	0.9953
D_{16}	86.83	0.9972	90.96	0.9983	92.34	0.9958	92.56	0.9943	94.27	0.9952
D_{18}	86.83	0.9966	90.15	0.9983	92.08	0.9952	92.88	0.9913	93.90	0.9949
D_{20}	86.88	0.9966	90.10	0.9984	92.08	0.9954	92.83	0.9920	93.90	0.9949

Table 2
Results of applying the proposed FWT and WPT methods to compress IR spectra of the twenty compounds under study with cutoff value $\varepsilon = 0.20$ at and $J = 4$ by using the Daubechies wavelet functions D_{14} and D_{16}

Compound name	FWT				WPT			
	D_{14}		D_{16}		D_{14}		D_{16}	
	R_{comp}	D_{corr}	R_{comp}	D_{corr}	R_{comp}	D_{corr}	R_{comp}	D_{corr}
1,4-Dichlorobenzene	92.99	0.9759	93.09	0.9730	93.04	0.9758	93.09	0.9750
1-Bromobutane	93.63	0.9958	93.52	0.9967	93.68	0.9958	93.58	0.9966
1-Chloro-2,4-dinitrobenzene	93.74	0.9692	93.79	0.9636	93.79	0.9672	93.79	0.9636
1-Chlorobutane	92.93	0.9965	93.20	0.9957	92.99	0.9967	93.25	0.9962
1-Iodobutane	92.18	0.9975	92.51	0.9970	92.13	0.9980	92.45	0.9978
2,4-Dichlorophenol	93.04	0.9739	92.93	0.9769	93.09	0.9738	92.99	0.9769
2-Chlorobutane	92.24	0.9963	92.24	0.9963	92.45	0.9960	92.51	0.9961
2-Chlorophenol	92.02	0.9975	92.08	0.9972	92.29	0.9974	92.18	0.9973
2-Nitrophenol	92.40	0.9823	92.61	0.9825	92.40	0.9823	92.67	0.9814
3-Chlorophenol	92.88	0.9982	93.09	0.9983	92.93	0.9983	93.15	0.9984
3-Nitrophenol	93.20	0.9838	93.09	0.9866	93.25	0.9853	93.09	0.9866
4-Nitrobenzyl chloride	93.20	0.9720	92.93	0.9693	93.25	0.9716	92.99	0.9700
4-Nitrophenol	91.49	0.9957	92.02	0.9944	91.54	0.9957	92.02	0.9946
4-Nitrotulene	92.45	0.9865	92.08	0.9869	92.45	0.9871	92.40	0.9859
Benzoic acid	92.02	0.9962	92.34	0.9958	92.13	0.9963	92.45	0.9959
Bromobenzene	92.08	0.9954	92.08	0.9941	92.40	0.9951	92.45	0.9946
Cyclohexyl bromide	91.70	0.9962	91.97	0.9943	91.18	0.9961	92.08	0.9948
Cyclopentyl chloride	93.52	0.9976	93.42	0.9979	93.52	0.9978	93.42	0.9982
Iodobenzene	91.92	0.9943	91.92	0.9939	92.02	0.99 43	91.92	0.9945
Toluene	92.67	0.9956	92.72	0.9957	92.72	0.9960	92.88	0.9958
Average	92.62	0.9898	92.72	0.9893	92.70	0.9780	92.77	0.9895
Standard derivation	±0.66	±0.0100	±0.63	±0.0106	±0.69	±0.0102	±0.54	±0.0106

Table 1 shows the results of compressing the IR spectrum of benzoic acid with different ε values and D_{2m} functions at resolution levels of $J = 3$, 4 and 5 by utilizing FWT. As the value of J increases, R_{comp} increases with a slight decrease in D_{corr} at a fixed cutoff value. The same trend is observed when both J and D_{2m} are fixed with different cutoff values. At a fixed resolution level and cutoff value, a maximum values of R_{comp} and D_{corr} have been found by employing the Daubechies wavelet functions D_{14} and D_{16}. Table 2 show results of compressing IR spectra of the twenty compounds under study with the proposed FWT and WPT treatments with $J = 4$, $\varepsilon = 0.20$, and using the D_{14} and D_{16} functions. From the result of this investigation, we recommend to use the D_{16} wavelet with $\varepsilon = 0.20$ and $J = 4$ to compress the IR spectrum in the range of 400–4000 cm^{-1} for data storage by FWT and WPT. Fig. 3 shows the reconstructed IR spectra from the compressed data as obtained from both FWT and WPT treatments on the IR spectrum of benzoic acid with, D_{16}, $\varepsilon = 0.20$ and $J = 4$.

4.2. Comparison of performance of FWT and WPT with that of fast fourier transform

The performance of FWT and WPT in spectral compression were compared with that by using FFT. R_{comp} and D_{corr} were chosen as the key parameters for comparison. As mentioned before, FFT and its inverse are the major techniques adopted currently to compress spectral data in chemical studies. Three methods namely the absolute cutoff, least-square-fitting and power spectrum methods coupled with FFT were employed for selecting Fourier coefficients. The chosen Fourier coefficients represent the compressed form of the original spectral data. Details of the FFT calculation and the three methods can be found in the work by Chau and Tam [15]. Table 3 shows results of compression for IR spectra of the twenty compounds as obtained by FFT coupled with the compression methods mentioned above. Although the FFT treatment using the least-square-fitting and power

