Anal. Chem. **1996,** *68,* 1742−1747

Application of Wavelet Packet Transform in Pattern Recognition of Near-IR Data

Beata Walczak,† Bas van den Bogaert, and Desiré Luc Massart*

ChemoAc-Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussel, Belgium

The wavelet packet transform is studied as a tool for improving pattern recognition based on near-infrared spectra. Application to the preprocessing of the spectra improves the classification when compared to using either the standard normal variate method or no pretreatment at all. Selecting features from a local discriminant basis instead of from a full decomposition does not improve the results.

In this paper we study the application of the wavelet packet transform (WPT) as a tool for improving pattern recognition based on near-infrared spectra. To position this tool in the pattern recognition process, it is useful to describe this process as consisting of different steps.

Construction of the pattern space: selection of representative training data.

Data pretreatment: reduction of measurement noise and other sources of error.

Dimension reduction: feature selection or reduction.

Classification: build a model to classify new samples.

Validation: check if the model is stable and not built on noise.

Methods are available for most steps. Only the construction of the pattern space is not methodical; it is a matter of common sense to select meaningful objects and variables. Data pretreatment depends on the type of effects to be reduced. Near-IR data generally suffer from the effects of differences in particle size or film thickness. These cause variation in optical path length, resulting in varying offset and curvature of the spectral baseline. A popular means of reducing this kind of variation in the data is standard normal variate (SNV).[1] For feature selection, several general methods are available. A simple and powerful approach uses Fisher's criterion,[2] the ratio of between-group to within-group variance. Variables for which this ratio is large can be considered useful for classification purposes. This univariate approach, however, cannot avoid selecting features carrying the same information. Multivariate methods, based, e.g., on a genetic algorithm,[3,4] do not have this disadvantage. In exchange, they are time-consuming, and their results are not easy to validate. For the classification step, the core of pattern recognition, methods are available that are both general and powerful, e.g., linear discriminant analysis (LDA), a parametrical statistical technique.[5] The last step can be performed as leave-one-out cross-validation. It is important that the validation should comprise both the classification and the feature selection.

The use of near-IR data creates some typical problems for pattern recognition. As already mentioned, it requires specific data pretreatment. Furthermore, there are many variables per object, stressing the need for dimension reduction. The spectral features are usually localized, meaning that there are large regions containing irrelevant information. However, there will still be many correlated variables.

The discrete wavelet transform (DWT)[6−8] and its generalization, the wavelet packet transform (WPT),[9,10] are promising tools for the pretreatment of near-IR data. They give another view of the data, which can be useful to detect features. In that sense, they are related to Fourier transformation, which has also been suggested for pretreating near-IR data.[11] DWT and WPT were developed to describe and clarify local data structures. It is this property that suggests their use for near-IR data. These transforms represent relatively recent mathematical developments, and they have not found many applications in chemistry yet. Bos and Vrielink[12] have used the DWT for the purpose of reducing data dimensionality in classification based on IR spectra. In DWT, a signal is decomposed into levels of a decreasing number of points by recursive application of a pair of filters. Bos used entire levels as the input to both linear and nonlinear classifiers. They experimented with the selection of both levels and filters and found that a strong data reduction could be accomplished without loss of predictive ability. In this paper, we will evaluate the use of the more general WPT. This is a very flexible tool, giving many different views of the data, of which the DWT is just one. To control this flexibility, a means of selecting from these views is welcome. Saito introduced such control by extending Coifman's best-basis[13] to what he calls a local discriminant basis (LDB).[14] We will compare the LDB to using the full flexibility without any selection.

THEORY

The wavelet packet transform is a generalisation of the discrete wavelet transform, and it is convenient to introduce the WPT from

† On leave from Silesian University, Katowice, Poland.

