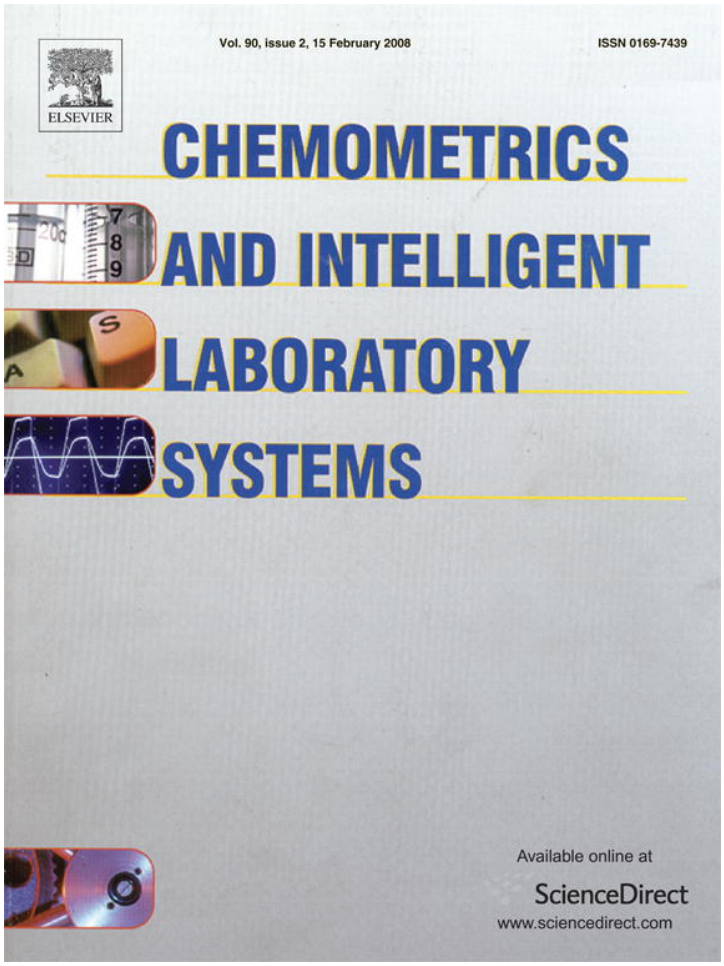

Available online at www.sciencedirect.com
ScienceDirect
Chemometrics and Intelligent Laboratory Systems 90 (2008) 161–168
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

Feature selection and classification of high-resolution NMR spectra in the complex wavelet transform domain

Seoung Bum Kim [a,*], Zhou Wang [b], Soontorn Oraintara [c], Chivalai Temiyasathit [a], Yodchanan Wongsawat [c]

[a] Department of Industrial and Manufacturing Systems Engineering 500 West First Street, Woolf Hall University of Texas at Arlington, Texas 76019, USA
[b] Department of Electrical and Computer Engineering University of Waterloo, Waterloo, ON N2L 3G1, Canada
[c] Department of Electrical Engineering 416 Yates Street, Nedderman Hall University of Texas at Arlington, Texas 76019, USA

Received 15 August 2007; received in revised form 21 September 2007; accepted 21 September 2007
Available online 29 September 2007

Abstract

Successful identification of the important metabolite features in high-resolution nuclear magnetic resonance (NMR) spectra is a crucial task for the discovery of biomarkers that have the potential for early diagnosis of disease and subsequent monitoring of its progression. Although a number of traditional features extraction/selection methods are available, most of them have been conducted in the original frequency domain and disregarded the fact that an NMR spectrum comprises a number of local bumps and peaks with different scales. In the present study a complex wavelet transform that can handle multiscale information efficiently and has an energy shift-insensitive property is proposed as a method to improve feature extraction and classification in NMR spectra. Furthermore, a multiple testing procedure based on a false discovery rate (FDR) was used to identify important metabolite features in the complex wavelet domain. Experimental results with real NMR spectra showed that classification models constructed with the complex wavelet coefficients selected by the FDR-based procedure yield lower rates of misclassification than models constructed with original features and conventional wavelet coefficients.

Keywords: Classification tree; Complex wavelet transforms; False discovery rates; Gabor coefficients; High-resolution NMR spectra; Metabolomics

1. Introduction

Metabolomics approaches that use high-resolution nuclear magnetic resonance (NMR) spectroscopy have been used to characterize metabolic variations in response to disease states, genetic medication, and nutritional intake. These include electrophoresis, chromatography, mass spectroscopy, and nuclear magnetic resonance spectroscopy. Of these, proton nuclear magnetic resonance (^{1}H-NMR) spectroscopy is efficient and cost effective because the analysis is either noninvasive or minimally invasive and requires little sample preparation [1,2]. A ^{1}H-NMR spectrum is a plot of the radio frequency applied against absorption. Fig. 1 shows a set of ^{1}H-NMR spectra generated by a 600 MHz ^{1}H-NMR spectroscopy. The *x*-axis indicates the chemical shift within units in parts per million (ppm), and the *y*-axis indicates the intensity values corresponding to each chemical shift. Traditionally, chemical shifts in the *x*-axis are listed from largest to smallest. The peaks with different scales in the spectra correspond to the specific resonance of chemical species in the samples.