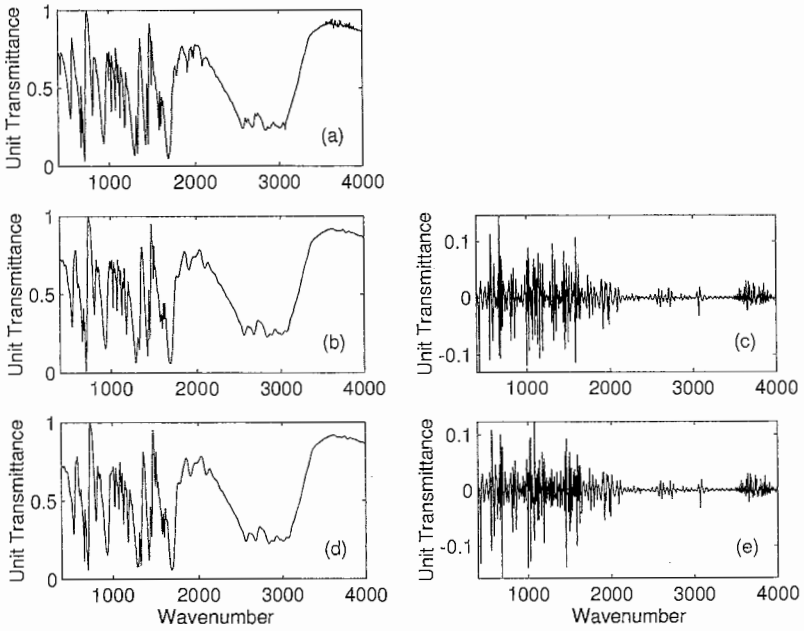

Fig. 3. The experimental IR spectrum (a) and reconstructed IR spectra of benzoic acid with the use of compressed data as obtained from FWT (b) and WPD (d) by utilizing D_{16} the function, $\varepsilon = 0.20$, and $J = 4$ with the corresponding error plot from FWT (c) and WPD (e).

spectrum methods give very high D_{corr} values, the compression ratio R_{comp} is very low. Spectral compression by FFT coupled with the absolute cutoff method ($\varepsilon = 0.10$) can attain the same average compression ratio as FWT and WPT (Table 2). But, the compression ratios for individual spectra fluctuate in a wide range by using FFT in the range of 85.87% (2,4-dichlorophenol) to 97.22% (1-bromobutane). In this investigation, our aim is to set up a good spectral library for IR spectra. It is desirable to have any method that can provide highly stable compression ratio and minimum distortion in the reconstructed spectra. From the results of Tables 2 and 3, it is clear that the performance of our proposed FWT and WPT methods in spectral compression is better than that of FFT.

In terms of computational time, the FWT calculation requires shorter time than that of FFT. A FFT computation requires $N \log_2 N$ operations while a single cycle of FWT require only N operations for a data length of $N(=2^P)$ [72,80]. In FWT, it is expected that the computation time increases when a higher J resolution level is required. Since FFT was specially designed for data length equal to 2^P, a slower algorithm is employed for FT calculation if the data length is not equal to 2^P. For a spectrum with data length equal to 2^P, FFT method is faster than FWT method if similar compression ratio is attained. For example, in compressing an IR spectrum with $N = 1024$ with similar compression ratio, FFT method requires 10,240 multiplication operations while FWT method needs 30,720 (from Eq. (18)) multiplication operations (up to the 4th resolution level with the D_{16} function ($L = 16$)) with L being the length of the filters H and G. But, the computational time would be close for the FFT and FWT methods if the data length of a spectrum is not equal to 2^P. To investigate the computational times for FWT and WPT, the following algorithm is adopted. The total number of multiplication operations after J iterations in FWT treatment given as follows [20].

$$L \times N + L \times N/2 + \ldots + L \times N/2^J = 2L \times N \times (1 - 1/2^J). \tag{18}$$

In WPT calculation, the total number of multiplication operations after J operations is.

$$L \times N + 2L \times N/2 + \ldots + 2^J L \times N/2^J = JL \times N. \tag{19}$$

Table 3
Results of applying the FFT method to compress IR spectra of the twenty compounds[a] under study by using the absolute cutoff, the least-square-fitting, and power spectrum methods

Compound name	Absolute cutoff method		Least-square-fitting method		Power spectrum method	
	R_{comp}	D_{corr}	R_{comp}	D_{corr}	R_{comp}	D_{corr}
1,4-Dichlorobenzene	90.26	0.9903	75.05	0.9997	52.14	1.0000
1-Bromobutane	97.22	0.9907	75.05	1.0000	53.69	1.0000
1-Chloro-2,4-dinitrobenzene	86.08	0.9976	77.84	0.9995	56.42	1.0000
1-Chlorobutane	93.15	0.9981	76.07	1.0000	54.76	1.0000
1-Iodobutane	95.61	0.9917	77.09	1.0000	53.21	1.0000
2,4-Dichlorophenol	85.87	0.9976	75.05	0.9995	57.98	1.0000
2-Chlorobutane	92.24	0.9975	75.05	1.0000	51.28	1.0000
2-Chlorophenol	89.72	0.9993	75.05	1.0000	52.41	1.0000
2-Nitrophenol	92.93	0.9837	75.05	0.9998	53.69	1.0000
3-Chlorophenol	93.52	0.9988	75.05	1.0000	54.01	1.0000
3-Nitrophenol	88.92	0.9975	76.93	0.9998	61.46	1.0000
4-Nitrobenzyl chloride	89.61	0.9926	75.05	0.9997	52.52	1.0000
4-Nitrophenol	87.10	0.9988	77.68	0.9999	53.16	1.0000
4-Nitrotulene	93.20	0.9819	76.39	0.9994	51.66	1.0000
Benzoic acid	96.84	0.9788	75.54	0.9999	57.07	1.0000
Bromobenzene	92.51	0.9903	75.05	0.9999	52.62	1.0000
Cyclohexyl bromide	93.25	0.9906	75.05	0.9999	52.89	1.0000
Cyclopentyl chloride	96.20	0.9963	77.57	1.0000	52.52	1.0000
Iodobenzene	90.04	0.9926	75.70	0.9999	50.70	1.0000
Toluene	95.77	0.9831	76.70	0.9999	52.52	1.0000
Average	92.00	0.9924	75.87	0.9998	53.84	1.0000
Standard derivation	±3.46	±0.0063	±1.04	±0.0002	±2.59	±0.0000

[a]The number of bytes of the original data was used in computing R_{comp}.

For example, to compress an IR spectrum with $N = 1868$ up to $J = 4$ using D_{16} function ($L = 16$), FWT and WPT require about 56,040 and 119,552 multiplication operations respectively. Hence, the computational time needed for FWT is about 47% shorter than that for WPT calculation.