(1) Barnes, R. J.; Dhanoa, M. S.; Lister, S. J. *Appl. Spectrosc.* **1989**, *43*, 772−777.
(2) Massart, D. L.; Vandeginste, B. G. M.; Deming, S. N.; Michotte, Y.; Kaufman, L. *Chemometrics; a textbook*, Elsevier: Amsterdam, 1988.
(3) Leardi, R.; Boggia, R.; Terrile, M. *J. Chemometr.* **1992**, *6*, 267−281.
(4) Lucasius, C. B.; Beckers, M. L. M.; Kateman G. *Anal. Chim. Acta* **1994**, *286*, 135−153.
(5) Coomans, D.; Massart, D. L.; Kaufman, L. *Anal. Chim. Acta* **1979**, *112*, 97−122.
(6) Press,W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical recipes in C*; Cambridge University Press: Cambridge, UK, 1992.
(7) Chui, C. K. *Introduction to Wavelets*; Academic Press: Boston, 1991.
(8) Strang, G. *SIAM Rev.* **1989**, *31*, 614−627.
(9) Coifman, R. R.; Meyer, Y.; Wickerhauser, V. In *Progress in wavelet analysis and applications*; Meyer, Y., Roques, S., Eds.; Editions Frontieres: Gif-sur-Yvette, France, 1993, 77−93.
(10) Cody, M. A. *Dr. Dobb's J.* **1994**, *17*, 16−28.
(11) Osborne, B. G.; Fearn, T. *Near infrared spectroscopy in food analysis*; Longman Scientific & Technical: Harlow, UK, 1986.
(12) Bos, M.; Vrielink, J. A. M. *Chemom. Intell. Lab. Syst.* **1994**, *23*, 115−122.
(13) Coifman, R. R.; Wickerhauser, M. V. *IEEE Trans. Inform. Theory* **1992**, *38*, 713−719.
(14) Saito, N. Local feature extraction and its applications using a library of bases. Ph.D. thesis, Yale University, New Haven, CT, 1994.

S0003-2700(95)01091-2 CCC: $12.00

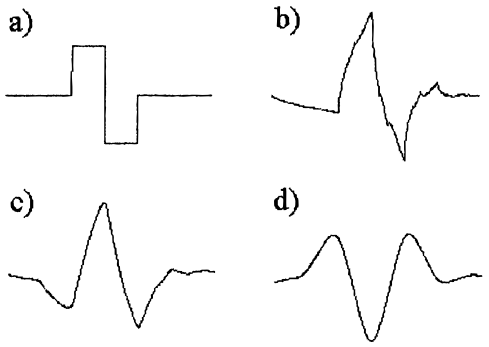

Figure 1. Daubechies wavelets 1, 2, 4, and 8.

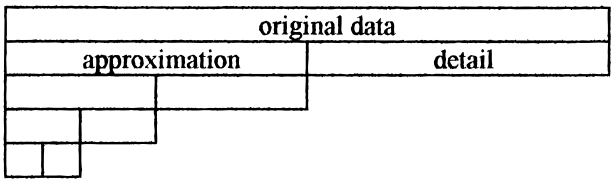

Figure 2. Scheme of Mallat's pyramid algorithm for calculating the DWT.

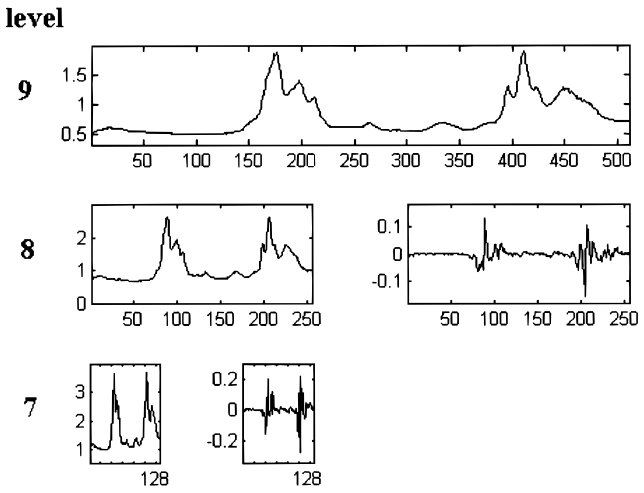

Figure 3. Top three levels of DWT of a single spectrum from data set 2. Top graph, original signal; middle, approximation and detail of original signal; bottom, approximation and detail of above approximation.