Analysis of high-resolution NMR spectra usually involves combinations of multiple samples, each with a large number of metabolite features with different scales. This leads to a huge number of data points and a situation that challenges analytical and computational capabilities. To simplify such a complexity in NMR spectra, data size reduction is critical. Data reduction can be done by selecting a small number of important features that preserve the most information contained in the original data. The widely used methods for identifying important metabolite features in spectral data include principal component analysis (PCA) and Partial Least Squares (PLS) [3,4]. Both PCA and PLS attempt to extract new features based on the transformation of the original features. In general, the first few transformed features obtained through PCA are sufficient to account for most of

* Corresponding author.
E-mail address: sbkim@uta.edu (S.B. Kim).

0169-7439/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2007.09.005

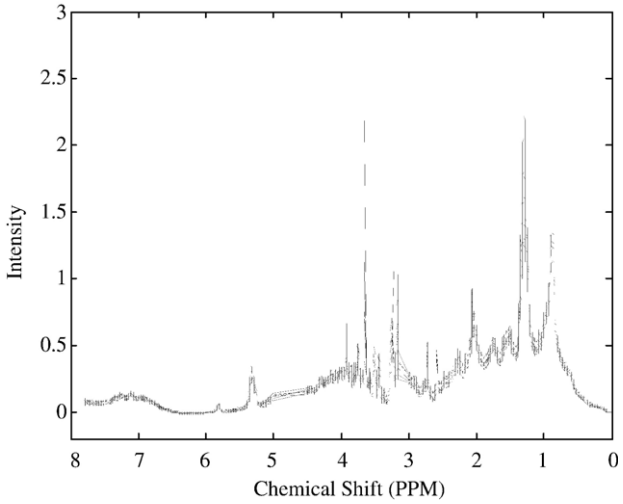

Fig. 1. Multiple spectra generated by a 600 MHz ^{1}H-NMR spectroscopy.

variability in the original data. Similarly, the first few transformed features obtained by PLS are sufficient to maximize class separability in the original data. These reduced dimensions can diminish the computational complexity for the analysis of NMR spectra. However, the reduced dimensions from PCA or PLS do not provide a clear interpretation with respect to the original features because they are linear combinations of a large number of original features. Interpretation problems posed by the transformation process in PCA and PLS can be overcome by selecting the best subset of given features in a spectrum. Recently, a two-stage genetic programming (GP) method was successfully used for selecting a subset of original metabolite features in NMR spectra for the classification of genetically modified barley [5]. However, GP may not always provide reliable results in high-dimensional and noisy data.

The present methods (e.g., PCA, PLS, and GP) each have their own advantages and disadvantages, and the choice among them depends upon the purpose of the application. However, the present methods have been conducted in the original frequency domain and have overlooked the fact that an NMR spectrum comprises a number of local bumps and peaks with different scales. This motivates the focus of this paper on the development of efficient analytical tools to handle the multiscale nature of NMR spectra.

Wavelets have the advantage of the locality of the analysis and their ability to handle multiscale information efficiently. Despite the numerous studies of wavelets that have been conducted in the general field of signal and image processing [6], wavelets have not been thoroughly studied for application to NMR spectra. Wavelet transform has been used to improve the detection of small chemical species in NMR spectra by suppressing the water signal [7,8]. Qu and his coauthors [9] used decimated discrete wavelet transform to analyze mass spectrometry data (similar to NMR spectra) with class information. They first conducted thresholding to reduce the size of data and constructed a classification model with a small subset of wavelet coefficients that best differentiated between the two classes. The major disadvantage of using decimated discrete wavelet transforms is that this method is insufficiently robust in terms of signal translation, which means that the representation of a signal by wavelet coefficients is highly dependent on the signal's relative position. This property is critical to the analysis of multiple NMR spectra because small variations in spectra caused by concentration, pH, temperature, and instrumental instabilities may influence the spectral alignment and thus can interfere with direct comparisons between spectra.

In the present study we propose the use of complex wavelet transforms for the analysis of multiple NMR spectra. A complex wavelet transform has the energy shift-insensitive property that leads to improvements in the comparability of multiple spectra. Among the various complex wavelets available, we choose the Gabor wavelets for three reasons: First, according to the Gabor uncertainty principle (which is a generalization of the Heisenberg uncertainty principle that originated in quantum mechanics), the time-frequency resolution of a signal is fundamentally limited by a lower bound on the product of its bandwidth and duration, and the Gabor filters are the only family of filters that achieve this lower bound [10]. In other words, the Gabor filters provide the best compromise between simultaneous time and frequency signal representations. Second, the Gabor wavelets are easily and continuously tunable for both the center frequencies and for bandwidths. Third, from the biological point of view, Gabor filters types have been widely used to model the profiles of the receptive filed of simple cells in the primate cortex [11].