4.3. Performance of the coefficient position retaining method

The data length of each IR spectrum recorded from our the FTIR instrument is 1868 which is not equal to power of 2. The FWT calculation will stop at the 2nd resolution level due to an odd number of data length being encountered. It is not possible to discard some data points at either end or both ends of the original spectral data to meet the 2^P requirement. It is because a significant amount of useful information will be lost from the IR spectrum in removing data. If the zero padding method is utilized to the extend data length to the next power of 2, side-lobe problem will occur in the reconstructed spectrum. The problem occurs when the periodic spectral signal has an abrupt level change within a short interval at the point of zero padding (Fig. 4a) [81]. Fig. 4b shows result of the reconstructed IR spectrum of benzoic acid with FWT coupled with zero padding method. It should be noted that the range of the two sets of spectra (a), (c), (e) and (b), (d), (f) are not the same. It is obvious that side lobes are produced at that position in the reconstructed spectrum. These lobes are also observed in the reconstructed IR spectrum when the spectral signal is extended without TRT treatment (Fig. 4c–d). Hence, zero padding method is not a good method to process spectral data via WT when the data length is not equal to the power of 2.

In order to achieve a WT calculation with higher J level, the CPR method was proposed and developed in this study to solve the problem associated with the data set with an odd data number in WT calculation. By

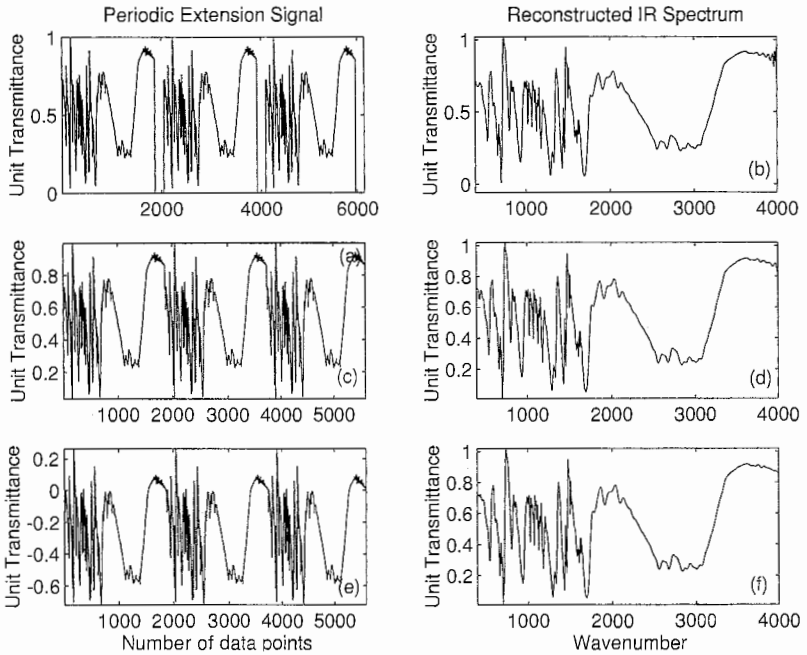

Fig. 4. Periodical extension signal of the IR spectrum of benzoic acid as generated by using the zero padding method (a), the CPR method without TRT treatment (c), and the CPR-TRT method (e) with the corresponding reconstructed spectra (b), (d) and (f) that were produced from the compressed data utilizing D_{16}, $\varepsilon = 0.20$ and $J = 4$.

comparing the CPR method with the zero padding method, it has two major advantages. Firstly, the former one can process spectral data with any data length. If the length is an odd number, WT calculation can still be performed. The CPR treatment can guarantee that the total data length is unchanged and will not affect the quality of the reconstructed spectrum. It is because data points at the ends of an IR spectrum usually represent the signal background and do not contain any importance information. Secondly, the CPR method does not involve any modification in the original spectral data compared with the zero padding method. Therefore, it can generate a smooth periodic spectral signal for the WT-TRT treatment (Fig. 4e). Hence, the side-lobe problem can be solved.

4.4. Setting up of a small size spectral library

After manipulating individual IR spectra through FWT and WPT, the compressed data can be used to set up a spectral library for future searching. The spectral library as generated by the SLCS software as developed in this work consists of four major types of data. They include the intensity data of a spectrum which is a combination of the suitable bases in the wavelet domain, the wavelength data of each intensity data point, the required data $c_0^{(0)}$ and $c_{N-1}^{(0)}$ for inverse TRT calculation and the compound name. If the spectral library is built up from WPT, extra space is needed for storing the best basis information for individual IR spectrum. Table 4 list the compressed file sizes, compression ratios and the computational times for setting up a spectral library containing IR spectra of the twenty selected compounds being processed by FWT, WPT and FFT. The results indicates that both FWT and WPT give higher compression ratios than that by FFT. The later method cannot achieve higher compression ratio because coefficients as obtained in the Fourier domain are complex numbers which require double amount of storage space. If a spectrum can be reduced with a high compression ratio by FFT, most of the high frequency signal will be removed and this will affect the quality of the reconstructed spectrum. However, the computational times for both FWT and WPT (Table 4) are longer than FFT because 4 and 15 WT

Table 4
Compression ratios and computational times for setting up a spectral library[a] with twenty IR spectra processed by FWT, WPT and FFT respectively

Compression methods	Compressed file size/KB	File compression ratio/%	Total compression time/s	Average compression time/s
FWT[b]	27.3	91.16	24.33	1.22
WPT[b]	30.0	90.30	160.22	8.01
FFT (absolute cutoff method)[c]	67.5	78.14	8.29	0.42
FFT (least square fitting method)[c]	117.2	62.05	9.89	0.49
FFT (power spectrum method)[c]	229.1	25.81	7.91	0.40

[a]The original spectral library size is 308.8 KB in the binary format.
[b]Calculation with the use of $\varepsilon = 0.20$, $J = 4$ and D_{16} function.
[c]Calculations were carried out using the FFT built-in function under the MATLAB® environment which is much faster than that by using the script file. So, the FFT computational time will be expected to longer than the listed values if a script file is utilized. Both FWT and WPT were carrying out using the SLCS script files.

iterations were required to process each IR spectrum, respectively. Since only a single data treatment is required to establish a spectral library, the longer computational time is not a big problem. Moreover, in spectral library searching, coefficients represented in either Fourier or wavelet domain can be employed for identification of an unknown spectrum. In the wavelet domain, the IR spectrum can be represented by a smaller number of coefficients compared to that in the Fourier domain owing to the higher compression ratios using the proposed WT method. So, the searching time in the wavelet domain is anticipated to be shorter. Therefore, spectral compression using FWT or WPT are a good choice.