the DWT. The DWT is a basis transformation, i.e., it calculates the coordinates of a data vector (signal or spectrum) in the so-called wavelet basis. A wavelet is a function that looks like a small wave, a ripple of the baseline, hence its name. The wavelet basis is generated by stretching out the wavelet to fit different scales of the signal and by moving it to cover all parts of the signal. The DWT is said to give a time-scale, or time−frequency, analysis of signals. For near-IR data, the word time should be replaced by wavelength or wavenumber. Several families of wavelets exist, the most popular one being that of Daubechies (see, e.g., ref 7). Wavelets can differ in smoothness and so-called compactness of support, which is essentially their width. In Figure 1 examples are given of some wavelets from the Daubechies family.

A computationally efficient implementation of the DWT is Mallat's pyramid algorithm (see, e.g., ref 8). The time−frequency analysis is performed by repeated filtering of the signal. In each filter step, the frequency domain is cut in the middle using a pair of filters. The low-frequency part is usually referred to as the approximation, the other as the detail. The number of points in the signal (n) should be an integer power of two. The first step produces $n/2$ low-frequency coefficients and $n/2$ high-frequency coefficients from the raw data. In every following step, the high-frequency components are kept, and the same filters are used to further subdivide the low frequencies, until only one point remains. The process is depicted in Figure 2, where the original data are represented by the top box, and the coefficients obtained in the analysis are represented by the boxes below. On each level, the approximation is in the box to the left and the detail in the box to the right. Note that the approximations do not have to be kept, except for the last one. They can be calculated from the approximation and detail on the level below by reversing the filter operation. Also note that it is not necessary to go down to the bottom; one can stop at any level and still have a basis.

Now let us look at what is inside the boxes, as depicted in Figure 2. For this purpose, a near-IR spectrum from one of the data sets studied (set 2, see Experimental Section) is analyzed using the pyramid algorithm for the simplest filter, Daubechies filter number 1, also known as the Haar wavelet. The top three levels are given in Figure 3. The top level, being the original data, contains 512 points. Because $512 = 2^9$, this level is referred to as level 9. The first approximation and the first detail each contain half as many points, i.e., there are 256 points, which is equal to 2^8; this is level 8. The approximations are smoothed versions of the data (see Figure 4a). On each level, the detail contains the information removed by the smoothing (see Figure 4b).

The values in a box can be seen as a correlation function of the original signal and some waveform, i.e., a waveform is moved along the signal, and the cross-product of signal and waveform is calculated on every position. This is depicted in Figure 5 for the Haar wavelet, given a signal of only eight points. Each box in this figure contains one waveform on the different positions it can take. Going down one level, the waveforms double in width, and the number of positions is halved. In this particular case, the waveforms for the approximations are simple blocks. The cross-product of a signal with a block on, e.g., points 1 and 2 gives the sum of the first two points of the signal, or their average when the block sums to unity. The next position of the block will be on points 3 and 4, and the cross-product with the signal will give the sum of points 3 and 4 in the signal. By taking these averages, the signal is being smoothed, and the number of points is being reduced at the same time. For the details, the waveforms have a positive and a negative part, meaning that a difference is calculated instead of a sum. In fact, for the Haar wavelet, the details are equivalent to the discrete first derivative of the approximation on the level above.

Another interpretation of the functions plotted in Figure 5 is that they are the basis functions making up the wavelet basis. Originally, the signal is described by coordinates with respect to the canonical basis, consisting of the vectors (10000000), (01000000), ... Calculating the first approximation and detail comes down to replacing the canonical basis by the functions in the two upper boxes in Figure 5. The functions in the upper left box span a part of data space that can also be spanned by the functions in the two boxes below, and they may be replaced by those functions. Indeed, this replacement is the next step in the pyramid algorithm. The algorithm keeps replacing the basis functions in the box to the left with those in the two boxes below, until the box to the left contains only a single function that is a constant.