One essential step in the use of NMR spectra for real applications is to identify the features associated with the problems being studied. NMR spectra shave a large number of features, and many of them can be considered redundant and irrelevant for the subsequent modeling processes. Thus, the process of identifying a small number of important features can be equivalent to the problem of optimal dimension reduction (i.e., data reduction) so that the remaining feature space has the strongest statistical significance for the purpose of pattern recognition. Most of the current dimension reduction processes have been carried out directly in the original frequency domain, and only a few have been performed in the transform domain. The present study proposes to use a multiple testing procedure based on a false discovery rate in the complex wavelet transform domain to improve feature selection and classification in NMR spectra.

The major purposes of this paper are: (1) To examine the feasibility of using the complex wavelet transform to efficiently analyze the multiscale nature of NMR spectra; (2) To identify the important metabolite features in the complex wavelet domain that play a significant role in discriminating between spectra under different experimental conditions; and (3) To evaluate the appropriateness of the identified metabolite features based on their classification capabilities.

2. Experimental data

We used plasma samples obtained from four healthy subjects under controlled metabolic conditions in the Emory General Clinical Research Center (GCRC). The subjects signed an informed consent approved by the Emory Institutional Review

Board. During the 12-day GCRC admission, the subjects consumed defined diets at standardized intervals. For the first two days (equilibration), the subjects consumed balanced meals from a plan in which foods were selected to ensure adequate energy, protein and sulfur amino acid (SAA) intake (SAA at 19 mg/kg/day). After this phase, subjects were placed on constant semipurified diets designed to alter SAA intake. The diets provided adequate energy and amino acid nitrogen to meet the estimated maintenance needs of individual subjects. The L-amino acid component of the diet was altered to provide zero SAA during the initial five days and 117 mg/kg per day during the latter five days of the GCRC stay. Blood was drawn serially 34 times from four subjects over ten days, and ^{1}H-NMR spectra were obtained by a Varian INOVA 600 MHz instrument. During the first 17 time points, blood was collected from each subject consuming zero SAA (zero-SAA phase) and 117 mg/kg per day SAA during the latter 17 time points (supplemented-SAA phase). Thus, the total number of spectra used in this study is 136 (=4 subjects × 34 spectra).

Raw NMR spectra require preprocessing, which includes phase/baseline correction, elimination of uninformative spectral regions containing no significant metabolite signals, alignment, and normalization relative to the internal standard. The NUTS software (Acron NMR Inc., Livermore, CA) was used for phase/baseline correction. To adjust for the variable suppression of the large water signal in NMR spectra and enhance the detection of metabolites, water signal and other uninformative spectral regions were eliminated. We used MATLAB (Mathwork Inc., Natick, MA) with the beam search algorithm [12] for initial spectral alignment. Finally, normalization of NMR spectra was achieved by scaling to the integral of the internal standard.

3. Methods

3.1. Complex wavelet transforms

One of the major aspects of NMR spectral analysis is how to associate the spectral region-of-interest (SROI) in NMR spectra with their corresponding chemical shifts. One critical observation is that the optimal SROI in NMR spectra varies with different chemical species. This motivates us to use a multiscale approach because no single scale (no matter if it is coarse or fine) can optimally capture all the SROIs in an NMR spectrum at the same time. Wavelet analysis provides a natural solution for this purpose and serves as a convenient and flexible framework for localized representation of signals simultaneously in space and frequency.

We consider complex wavelets as dilated/contracted and translated versions of a complex-valued "mother wavelet" $w(x) = g(x)e^{j\omega_c x}$, where ω_c is the center frequency of the modulating band-pass filter, and $g(x)$ is a slowly varying and symmetric real-valued function. The family of wavelets derived from the mother wavelet can be expressed as:

$$w_{s,p}(x) = \frac{1}{\sqrt{s}} w\left(\frac{x-p}{s}\right) = \frac{1}{\sqrt{s}} g\left(\frac{x-p}{s}\right) e^{j\omega_c (x-p)/s}, \quad (1)$$

where $s \in R^+$ and $p \in R$ are the scale and translation factors, respectively. Considering the fact that $g(-x)=g(x)$, the wavelet transform of a given real signal $f(x)$ can be written as:

$$F(s,p) = \int_{-\infty}^{\infty} f(x)\, w^*_{s,p}(x)\, dx = \left[f(x) * g\left(\frac{x}{s}\right) e^{j\omega_c x/s} \right]_{x=p}. \quad (2)$$

In other words, we can use this to compute the wavelet coefficient $F(s,p)$ at any given scale s and location p. In this paper, we are specifically interested in the Gabor wavelet transform, which is a special case of the complex wavelet transform described above. In particular, we define $g(x)$ to have a Gaussian shape:

$$g(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-x^2}{\sigma^2}}. \quad (3)$$