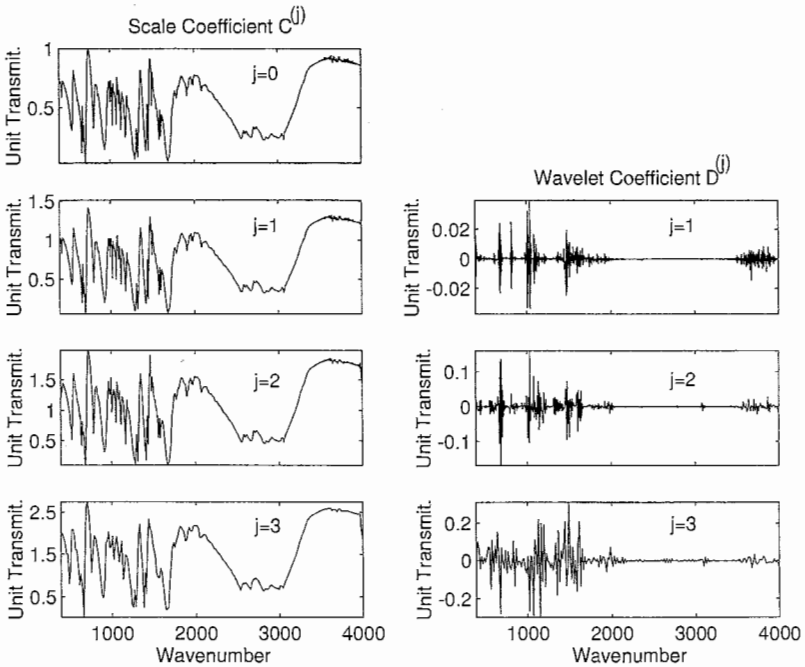

Fig. 5. Plots of the wavenumber against the scale and wavelet coefficients in unit transmittance as obtained from compressing the IR spectrum of benzoic acid at different resolution levels using the D_{16} function.

In compressing IR data, WPT performs better than FWT. In general, the IR spectrum usually has a lot of peaks in the fingerprint region. After a single cycle of WT calculation, a portion of the high frequency signals are preserved at the wavelet coefficients together with the noise (Fig. 5). In order to extract such high frequency signals, the noise present in D must be removed. In WPT treatment, WT calculation was performed on all bases so that the noise signal can be separated from the required high frequency data. By employing the Shannon–Weaver entropy calculation as mentioned previously, the best basis can be selected out to represent the original signal. However, in FWT treatment, no further WT calculation can be performed on $D^{(1)}$ (Fig. 1a) to extract the required signal. In such case, error would be observed in the reconstructed IR spectrum if a very high cutoff value is employed for noise removal and data compression treatment.

4.5. Spectral library searching

Spectral compression coupled with prefilter technique is the most common way for searching a large spectral library. By using this technique, a small group of spectra will be extracted from the library for further processing. For example, Lo and Brown [82] proposed to compress a spectral library by fast Fourier transform and adopted principal component regression for prefiltering. In this work, the IR spectrum was compressed by WT and represented by a smaller group of scale coefficients $C^{(J)}$ at the assigned resolution level J. Coefficients of $C^{(J)}$ represent the approximation of the original signal $C^{(0)}$ with a resolution of one point per every 2^J point of the original one. So, most of the characteristic peaks in the original IR spectrum can be retained in this approach. Fig. 6a and b depicts, respectively, the scale coefficients $C^{(J)}$ of the IR spectra of 4-nitrophenol and cyclophentyl chloride at various resolution levels. $C^{(J)}$ was utilized as a reference spectrum in the library searching process. In this investigation, each IR spectrum in the library was represented by 166 coefficients with the inclusion of 116 coefficients from $C^{(4)}$ and 50 coefficients from $D^{(1)}$ to $D^{(4)}$ after data compression. Up to 90% of the searching time can be saved for using the compressed spectra compared with that for the uncompressed spectra in the direct matching method.

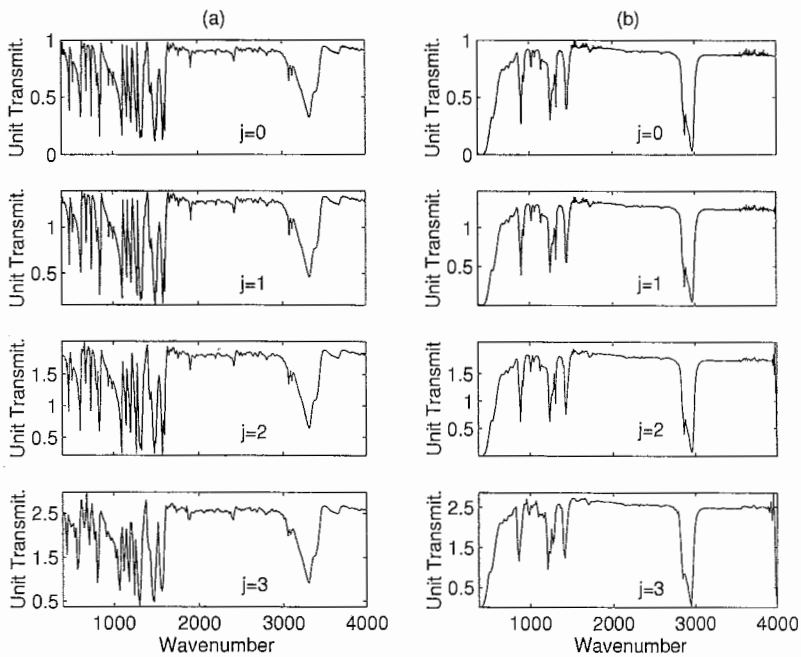

Fig. 6. A diagram to show the scale coefficient $C^{(J)}$ of the IR spectrum of 4-nitrophenol (a) and cyclophentyl chloride (b) as obtained from resolution levels 0 to 3 via FWT treatment using the D_{16} function and a cutoff value of $\varepsilon = 0.20$.