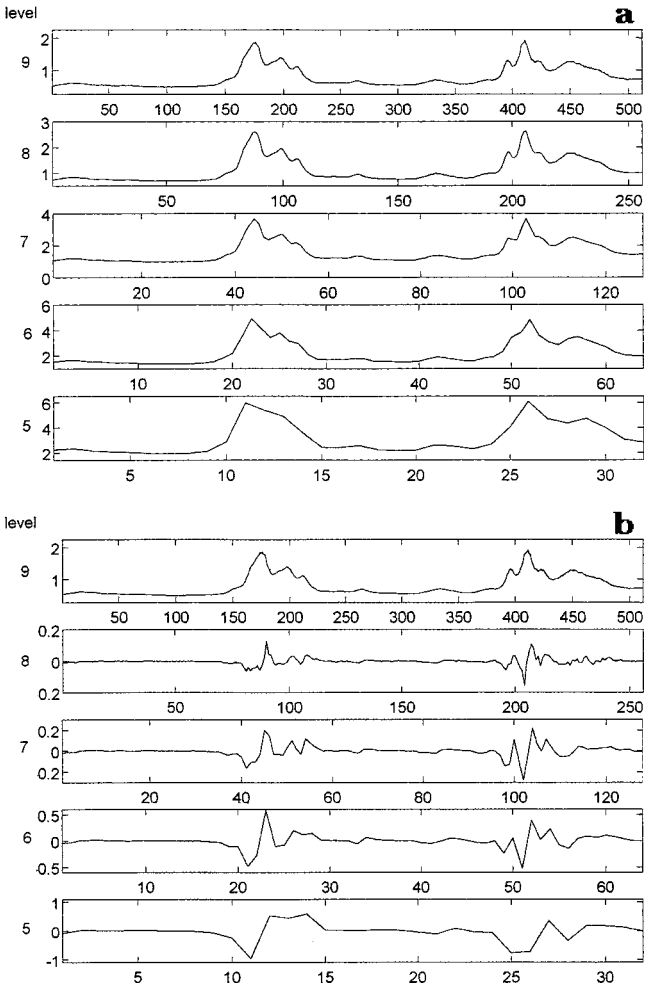

Figure 4. DWT using Haar wavelet applied to a single spectrum from data set 2: (a) approximations on levels 8 to 5 and (b) corresponding details. Level 9 represents original data.

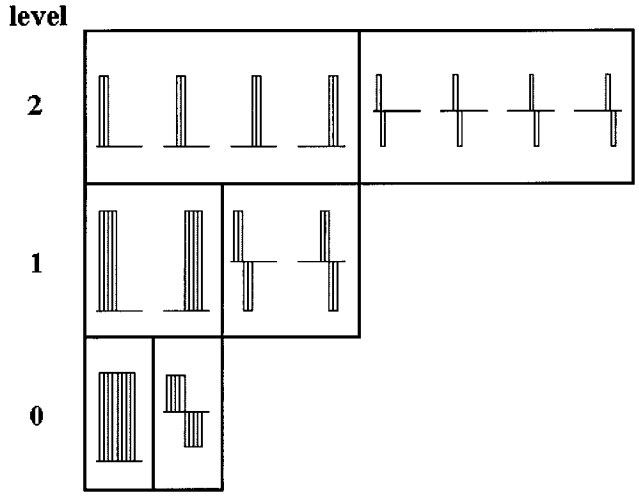

Figure 5. Waveforms for the DWT using the Haar wavelet for an 8-points-long signal.

In the pyramid algorithm, as described above, the details are not further analyzed. When we do analyze them, in the same way as the approximations, a tree of possible decompositions grows (Figure 6). The pyramid algorithm is just one branch of this tree. The full decomposition is the framework of the WPT. Just as in the pyramid algorithm, the boxes contain correlation functions of the original data and waveforms, but in the WPT,

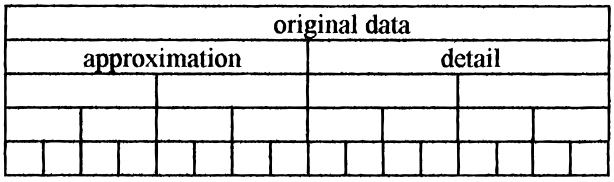

Figure 6. Scheme of the full WPT decomposition framework.