One interesting property of the Gabor wavelet transform is that it is energy shift-insensitive [13,14]. This can be easily shown by the Fourier domain analysis: Using the convolution theorem and the shifting and scaling properties of the Fourier transform, it is not difficult to derive that

$$F(s,p) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega)\, \sqrt{s} G(s\omega - \omega_c)\, e^{j\omega p} d\omega, \quad (4)$$

where $F(\omega)$ and $G(\omega)$ are the Fourier transforms of $f(x)$ and $g(x)$, respectively. Now suppose that the function $f(x)$ has been shifted by a small amount Δx, i.e., $f'(x)=f(x+\Delta x)$. This corresponds to a linear shift in the Fourier domain: $F'(\omega) = F(\omega)e^{j\omega_c \Delta x}$. Substitute this into Eq. (4), we obtain

$$\begin{aligned} F'(s,p) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega)\, \sqrt{s} G(s\omega - \omega_c)\, e^{j\omega(p+\Delta x)} d\omega \\ &= \frac{e^{j\omega_c \Delta x/s}}{2\pi} \int_{-\infty}^{\infty} F(\omega) \sqrt{s} G(s\omega - \omega_c)\, e^{j\omega p} e^{j(\omega-\omega_c/s)\Delta x} d\omega \\ &\approx F(s,p) e^{j\omega_c \Delta x/s}. \end{aligned} \quad (5)$$

Here the approximation is valid when Δx is small relative to the Gaussian function $g(x)$. The information we are particularly interested in here is the magnitude of the Gabor coefficient $|F(s,p)|$. From Eq. (5), we have

$$|F'(s,p)| \approx |F(s,p)|.$$

In other words, the magnitude (or energy) of the Gabor wavelet coefficient does not change significantly with a small translation. Such an energy shift-insensitive property is very important in the analysis of NMR spectra because a small misalignment between multiple NMR spectra is unavoidable (even after preprocessing), and the misalignment may interfere with direct comparisons between NMR spectra.

3.2. A multiple testing procedure based on a false discovery rate

In the wavelet transform domain, significant data reduction of a signal can be achieved by applying a thresholding method.

The basic idea of thresholding is to zero out the small-magnitude wavelet coefficients in the wavelet transform domain [15,16]. The fundamental assumption behind the thresholding is that informative signals result in large-magnitude wavelet coefficients, and the small-magnitude coefficients most likely come from noise. Although the thresholding algorithm can be efficient for reconstructing an original signal with a few important wavelet coefficients, these coefficients may not always produce maximum discrimination between sample conditions. For example, some wavelet coefficients with small magnitude may be neglected even though they are indeed important for classification. In the present study we used a feature selection procedure to identify wavelet coefficients to maximize the separation of the classes in NMR spectra. More precisely, a multiple testing procedure that controls the false FDR was used to identify significant Gabor coefficients that discriminate between the spectra under different conditions. The FDR is the error rate in multiple hypothesis tests and is defined as the expected proportion of false positives among all the hypotheses rejected [17]. In our problem, the rejected hypotheses can be interpreted as the significant Gabor coefficients necessary for classification.

The FDR-based procedure is explained with our experimental data. Let δ_{jk} be the magnitude of the Gabor coefficient at the kth position (for $k=1,2,\ldots,K$) of the jth class (for $j=1,\ldots,J$). As illustrated in Section 2, our experimental data comprise 136 NMR spectra in which half of the spectra were taken from the zero-SAA phase and the other half were taken from the supplemented-SAA phase. The goal is to identify a set of δ_k that maximizes the separability between the two SAA phases. For each wavelet coefficient, a null hypothesis states that the average magnitudes of Gabor coefficients are equal between the two SAA phases, and the alternative hypothesis is that they differ. The two-sample t statistic for the δ_k is

$$t_k = \frac{\overline{\delta}_{1k} - \overline{\delta}_{2k}}{\sqrt{\frac{\hat{\sigma}^2_{1k}}{n_1} + \frac{\hat{\sigma}^2_{2k}}{n_2}}}, \tag{6}$$

where $\overline{\delta}_{1k}$, $\hat{\sigma}^2{}_{1k}$, and n_1 are the sample mean, variance, and the number of samples from the first condition, respectively. Similarly, $\overline{\delta}_{2k}$, $\hat{\sigma}^2{}_{2k}$, and n_2 were obtained from the second condition. By asymptotic theory, t_k in Eq. (6) approximately follows a t-distribution on the assumption that the null hypothesis is true. Using this, the p-values for δ_k for $k=1,2,\ldots,K$ can be obtained. In multiple testing problems, it is well-known that applying a single testing procedure leads to an exponential increase of false positives. To overcome this, the methods that control family-wise error rates have been proposed. The most widely used one is the Bonferroni method that uses a more stringent threshold [18] . However, the Bonferroni method is too conservative, and it often fails to detect the "true" significant features. A more recent multiple testing procedure that controls FDR was proposed by Benjamini and Hochberg [17]. The advantage of the FDR-based procedure is that it identifies as many significant hypotheses as possible while keeping a relatively small number of positives [19,20]. A summary of the FDR-based procedure is as follows:

- Select a desired FDR level α between 0 and 1.
- Order the p-values: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(K)}$
- Find the largest s denoted as w, where $w = \max\left[s : p_{(s)} \leq \frac{s}{k}\frac{\alpha}{\pi}\right]$, where k is the total number of hypotheses, and π denotes the proportion of true null hypothesis. The present study uses $\pi=1$, the most conservative choice.
- Let the p-value threshold be $p_{(w)}$, and declare the Gabor coefficient δ_k as significant if and only if $p_s \leq p_{(w)}$.