The IR spectrum of 4-nitrophenol was chosen as our unknown one to test the performance of the proposed method for library search. Tables 5 and 6 show the results of the preliminary and detail search. In the preliminary one, all reference spectra in the library are involved in the direct matching process. After this step, hit indices are generated in descending order to indicate the similarity between reference spectra in the library and the unknown spectrum. Hit index 1 will be assigned to the reference spectrum being found to have the highest D_{corr} and the lowest $D_{absdiff}$ and D_{absder} values from the direct matching calculations, and index 2 for the next highest D_{corr} and lowest $D_{absdiff}$ and D_{absder} values and so on. Thus, a lower hit index value indicates higher similarity between the reference and the unknown spectra. The number of selected IR spectra in the hit index can be adjusted according to the size of the spectral library. In this work, only five hit indices of 1 to 5 of the corresponding reference IR spectra are listed and used in the detail search. Table 5 shows the search results of the unknown IR spectrum of 4-nitrophenol of the FWT and WPT compressed libraries. All the three direct matching schemes could identify the unknown spectrum correctly. In the detail search, the suspected reference spectra are reconstructed from the compressed data up to level 0. At this level, the reconstructed spectrum does not exactly represent the spectrum in its original form, but, a smoothed version of the original spectrum. It is because part of the noise signal was removed in the WT compression process. The overall results for the preliminary and detail search are slightly different. It is because the resolution of the reconstructed spectra increases for lower resolution levels. Thus, more information is contained in the reconstructed spectrum. Since it is very time consuming to carry out detail search on a huge spectral library with spectra in their original form, the preliminary search provides a fast way to find out a group of reference spectra with high similarity. Then, the selected spectra are reconstructed via either inverse FWT or inverse WPT depending on the compression method used for further confirmation. By adopting this approach, an unknown IR spectrum can be identified within very short time compared to the conventional methods such as full spectrum searching.

Library search by using FFT was also performed in this work for comparison. Similar to that in WT treatment, the IR spectral library was first compressed by FFT with absolute cutoff, least-square-fitting and power spectrum methods [15]. Then, coefficients in the Fourier domain were employed for library searching. Table 7 gives the results of library search by using FFT. The results obtained indicates that coefficients from FFT coupled with absolute cutoff method is not good for spectral searching. Although the absolute cutoff method by discarding Fourier coefficients with values less that the cutoff value can compress a spectrum with very high compression ratio, some features are removed in the process. Of course the bad performance may also be due to the use of only the real part of the coefficients in the FFT domain for the two library searches. As a result, the compressed reference spectra do not provide enough information for identifying the unknown IR spectrum correctly. For the other two FFT methods, the search results are quite similar to those from FWT and WPT. But, spectral library treated with FWT and WPT is better than that by FFT especially in visual comparison and mem-

Table 5
Results of the preliminary library search[a] of the unknown IR spectrum (spectrum of 4-nitrophenol) from the library being set up using FWT and WPT compressed spectra (see text)

Hit index	Correlation coefficient method		Absolute difference method		Absolute derivative method	
	Compound identified	Value	Compound identified	Value	Compound identified	Value
1	4-nitrophenol	0.9698	4-nitrophenol	5.64	4-nitrophenol	2.24
2	3-nitrophenol	0.7453	3-nitrophenol	11.58	4-nitrotulene	9.64
3	4-nitrotulene	0.6164	1-cholor-2, 4-dinitrobenzene	14.36	2-chlorophenol	11.49
4	1-cholor-2, 4-dinitrobenzene	0.5712	4-nitrotulene	14.59	1-bromobutane	11.96
5	2-nitrophenol	0.5354	2-nitrophenol	15.72	1-cholor-2, 4-dinitrobenzene	12.31

[a]Both FWT and WPT methods give the same result.

Table 6
Results of the detail search of the unknown IR spectrum (spectrum of 4-nitrophenol) from the library being reconstructed at resolution level 0 (see text)

Method use	Hit index	Correlation coefficient method		Absolute difference method		Absolute derivative method	
		Compound identified	Value	Compound identified	Value	Compound identified	Value
FWT	1	4-nitrophenol	0.9639	4-nitrophenol	103.26	4-nitrophenol	15.88
	2	3-nitrophenol	0.6868	2-nitrophenol	210.95	1-cholor-2, 4- dinitrobenzene	23.93
	3	3-chlorophenol	0.5637	4-nitrotulene	211.84	3-nitrophenol	25.20
	4	4-nitrotulene	0.5611	3-nitrophenol	222.91	1-bromobutane	25.36
	5	1-cholor-2, 4-dinitrobenzene	0.5534	2,4-dichloro-phenol	252.68	4-nitrobenzyl chloride	25.47
WPT	1	4-nitrophenol	0.9643	4-nitrophenol	103.18	4-nitrophenol	15.66
	2	3-nitrophenol	0.6864	2-nitrophenol	211.21	1-cholor-2, 4- dinitrobenzene	23.93
	3	3-chlorophenol	0.5641	4-nitrotulene	211.66	3-nitrophenol	24.96
	4	4-nitrotulene	0.5600	3-nitrophenol	222.75	1-bromobutane	25.37
	5	1-cholor-2, 4-dinitrobenzene	0.5534	2,4-dichloro-phenol	252.64	4-nitrobenzyl chloride	25.55

ory storage. In some cases, the search routine cannot identify the unknown spectrum correctly for compounds selected in our library. Then, users need to compare the reference and unknown spectra by their own experience. One of the advantages in using both FWT and WPT methods for processing spectral library is that they can