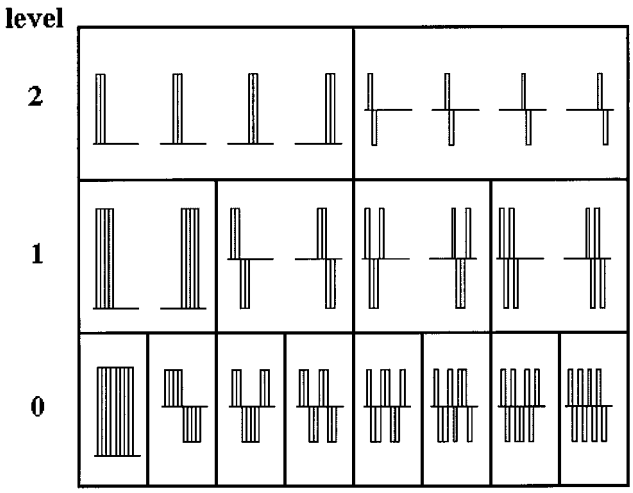

Figure 7. Waveforms for the WPT using the Haar wavelet for an 8-points-long signal.

the number of different waveforms is much larger and the shapes are more complex. The waveforms involved in the WPT of a signal of eight points using the Haar wavelet are shown in Figure 7. Compared to the pyramid algorithm, the WPT gives more flexibility. Instead of zooming in on lower and lower frequencies, it allows focusing on any part of the time−frequency domain. This aspect can be observed in the waveforms shown in Figure 7. The frequencies of the forms in the detail boxes in the bottom row increase from left to right. This illustrates that, when we turn to the right in the WPT tree, we focus on higher frequencies.

From the point of view of providing a new basis for the data, it should be realized that the full tree contains redundancy and no longer represents a basis like the pyramid algorithm does. Instead, it comprises many different bases. Each division into approximation and detail creates two orthogonal subspaces. To obtain a basis, one has to select boxes from the full decomposition of Figure 6 in such a way that the signal is covered horizontally without overlapping vertically. Examples are given in Figure 8.

An elegant way of selecting a basis from the full WPT is the best-basis algorithm, developed for data compression by Coifman.[13] To explain the principle of his approach, let us introduce the concepts energy and entropy. The energy of a signal is the sum of squares of its elements, i.e., of the coordinates of the signal in data space. If we choose a different basis for the data space, the energy of the signal will remain the same, but the distribution over its coordinates may differ. This distribution may be characterized using the entropy of the squared coordinates of the signal. When the basis is such that the energy is spread out over all basis functions, this entropy will be high. When, on the other hand, only a few basis functions account for most of the energy, the entropy will be low. For the purpose of data compression, a basis resulting in low entropy is an efficient one, since the signal may be described by a small number of features. This is the principle of Coifman's best-basis algorithm: to search for the basis

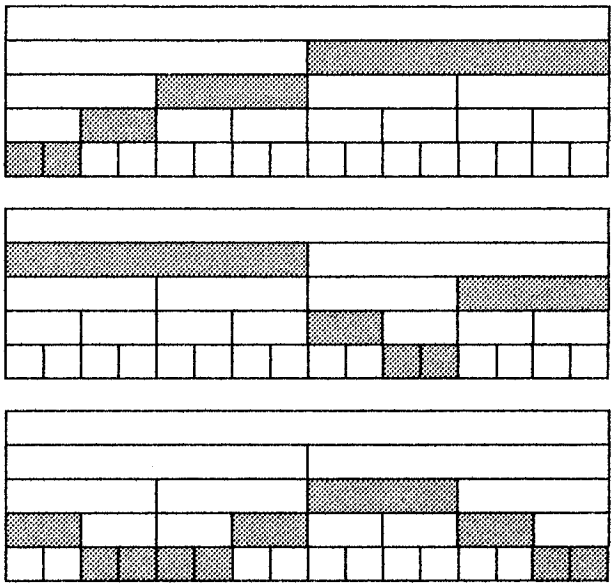

Figure 8. Examples of bases selected from the full WPT.

giving minimum entropy. The search makes use of the tree structure of the full decomposition of the signal. The entropy is calculated for each box and for the two boxes below it. When the entropy of a box is less than the sum of the entropies of the boxes below, the box is kept as part of the best-basis and not further decomposed.