3.3. Classification tree and cross validation

A classification model was used to examine the advantage of using the complex wavelet transform and FDR-based feature selection in NMR spectra. We used a classification tree, one of the widely used classification methods. Classification trees partition the input (feature) space into disjoint hyper-rectangular regions according to performance measures such as misclassification errors, Geni index, and cross-entropy and then fit a constant model in each disjoint region [21]. The number of disjoint regions (equivalent to the number of terminal nodes in a tree) should be determined appropriately because a very large tree overfits the training set, while a small tree cannot capture important information in the data. In general, there are two approaches to determining the tree size. The first approach is the direct stopping methods that attempt to stop tree growth before the model overfits the training set. The second approach is tree pruning that removes the leaves and branches of a full-grown tree to find the right size of the tree. In the present study the Geni index was used as a performance measure. To determine tree size, we stop the growth of a tree when the number of data points in the terminal node reaches five.

In order to estimate the true misclassification rate of classification tree models, we used a cross-validation technique. Specifically, we used a four-fold cross validation in which the experimental data were split into four groups corresponding to four subjects. Three subjects were used for training the models, and the one remaining subject was used for testing. This process was repeated three more times. The final classification results from the four different testing samples were then averaged to obtain the misclassification rates (or cross-validated error rates) of the classification tree models.

4. Results and discussion

4.1. Multiscale modeling by the Gabor and Symlet wavelet transforms

We used the Gabor and Symlet wavelet transforms to obtain the wavelet coefficients. The parameters ω_c and σ in the Gabor wavelet decomposition control, respectively, the frequency band and the locality. In order to have a fair comparison, these parameters are tuned so that their frequency responses match those of the Symlet, which can be viewed as

a real wavelet transform. In our implementation, we set $\omega_c = 0.75\pi$ and $\sigma = 2.1$, and compute the Gabor coefficients at five discrete levels with corresponding decimation factors of 2, 4, 8, 16, and 32. The maximum number of levels of the Gabor decomposition (Level 5 with a decimation factor of 32) is determined by the maximal scale of the features in the NMR spectrum. For example, a single chemical specie may correspond to a SROI in the NMR spectrum. The coarsest scale (corresponding to the maximum level) of Gabor decomposition should be able to fully cover the SROI with one Gabor profile. Fig. 2 shows an example of the Gabor coefficients $|F(s,p)|$ computed at five different levels from our implementation.

To demonstrate the advantage of energy-shift insensitivity of the complex wavelet transform, we compare with the Symlet, which is one of the commonly used real wavelet transforms. Consequently, the Symlet wavelet of order 16 (Symlet-16) is selected, and the maximum level of Symlet decomposition is set to five. Fig. 3 shows the similarity of the frequency responses of all five levels of the Gabor and Symlet wavelets.

Furthermore, to demonstrate the improvement of processing in the transform domain, the resulting wavelet coefficients from both transforms are downsampled so that the total number of transform coefficients is approximately the same as the total number of original features. Fig. 4 illustrates a downsampling process of five levels in the Gabor and Symlet wavelet transform domains.

The original NMR spectrum, $f_{\text{original}}(x)$, is decomposed (by Gabor or Symlet wavelet transform) into five levels. The resulting wavelet coefficients $f_i(x)$ of the ith level are downsampled by 2^i, where $i = 1, \ldots, 5$.

The Gabor filters are defined by $h_i(x) = g\left(\frac{x}{s}\right)e^{j\omega_c x/s}$, where $s = 2^i$, $\omega_c = 0.75\pi$, and $g(x)$ is defined by Eq. (3). The Symlet filters are defined as follows. Let $H_i(\omega)$ be the Fourier transform of $h_i(x)$, then

$$H_1(\omega) = \hat{H}_0(\omega)\hat{H}_1\left(\omega^2\right),$$

$$H_2(\omega) = \hat{H}_0(\omega)\hat{H}_0\left(\omega^2\right)\hat{H}_1\left(\omega^4\right),$$

$$H_3(\omega) = \hat{H}_0(\omega)\hat{H}_0\left(\omega^2\right)\hat{H}_0\left(\omega^4\right)\hat{H}_1\left(\omega^8\right),$$

$$H_4(\omega) = \hat{H}_0(\omega)\hat{H}_0\left(\omega^2\right)\hat{H}_0\left(\omega^4\right)\hat{H}_0\left(\omega^8\right)\hat{H}_1\left(\omega^{16}\right),$$

$$H_5(\omega) = \hat{H}_0(\omega)\hat{H}_0\left(\omega^2\right)\hat{H}_0\left(\omega^4\right)\hat{H}_0\left(\omega^8\right)\hat{H}_0\left(\omega^{16}\right)\hat{H}_1\left(\omega^{32}\right),$$

where $\hat{H}_0(\omega)$ and $\hat{H}_1(\omega)$ are the lowpass and highpass filters corresponding to the scaling and wavelet functions of the Symlet wavelet transform, respectively. The impulse responses $h_i(x)$ can be obtained by taking the inverse Fourier transform of $H_i(\omega)$. As a result of downsampling, we obtained 8181 wavelet coefficients from five levels of the Gabor and Symlet transforms. These coefficients will be used for feature selection described in Section 4.2. Gabor and Symlet wavelet transforms were performed using MATLAB (Mathwork Inc., Natick, MA).