Table 7
Results of the preliminary search of the unknown IR spectrum of 4-nitrophenol through the spectral library being compressed by FFT with the use of absolute cutoff, least-square-fitting and power spectrum methods

Method use	Hit index	Correlation coefficient method		Absolute difference method		Absolute derivative method	
		Compound identified	Value	Compound identified	Value	Compound identified	Value
Absolute cutoff method	1	4-nitrophenol	0.9373	1-cholor-2, 4-dinitrobenzene	856	1-cholor-2, 4-dinitrobenzene	979
	2	3-nitrophenol	0.8215	3-nitrophenol	898	3-nitrotulene	1059
	3	2-nitrophenol	0.7849	1,4-dichloro-benzene	952	1,4-dichloro-benzene	1125
	4	benzonic acid	0.6898	2-nitrophenol	1000	2-nitrophenol	1219
	5	3-chlorophenol	0.6393	2,4-dinitro-phenol	1016	2,4-dinitro-phenol	1244
Least square method	1	4-nitrophenol	0.9808	4-nitrophenol	585	4-nitrophenol	657
	2	3-nitrophenol	0.8160	3-nitrophenol	1700	3-nitrophenol	1991
	3	2-nitrophenol	0.7368	1-cholor-2, 4-dinitrobenzene	1818	4-nitrotulene	2070
	4	3-chlorophenol	0.5937	2,4-dinitro-phenol	1843	1-cholor-2, 4-dinitrobenzene	2121
	5	4-nitrotulene	0.5868	2-nitrophenol	1845	4-nitrobenzyl chloride	2163
Power spectrum method	1	4-nitrophenol	0.9808	4-nitrophenol	624	4-nitrophenol	672
	2	3-nitrophenol	0.8159	3-nitrophenol	1769	3-nitrophenol	2042
	3	2-nitrophenol	0.7367	1-cholor-2, 4-dinitrobenzene	1880	4-nitrotulene	2142
	4	3-chlorophenol	0.5937	2,4-dinitro-phenol	1904	1-cholor-2, 4-dinitrobenzene	2171
	5	4-nitrotulene	0.5865	2-nitrophenol	1907	4-nitrobenzyl chloride	2209

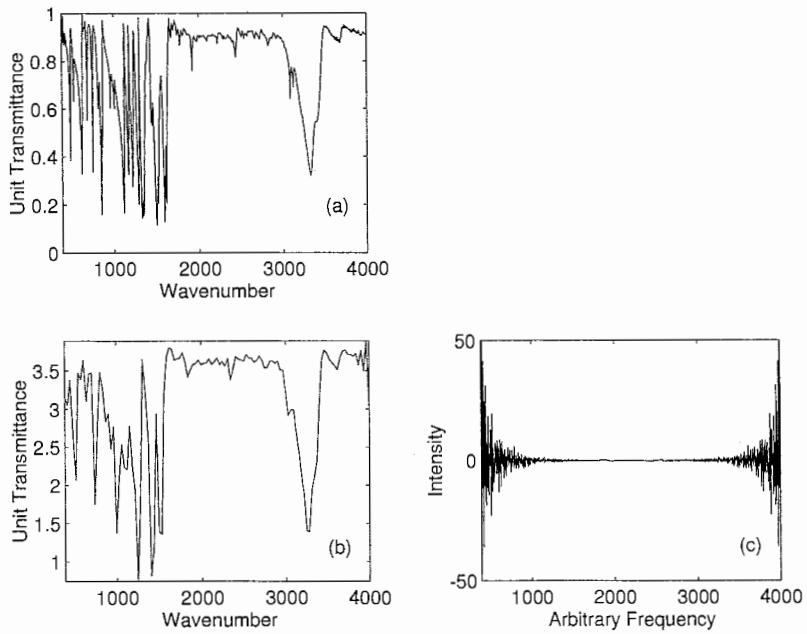

Fig. 7. (a) The IR spectrum of 4-nitrophenol, (b) plot of the scale coefficients as obtained from a FWT treatment on the IR spectrum of 4-nitrophenol using D_{16} at $J = 4$ against the wavenumber, and (c) plot of the Fourier coefficients of the IR spectrum of 4-nitrophenol as generated from FFT against frequency.

provide an easier way for visual comparison. Users can just make use of the scale coefficients at a particular resolution level to do so (e.g., Fig. 7a and b) because most of the characteristic peaks in the IR spectrum are still retained in the scale coefficients even in higher resolution levels. In FFT, only the real part of the coefficients are utilized for searching. Fig. 7c shows the real part of the FFT coefficients of the IR spectrum of 4-nitrophenol. These coefficients forms a frequency spectrum that consists of a lot of peaks. It is not easy to identify different functional groups directly by inspecting the frequency spectrum. Hence, the FFT coefficients are not suitable for visual comparison. In order to perform visual comparison via FFT coefficients, inverse FFT must be applied. With regard to our proposed methods, inverse FWT, inverse WPT and inverse CPR are not necessary in the preliminary search process while they are performed only in the detail search process. Besides, FFT calculation requires extra amount of memory space for data storage because all the FFT coefficients are stored in the form of complex number while WT coefficients in form of real number. Therefore, spectral library search is more efficient for a library that is set up by either FWT or WPT compression and is better than that using FFT compression.

5. Conclusion

Fast wavelet transform and wavelet packet decomposition has been proposed in this work for compressing IR spectra. IR spectra of twenty organic compounds with similar structure were compressed at the 4th resolution level with the use of the Daubechies wavelet function D_{16} and cutoff value $\varepsilon = 0.20$ by using these WT methods. The results indicate that compression by FWT and WPT can reduce the spectral library file size by 90% which is better than that by FFT. Besides, the coefficient position retaining method was developed in this study to process spectral data with odd number data length, which cannot be handled by conventional FWT and WPT methods.