In pattern recognition, we do not work with individual signals, but with groups or classes of signals. We can decompose the spectra of all objects separately and plot the results in one scheme, like the one in Figure 6. The decomposed signals can be used for pattern recognition. However, the full decompositions are larger than the original signals. Take, for instance, Figure 6, where the original data consist of 64 (2^6) variables, and the decomposition gives $6(2^6) = 384$ variables. This increase stresses the need for feature selection. Saito suggests making a preselection by first selecting boxes from the decomposition tree and then selecting features from within these boxes.The principle of the first selection is to take those boxes that are most efficient in describing the difference between classes. For this purpose, Saito has extended the best-basis algorithm to the so-called local discriminant basis algorithm (LDB).[14]

Instead of the usual entropy, which describes the distribution of a single vector, Saito uses a cross-entropy, which expresses the distribution of the difference between two vectors. In general, for two vectors **x** and **y**, the cross-entropy would be calculated as

$$E_{xy} = \sum \mathbf{x} \ln(\mathbf{x}/\mathbf{y})$$

Being a measure of difference, cross-entropy should be maximized: it is large when the difference between two vectors is concentrated in a few elements. Cross-entropy relates only two vectors. The difference between more than two vectors is characterized by the sum of the cross-entropies of all vector pairs. The vectors to be compared are the squared coordinates of the signals, described with respect to all possible bases. To discriminate between classes, Saito sums the vectors within each class and normalizes the result to account for different class sizes. The basis giving the best discrimination between the classes is the one giving the highest cross-entropy. As described before, each basis consists of a set of boxes in the decomposition scheme, and each box may be replaced by the two boxes below. Cross-entropy can be used to decide if this replacement gives a basis with better discriminating power. The LDB algorithm tests this from top to bottom.

Saito's procedure can be listed as follows:[14]

(i) For each signal, calculate the sum of squares;

(ii) for each class, sum the sums of squares (this will be the normalization factor);

(iii) decompose all signals;

(iv) calculate square of all values;

(v) calculate sum over classes and divide by the normalization factor;

(vi) for each box and each pair of classes, calculate the cross-entropies;

(vii) for each box, calculate sum of cross-entropies over all pairs;

(viii) starting from the top, compare each box with the two boxes directly below it. When the value of the top box is less than the sum of the boxes below, keep the top box and do not go further down.

Note that it is not required to do a full decomposition before determining the basis; the algorithm can also guide the decomposition itself.

EXPERIMENTAL SECTION

Four near-IR data sets were selected for the evaluation of the WPT: (1) spectra (1376−2398 nm, 2-nm bandwidth) for three groups of 20 samples, being mixtures of cellulose, mannitol, sucrose, saccharin sodium salt, and citric acid (60 spectra in total); (2) spectra (1318−2340 nm, 2-nm bandwidth) for two groups of 20 samples, being pure *p*-xylene and *p*-xylene spiked with 0.3% *o*-xylene (40 spectra in total); (3) spectra (1330−2352 nm, 2-nm bandwidth) for four groups of polymer samples of differing quality, with group sizes 22, 21, 20, and 20 (83 spectra in total); and (4) spectra (1330−2352 nm, 2-nm bandwidth) for four groups of polymer samples of differing quality, with group sizes 20, 20, 10, and 10 (60 spectra in total).

Data set 1 is a typical example of the kind of problem that may be encountered in pharmaceutical analysis. Set 2 was created in order to have clear spectral bands and to be rather noisy. Sets 3 and 4 contain the strong variation in spectral baseline that is typical for the near-IR of granular samples; they are similar and may be expected to yield similar results.

Four procedures were followed in the pattern recognition of these data sets, differing only in the pretreatment. We distinguished (i) no pretreatment; (ii) standard normal variate (SNV);[1] (iii) full WPT decomposition; and (iv) local discriminant basis.

Derivation, a procedure commonly used for data pretreatment, is not tested separately, since it is also part of the WPT using the Haar wavelet, i.e., Debauchies filter number 1.