4.2. Selection of important metabolite features

We applied the FDR-based multiple testing procedure to identify significant metabolite features in the original domain and in the wavelet domains (e.g., Gabor and Symlet). The total number of metabolite features in the original domain is 8444, and the number of features selected by the FDR-based procedure is 1278 (Table 1). The number of Symlet and Gabor coefficients selected by the FDR-based procedure is reported in Table 1, showing that significant data reduction is achieved.

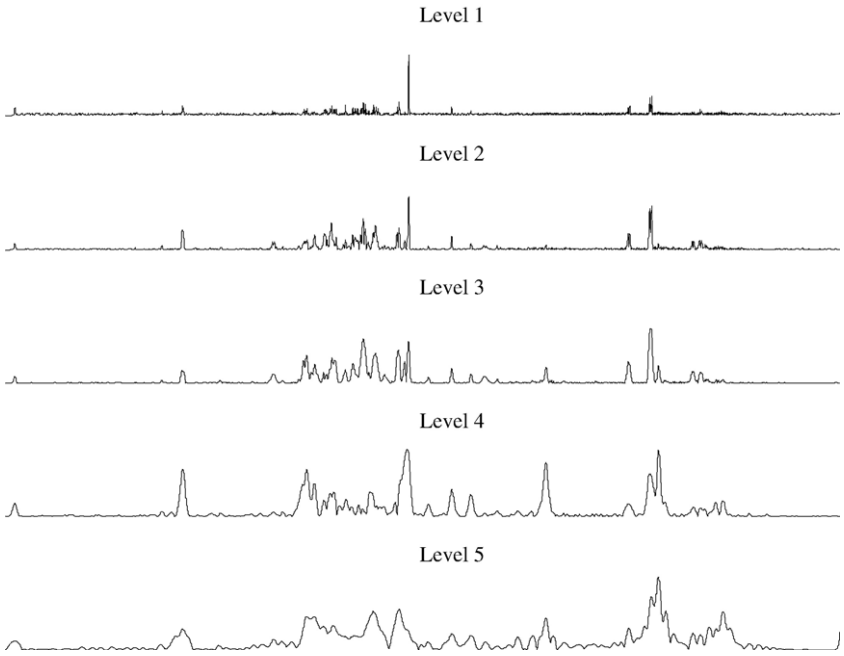

Fig. 2. The magnitude of the Gabor wavelet coefficients at five levels (cropped from a long MNR spectrum for better visualization).

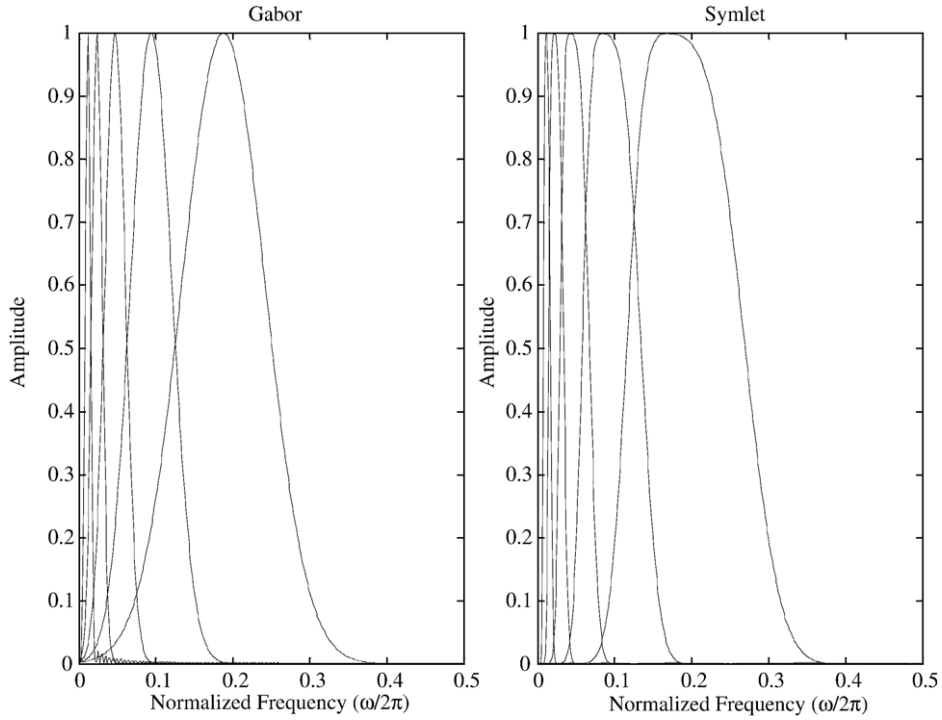

Fig. 3. Frequency responses of five levels of Gabor wavelet transforms and Symlet-16 wavelet transforms.