After data compression, the coefficients obtained were selected and employed to build a spectral library for future searching. Spectral library searching of this database was found to be better than that treated by FFT especially in the aspect of visual comparison in some cases. The scale coefficients as obtained from FWT and WPT can be used effectively for preliminary searching in a large spectral library. Our proposed methods can minimize the search time in the search by using the direct matching method. Since only a limited number of IR spectra were involved in the study, we only point out the advantages of library search for IR spectral database that is treated by the proposed WT methods. Further experiments are required to determine the optimal parameters for compression of spectra and the performance of library search with WT in a large library system.

Acknowledgements

This work was supported by Research Grant Council (RGC) of Hong Kong Special Administration Region (Grant No. HKP 45/94E) and the Research Committee of the Hong Kong Polytechnic University (Grant Nos. 350/529 and 351/577).

References

[1] W.A. Warr, C. Suhr (Eds.), Chemical Information Management, VCH Publishers, New York, 1992, p. 10.
[2] S.J. Haswell (Ed.), Practical Guide to Chemometrics, Marcel Dekker, New York, 1992.
[3] W.A. Warr, Anal. Chem., 65 (1993) 1045A–1050A and 1087A–1095A.
[4] W.A. Warr, Chemom. Intell. Lab. Syst. 10 (1991) 279–292.
[5] H. Hobert, in: J. Einax (Ed.), Chemometrics in Environmental Chemistry—Applications, Springer-Verlag Berlin Heidelberg, Germany, 1995, pp. 1–23.
[6] C.S. Rann, Anal. Chem. 44 (1972) 1669–1672.
[7] M.H. Adam, I. Black, Anal. Chim Acta 189 (1986) 353–363.
[8] D.R. Scott, Chemom. Intell. Lab. Syst. 4 (1988) 47–63.
[9] B.K. Alsberg, J. Chemom. 7 (1993) 177–193.
[10] B.K. Alsberg, O.M. Kvalheim, J. Chemom. 7 (1993) 61–73.
[11] B.K. Alsberg, E. Nodland, O.M. Kvalheim, J. Chemom. 8 (1994) 127–145.
[12] E.F. Crawford, R.D. Larsen, Anal. Chem. 49 (1977) 508–510.
[13] R.B. Lam, S.J. Foulk, T.L. Isenhour, Anal. Chem. 53 (1981) 1679–1684.
[14] P.M. Owens, T.L. Isenhour, Anal. Chem. 55 (1983) 1548–1553.
[15] F.T. Chau, K.Y. Tam, Comput. Chem. 18 (1994) 13–20.
[16] C.P. Wang, T.L. Isenhour, Appl. Spectrosc. 41 (1987) 185–194.
[17] E.R. Malinowski (Ed.), Factor Analysis in Chemistry, 2nd edn., Wiley, New York, 1991.
[18] G. Hangac, R.C. Wieboldt, R.B. Lam, T.L. Isenhour, Appl. Spectrosc. 36 (1982) 40–47.
[19] X.D. Dai, B. Joseph, R.L. Motard, in: R.L. Motard, B. Joseph (Eds.), Wavelet Application in Chemical Engineering, Kluwer Academic Publishers, MA, 1994, pp. 1–32.
[20] F.T. Chau, T.M. Shih, J.B. Gao, C.K. Chan, Appl. Spectrosc. 50 (1996) 339–349.
[21] K.M. Leung, F.T. Chau, J.B. Gao, Chemom. Intell. Lab. Syst. 43 (1998) 165–184.
[22] M. Bos, E. Hoogendam, Anal. Chim. Acta 267 (1992) 73–80.
[23] Z.X. Pan, X.G. Shao, H.B. Zheng, W. Liu, H. Wang, M.S. Zhang, Chin. J. Anal. Chem. 24 (1996) 149–153, (in Chinese).
[24] E.R. Collantes, R. Duta, W.J. Welsh, Anal. Chem. 69 (1997) 1392–1397.
[25] X.G. Shao, W.S. Gai, P.Y. Sun, M.S. Zhang, G.W. Zhao, Anal. Chem. 69 (1997) 1722–1725.
[26] P.B. Stark, M.M. Herron, A. Matteson, Appl. Spectrosc. 68 (1993) 1820–1829.
[27] M. Bos, J.A.M. Vrielink, Chemom. Intell. Lab. Syst. 23 (1994) 115–122.
[28] B. Walczak, B. Bogaert, D.L. Massart, Anal. Chem. 68 (1996) 1742–1747.
[29] B. Walczak, E. Bouveresse, D.L. Massart, Chemom. Intell. Lab. Syst. 36 (1997) 41–51.
[30] B.K. Alsberg, A.M. Woodward, W.K. Winson, J. Rowland, D.B. Kell, Analyst 122 (1997) 645–652.
[31] S.L. Shew, US Patent 5,436,477, July 25, Government Printing Office, Washington, DC, 1995.
[32] S.G. Nikolov, H. Hutter, M. Grasserbauer, Chemom. Intell. Lab. Syst. 34 (1996) 263–273.