The WPT was performed using the public domain Matlab toolbox by Taswell.[15] The WPT-based procedures were repeated for the first 10 members of the Daubechies family of wavelets in order to find the best one. The Fisher criterion, the ratio of between-group to within-group variance, was calculated for all features in the pretreated data. Only the best features, i.e., those

(15) Taswell, C. WavBox 3; Stanford University, Stanford, CA. Available via taswell@sccm.stanford.edu.

Table 1. Estimated Probability of Correct Classification for Those Combinations of Filter (Wavelet) and Number of Features Where This Value Is Maximal

data set	raw	SNV	LDB	full-WPT
1	0.658	0.927	1.000	1.000
2	0.792	0.858	0.875	0.972
3	0.924	0.969	1.000	1.000
4	0.891	0.870	1.000	1.000

Table 2. Percentage of Correctly Classified Objects for Those Combinations of Filter (Wavelet) and Number of Features Where This Value Is Maximal

data set	raw	SNV	LDB	full-WPT
1	61.7	91.7	100	100
2	75.0	87.5	92.5	97.5
3	92.8	97.6	100	100
4	90.0	85.0	100	100

with the highest value for the Fisher criterion, were used in the LDA classification that followed. It is a rule of thumb not to use more features for classification than one-third of the number of objects in the data set to prevent overfitting.[2] We kept to this rule, although leave-one-out cross-validation was used as the final stage, which should detect overfitting when it occurs. Classification was repeated while increasing the number of features from one to the maximum mentioned in order to find the optimal number. The validation results in an estimation of the predictive ability of the classification, expressed as a probability of correct classification.[16] This measure gives more detail than a percentage of correctly classified objects in case of small classes. It is calculated as

$$p = (1/N)\sum_k\sum_i \{P(w_k|x_i) - 0.5[P(w_k|x_i) + (1 - P(w_k|x_i))^2]\} + 0.5$$

where: N is the total number of objects, k is the counts over classes, i is the counts over objects within a class, w is the class, x is the object, and $P(w_k|x_i)$ is the conditional probability of object i belonging to class k, as estimated from the discriminant scores in LDA.[5]

This probability is especially useful for comparing classification procedures with similar performance, since it can take values on a continuous scale from 0 to 1, whereas percentages are always discrete values in the range from 0 to 100, with an interval depending on the number of objects in the data set.

RESULTS AND DISCUSSION

The results are presented in the Tables 1 and 2, expressed as probabilities of correct classification and percentages of correctly classified objects, respectively. For the WPT-based procedures, only the best results over all wavelet types (filters) and all numbers of features are given for each data set. The optimal combination of filter and features may differ between the two performance measures, because the numerical resolution is higher for the probability measure. Overall, however, the percentages do not deviate considerably from the probabilities, and the conclusions will not depend on the choice of the performance measure.

The use of the WPT clearly improves classification compared to both raw data and data pretreated by SNV for all data sets. No gain is brought about by the local discriminant basis approach when compared to the use of the full wavelet decomposition.