Interpretation of the FDR results can be made as follows: In the original domain, for example, the FDR-based procedure identified 1278 significant features. This implies that there are on average, 13 ($12.78 = 1278 \times 0.01$) false discoveries out of the 1278 coefficients discovered (identified as significant) through the FDR-based procedure. In the wavelet domains, the FDR-based procedure selected 20 significant Symlet coefficients and 21 significant Gabor coefficients. This implies that there is less than one false discovery in the 21 and 20 features selected at FDR level = 0.01. Some remarks on the FDR-based multiple testing procedure are given below.

Remark 1. A higher level of FDR increases the number of significant features. This yields higher statistical power but produces more false positives.

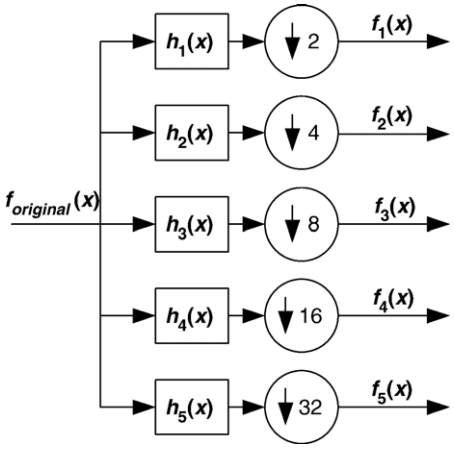

Fig. 4. Downsampling process of five levels of Gabor and Symlet wavelet transforms.

Remark 2. Metabolite features in NMR spectrum are correlated. Even though wavelet transforms alleviate the serial correlation between the features, an appropriate procedure is needed to take into account the correlation between the wavelet coefficients. The original FDR-based procedure proposed by [17] assumed that all hypotheses are independent. Later work by [22] revealed that the conclusion still holds if the tests are positively correlated. Furthermore, they extended the original FDR-based procedure, with minor modification, to handle any correlation structures.

Remark 3. The FDR-based procedure for feature selection can be generalized into problems with more than three classes. As an alternative to two-sample statistics for two class problems, an analysis of variance table is constructed for each wavelet coefficient and its significance is tested based on an *F*-test.

4.3. Classification results

To evaluate the adequacy of the metabolite features obtained from Sections 4.1 and 4.2, the following six data sets with the

Table 1
Number of significant metabolite features from the FDR-based multiple testing procedure (FDR level = 0.01) in the original and wavelet domains

Types of features	Total number of features	Number of selected features	Reduction rate
Original	8444	1278	84.86%
Symlet-16 (Real Wavelet)	8181	20	99.74%
Gabor(Complex Wavelet)	8181	21	99.76%

different sets of features were used for the tree classification method. The values in parentheses indicate the numbers of features.

- Data set 1: All metabolite features in the original domain (8444).
- Data set 2: Coefficients downsampled from five levels of the Symlet wavelet transform (8181).
- Data set 3: Coefficients downsampled from five levels of the Gabor wavelet transform (8181).
- Data set 4: Metabolite features selected from Data set 1 by the FDR-based procedure (1278).
- Data set 5: Coefficients selected from Data set 2 by the FDR-based procedure (21).
- Data set 6: Coefficients selected from Data set 3 by the FDR-based procedure (20).

Data set 1 is the full data set containing all metabolite features in the original domain. Data sets 2 and 3 contain the downsampled Symlet and Gabor wavelets. Data sets 3, 4, and 5 consist of the features selected from Data sets 1, 2, and 3 using the FDR-based feature selection method (FDR level = 0.01). Classification error rates obtained from cross validation are shown to evaluate the efficacy of the data sets with different sets of features (Fig. 5). Comparing the features in the original domain with the features in the wavelet domain, it is clear that the wavelet transform approaches improve classification accuracy. This result demonstrates the advantage of using wavelet transforms in the analysis of the multiscale nature of NMR spectra. Between the complex wavelet and real wavelet transforms, the classification tree models with Gabor wavelet coefficients yield a lower misclassification rate than the models with Symlet coefficients. This implies that the energy shift-insensitive property of the complex wavelets results in the reduction of the misclassification rate. Furthermore, the classification tree models constructed with the features selected by the FDR-based feature selection method produce smaller misclassification rates than those without applying the FDR-based method; this demonstrates that feature selection by FDR was adequate and that it successfully eliminated non-informative features and improved overall classification accuracy.

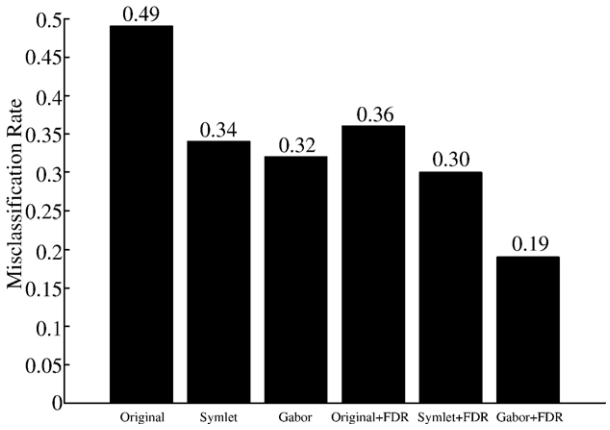

Fig. 5. Misclassification rates (cross-validation error rates) of the classification trees for six different sets of features.