[33] M. Wolkenstein, H. Hutter, S.G. Nikolov, M. Grasserbauer, Fresenius J. Anal. Chem. 357 (1997) 783–788.
[34] P. Guillemain, R. Kronland-Martinet, B. Martens, in: Y. Meyer (Ed.), Wavelets and Application: Proceedings of the Second International Conference, May 1989, Marseille, France, Springer Verlag, Pairs, 1992, pp. 38–60.
[35] H. Wang, Z.X. Pan, W. Liu, M.S. Zhang, S.Z. Si, L.P. Wang, Chem. J. Chin. Univ. 18 (1997) 1286–1290, (in Chinese).
[36] X.Q. Lu, J.Y. Mo, Analyst 121 (1996) 1019–1024.
[37] W. Liu, J.H. Xiong, H. Wang, Y.M. Wang, Z.X. Pan, M.S. Zhang, Chem. J. Chinese Universities 6 (1997) 860–863, (in Chinese).
[38] L. Yan, J.Y. Mo, Chinese Sci. Bull. 17 (1995) 1567, (in Chinese).
[39] J. Chen, H.B. Zhong, Z.X. Pan, M.S. Zhang, Chin. J. Anal. Chem. 24 (1996) 1002–1006, (in Chinese).
[40] H. Fang, H.Y. Chen, Anal. Chim. Acta 346 (1997) 319–325.
[41] L.J. Bao, J.Y. Mo, Z.Y. Tang, Anal. Chem. 69 (1997) 3053–3057.
[42] X.Y. Zou, J.Y. Mo, Anal. Chim. Acta 340 (1997) 115–121.
[43] D.N.S. Permann, H. Teitelbaum, J. Phys. Chem. 97 (1993) 12670–12673.
[44] M.V. Wickerhauser, Croat. Chem. Acta 68 (1995) 1–27.
[45] C.R. Mittermayr, S.G. Nikolov, H. Hutter, M. Grasserbauer, Chemom. Intell. Lab. Syst. 34 (1996) 187–202.
[46] S. Qian, H. Sun, Spectroscopy and Spectral Analysis 16 (1996) 1–8, (in Chinese).
[47] B. Walczak, D.L. Massart, Chemom. Intell. Lab. Syst. 36 (1997) 81–94.
[48] C.R. Mittermayr, E. Rosenberg, M. Grasserbauer, Anal. Commum. 34 (1997) 73–75.
[49] V.J. Barclay, R.F. Bonner, I.P. Hamilton, Anal. Chem. 69 (1997) 78–90.
[50] D. Permann, I. Hamilton, Phys. Rev. Letters 69 (1992) 2607–2610.
[51] P. Fischer, M. Defranceschi, Int. J. Quantum Chem. 45 (1993) 619–636.
[52] Z.M. Li, A. Borrmann, C.C. Martens, Chem. Phys. Letters 214 (1993) 362–366.
[53] D. Permann, I. Hamilton, J. Chem. Phys. 100 (1994) 379–386.
[54] A. Askar, A.E. Cetin, H. Rabitz, J. Phys. Chem. 100 (1996) 19165–19173.
[55] J.P. Modisette, P. Nordlander, J.L. Kinsey, B.R. Johnson, Chem. Phys. Letters 250 (1996) 485–494.
[56] F.T. Chau, T.M. Shih, J.B. Gao, C.K. Chan, Appl. Spectrosc. 50 (1996) 339–349.
[57] J.B. Gao, F.T. Chau, T.M. Shih, SEA Bull. Math. 20 (1996) 85–90.
[58] F.T. Chau, J.B. Gao, T.M. Shih, J. Wang, Appl. Spectrosc. 51 (1997) 649–659.
[59] W.C. Penski, D.A. Padowski, J.B. Bouck, Anal. Chem. 46 (1974) 955–957.
[60] G.T. Rasmussen, T.L. Isenhour, Appl. Spectrosc. 33 (1979) 371–376.
[61] G.W. Small, G.T. Rasmussen, T.L. Isenhour, Appl. Spectrosc. 33 (1979) 444–450.
[62] J.A. de Haseth, L.V. Azarraga, Anal. Chem. 53 (1981) 2292–2296.
[63] L.V. Azarraga, R.R. Williams, J.A. de Hasete, Appl. Spectrosc. 35 (1981) 466–469.
[64] I. Daubechies (Ed.), Ten Lectures on Wavelets, SIAM Press, Philadelphia, 1992.
[65] S. Palavajjhala, R.L. Motard, B. Joseph, in: R.L. Motard, B. Joseph (Eds.), Wavelet Application in Chemical Engineering, Kluwer Academic Publishers, MA, 1994, pp. 33–83.
[66] B.L. Borde, in: H.H. Szu (Ed.), Wavelet Application II: Proceedings of SPIE-the International Society for Optical Engineering, Vol. 2491, April 1995, Orlando, FL, SPIE-the International Society for Optical Engineering, Washington, DC, 1995, pp. 1073–1085.
[67] I. Daubechies, Commun. On Pure Appl. Math. 41 (1988) 909–996.
[68] I. Daubechies, IEEE Trans. Inf. Theory 36 (1990) 961–1005.
[69] S. Mallat, IEEE Trans. Acoust. 37 (1989) 2091–2110.
[70] S. Mallat, Trans. Am. Math. Soc. 315 (1989) 69–88.
[71] R.R. Coifman, Y. Meyer, S. Quake, M.V. Wickerhauser, in: J.S. Byrnes, J.L. Byrnes, K.A. Hargreaves, K. Berry (Eds.), Wavelets and their Applications, Kluwer Academic Publishers, The Netherlands, 1994, pp. 363–379.
[72] M.A. Cody, Dr. Dobbs J., 19(4) (1994) 44–54 and 100.
[73] S.R. Lowry, D.A. Huppler, C.R. Anderson, J. Chem. Inf. Comput. Sci. 25 (1985) 235–241.
[74] N. Saito, R.R. Coiffman, in: A.F. Laine, M.A. Unser (Eds.), Wavelet Applications in Signal and Image Processing II, Proceedings of SPIE-the International Society for Optical Engineering, Vol. 2303, July 1995, San Diego, CA, SPIE-the International Society for Optical Engineering, Washington, DC, 1994, pp. 2–14.
[75] J.W. Hayes, D.E. Glover, D.E. Smith, M.W. Overton, Anal. Chem. 45 (1973) 277–284.
[76] N. Morrison (Ed.), Introduction to Fourier Analysis, Wiley, New York, 1994, p. 388.
[77] P.B. Harrington, T.L. Isenhour, Appl. Spectrosc. 41 (1987) 1298–1302.
[78] L. Glasser, J. Chem. Edu. 64 (1987) A260–A266.
[79] The MathWorks, MATLAB for Windows Version 4.2 Reference Guide, Massachusetts, 1992.
[80] I. Daubechies, in: L.L. Schumaker, G. Webb (Eds.), Recent Advances in Wavelet Analysis, Academic Press, San Diego, 1994, pp. 237–257.
[81] N. Morrison (Ed.), Introduction to Fourier Analysis, Wiley, New York, 1994, p. 388.
[82] S.C. Lo, C.W. Brown, Appl. Spectrosc. 45 (1991) 1628–1632.