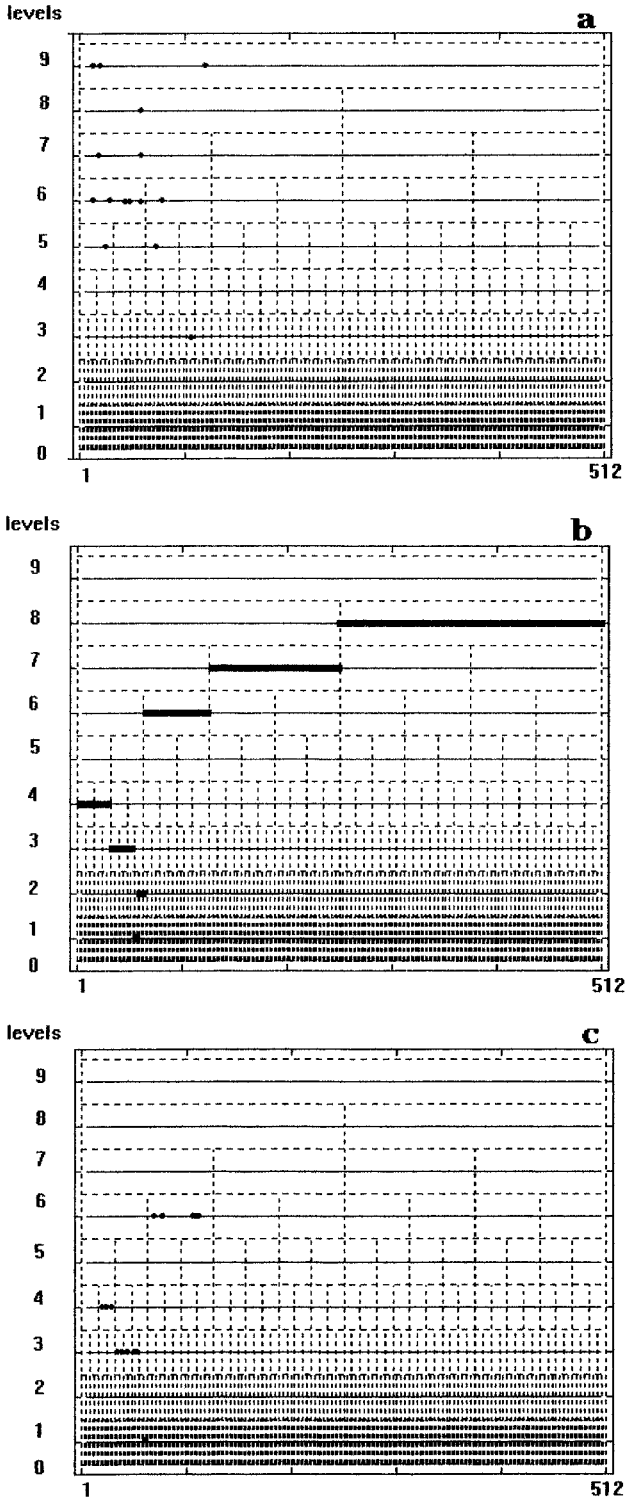

Figure 9. Application of the WPT to data set 1. Observations for the filter (wavelet) from the Daubechies family that was found to give the best results: (a) features selected from the complete WPT framework; (b) the local discriminant basis; and (c) features selected from the LDB.

(16) Hilden, J.; Habbema, J. D. F.; Bjerregaard, B. *Methods Inf. Med.* **1978**, *17*, 238−246.

We found no apparent logic in the features obtained from the different filters giving the best results. We expected that, for the same data set, all filters would focus on the same features in the time−frequency domain, one filter with more success than another. It was observed that all features were found in roughly the same region of the time−frequency domain, representing low frequencies of intermediate localization, but not on exactly the same positions. This suggests that, when you know more about a data set, you can direct the search by selecting a time−frequency region of interest. The question is how such knowledge should be acquired. Something like an objective time−frequency analysis would be useful.

Figure 9 is given as an example of the location of features on the time−frequency map. Many features fall outside the DWT basis, showing the added flexibility of the WPT. Comparing the LDB features (Figure 9c) with those obtained using the full decomposition (Figure 9a), it is observed that features are not necessarily located within the LDB (Figure 9b). This sheds some doubts on the justification of the LDB. For the univariate feature selection that follows, it is an important shortcoming that an orthogonal basis does not guarantee orthogonal features.

In this work, we focused on the usefulness of the local discriminant basis approach to using wavelets in pattern recognition. Wavelets were used because the features in the spectra on which the classification has to be performed are local, and wavelets are considered an efficient way of representing local phenomena. For the data sets investigated, wavelets have proved to be a reasonable approach. However, we are not able to specify the exact conditions under which wavelets may be expected to give good results. A comparison with global methods such as Fourier transformation should be performed to obtain this knowledge.

CONCLUSIONS

Use of the wavelet packet transform for preprocessing near-IR spectra improves the classification when compared to using either SNV or no pretreatment at all. Selecting features from a local discriminant basis instead of from a full decomposition does not improve the results.

Received for review November 2, 1995. Accepted February 22, 1996.[⊗]

AC951091Z

[⊗] Abstract published in *Advance ACS Abstracts,* April 1, 1996.