In particular, the classification accuracy was significantly improved by using the Gabor coefficients selected by the FDR-based procedure.

5. Conclusions

We have proposed to use a complex wavelet transform combined with the FDR-based feature selection method to improve feature selection and classification of high-resolution NMR spectra. The ability of wavelet transforms to break down the original spectrum into different resolution levels allows us to investigate the metabolite feature with different scales. Moreover, the energy shift-insensitive property in the complex wavelet transform can efficiently handle misalignment and enables direct comparison among multiple NMR spectra. The FDR-based feature selection procedure treats all the wavelet coefficients simultaneously and systematically identifies important features in NMR spectra.

The effectiveness of the complex wavelet transform and the FDR-based procedure was demonstrated using real NMR spectra in which the ultimate goal was to determine major metabolite features that contribute to distinguishing between the zero-SAA and supplemented-SAA phases. We compared the classification capabilities of the proposed approach with the original features and with the noncomplex wavelet transform. The results from classification tree models have shown that the proposed approach significantly increases overall classification performance.

Our study extends the application scope of both the complex wavelet transform and the FDR methodology and demonstrates that their systematic application to controlled clinical study provides a useful means to extract meaningful information from high-resolution NMR spectra. We hope that the procedure presented here stimulates further investigation into the development of better procedures for multiscale modeling and analysis of high-resolution NMR spectra.

Acknowledgments

We thank the referee for the constructive comments and suggestions, which greatly improved the quality of the paper. We are grateful to Dean P. Jones and Thomas R. Zeigler in the Emory University Medical School for their useful comments. We also thank the nursing and laboratory staff of the Emory General Clinical Research Center for their valuable helps in collecting samples. This work was supported in part by NSF Grant ECS-0528964.

References

[1] H. Antti, M.E. Bollard, T. Ebbels, H. Keun, J.C. Lindon, J.K. Nicholson, E. Holmes, Journal of Chemometrics 16 (2002) 461–468.
[2] J.C. Lindon, Business briefing: Future Drug Discovery 9 (2004) 1–6.
[3] R. Goodacre, E.V. York, J.K. Heald, I.M. Scott, Phytochemistry 62 (2003) 859–863.
[4] H.S. Tapp, M. Defernez, E.K. Kemsley, Journal of Agricultural And Food Chemistry 51 (2003) 6110–6115.
[5] R.A. Davis, A.J. Charlton, S. Oehlschlager, J.C. Wilson, Chemometrics and Intelligent Laboratory Systems 81 (2006) 50–59.

[6] S. Mallat, A Wavelet Tour of Signal Processing, Academic Press, New York, 1999.
[7] D. Barache, J. Antoine, J. Dereppe, Journal of Magnetic Resonance 128 (1997) 1–11.
[8] U.L. Gunther, C. Ludwig, H. Ruterjans, Journal of Magnetic Resonance 156 (2002) 19–25.
[9] Y. Qu, B.-L. Adam, M. Thornquist, J.D. Potter, M.L. Thompson, Y. Yasui, J. Davis, P.F. Schellhammer, L. Cazares, M. Clements, G.L. Write, Z. Feng, Biometrics 59 (2003) 143–151.
[10] D. Gabor, Journal of Institution Electrical Engineering (1946) 429–457.
[11] D.A. Pollen, S.F. Ronner, Science 212 (1981) 1409–1411.
[12] G.C. Lee, D.L. Woodruff, Analytica Chimica Acta 513 (2004) 413–416.
[13] I.W. Selecnick, R.G. Baraniuk, N.C. Kingsbury, IEEE Signal Processing Magazine 2 (2005) 123–151.
[14] Z. Wang, E.P. Simoncelli, IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceeding of the Conference, March 18–23, 2005, Philadelphia, USA, 2005, pp. 573–576, PA.
[15] D.L. Donoho, I.M. Johnstone, Biometrika 81 (1994) 425–455.
[16] D.L. Donoho, I.M. Johnstone, Journal of American Statistical Association 90 (1995) 1200–1224.
[17] Y. Benjamini, Y. Hochberg, Journal of The Royal Statistical Society Series B. Methodological 57 (1995) 289–300.
[18] J.P. Shaffer, Annual Review of Psychology 46 (1995) 561–584.
[19] S.B. Kim, K.-L. Tsui, M. Borodovsky, International Journal of Bioinformatics Research and Applications 2 (2006) 193–217.
[20] J.D. Storey, Annals of Statistics 31 (2003) 2013–2035.
[21] T. Hastie, R. Tibshirani, J. Friedman, The Element of Statistical Learning, Springer, New York, 2001.
[22] Y. Benjamini, D. Yekutieli, Annals of Statistics 29 (2001) 1165–1188.