ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/286568486

Wavelets in chemistry and cheminformatics

Article *in* Reviews in Computational Chemistry · January 2006

CITATIONS
11

READS
3,464

5 authors, including:

C Matthew Sundling
Oregon Health and Science University
8 PUBLICATIONS **166** CITATIONS
SEE PROFILE

N. Sukumar
Amrita Vishwa Vidyapeetham
100 PUBLICATIONS **1,331** CITATIONS
SEE PROFILE

M. Embrechts
Rensselaer Polytechnic Institute
188 PUBLICATIONS **4,351** CITATIONS
SEE PROFILE

Curt M Breneman
Rensselaer Polytechnic Institute
136 PUBLICATIONS **9,334** CITATIONS
SEE PROFILE

All content following this page was uploaded by N. Sukumar on 14 December 2017.

The user has requested enhancement of the downloaded file.

Wavelets in Chemistry and Cheminformatics

C. Matthew Sundling, Nagamani Sukumar, Hongmei Zhang, and Curt M. Breneman*
Department of Chemistry and Chemical Biology
Center for Biotechnology and Interdisciplinary Studies
Rensselaer Polytechnic Institute
110 8th St, Troy, NY 12180

and

Mark J. Embrechts
Department of Decision Science and Engineering Systems
Rensselaer Polytechnic Institute
110 8th St, Troy, NY 12180

*Corresponding author

Preface

Wavelet transform methods developed quickly during the 1990s, and have since become widely used in various fields of science and engineering, including a number of important applications in chemistry. The ability of wavelet methods to rapidly dissect signals into meaningful components makes them invaluable tools for data analysis and information compression. Unlike traditional signal processing methods, the wavelet transform offers simultaneous localization of information in both frequency and time or property domains. This makes the wavelet transform powerful in its ability to succinctly distill the details of complex and irregular property distributions or waveforms into meaningful and simple components. Consequently, wavelet transform methods are well suited to processing experimental data collected throughout the various areas of chemistry, as well as leading to new types of computationally-generated molecular property descriptors. The diverse utility of wavelets has caused an explosion of application papers covering many areas of chemical analysis, most of which land squarely in the realm of chemometrics or chemical spectroscopy, but significant applications in quantum chemistry, and recently, cheminformatics and computational chemistry show how wavelets can be applied in any situation where data analysis is needed. Wavelet transform methods are a proven technology in signal cleaning and signal feature isolation, and have provided chemists with better methods for analyzing and understand their experimental data—by distilling *chemical information* from raw experimental or computational data. This chapter is a pedagogically driven overview of basic wavelet transformation methods and their applications in chemistry and cheminformatics.

Introduction to Wavelets

We provide here a brief introduction to the concept of wavelets, intentionally glossing over much of the technical and mathematical details in favor of conveying a simple and conceptual perspective of wavelet techniques. For a thorough introduction to the theory and history of the wavelet transform, readers should examine the pertinent literature.[1-8] Other discussions of wavelets and their applications in chemistry may be found in references 9-17. The predecessor of the wavelet transform (WT) is the Fourier transform (FT), which has its own successful history of analyzing chemical spectra. In this section, we discuss the fundamentals of the Fourier transform in order to illustrate the advantages and disadvantages of this technique relative to wavelets and to show how the wavelet transform picks up where the Fourier transform leaves off. We also show that the wavelet transform acts as a "mathematical microscope" to reveal and focus on spectral features that are buried in the original signal but which are ignored by the Fourier transform. Wavelets thus allow for a thorough examination of the character of a signal, including high frequency noise, asymmetric broad regions, short-term spikes and other features of interest.

Throughout this chapter we will describe how wavelets can be used to analyze, clean and encode molecular information in a dense and usable format, and show how wavelets provide a useful and stable means for representing molecular electronic property distributions for use in Quantitative Structure-Activity Relationships/Quantitative Structure-Property Relationships (QSAR/QSPR) modeling. Before that, however, some history of their place in signal analysis is appropriate.

Like the Fourier transform, the wavelet transform converts a signal from its normal time- or property-domain representation into another representation—in wavelet space—which reveals the frequency content of the original signal. A wavelet space representation not only separates frequency components, but unlike the Fourier transform, also gives their exact position and identifies their effective domain. This allows wavelet transform methods to separate, isolate and analyze the individual components of a signal. The Fourier and wavelet transforms are useful when analyzing many different types of chemistry data, regardless of the domain to which they belong. Even though many important chemical "signals" exist within a variety of domains, we will discuss these methods in terms of time-domain signals, as well as their resulting frequency-domain and wavelet-domain transforms.

1. Fourier Transform

A time-domain signal contains a relationship between temporal information and amplitude information, but gives no explicit frequency information. The key assumption in Fourier analysis is that a signal can be considered a composite of sinusoidal components, each having a specific frequency. The individual components are convoluted together in the original signal and are therefore generally immune to interpretation by inspection. Fourier transformation (FT) converts the signal from a time-

domain into a frequency-domain, giving us access to frequency and amplitude information, as illustrated in Figure 1. FT is a reversible process and gives two entirely different perspectives of the same data. However, we cannot get both the time domain and frequency domain information simultaneously, which is often critical for a variety of data analyses.

[FIGURE 1 HERE]

2. Continuous Fourier Transform

The continuous Fourier transform (CFT) of a real or complex continuous function is defined as:

$$\ddot{f}(\omega) = \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt \qquad [1]$$

and the inverse transform is defined as:

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \ddot{f}(\omega) e^{+i\omega t} d\omega \qquad [2]$$

where:

$$e^{-i\omega t} = \cos\omega t + i \sin\omega t \qquad [3]$$

In Equations [1-3], function $f(t)$ is a continuous time-domain function, that is transformed into the frequency domain function $\hat{f}(\omega)$.[1, 4] Since our basis functions are sinusoidal, as defined by Equation [3] and illustrated in Figure 2, a single component with frequency ω will affect the entire domain of the signal $f(t)$ equally. This makes the Fourier transform not entirely suitable for analyzing non-stationary (non-periodic) functions (as depicted in Figure 3). Isolation of a given frequency component to a finite time region is necessary when dealing with localized signal features (such as a sharp spike) which are, by nature, non-stationary. The Fourier basis function entirely lacks the ability to distinguish or isolate time-domain information.

[FIGURE 2 HERE]

[FIGURE 3 HERE]

3. Short Time Fourier Transformation

Given the limitations of the Fourier transform, some approximations are needed to handle non-stationary signals. The discrete Fourier transform (DFT) and the short-time Fourier transform (STFT, a.k.a. the Gabor transform) are two alternative transformation methods that address this issue.[1, 3-5, 9, 18] In the mid-20th century, Jean Ville pointed out that two

basic approaches to time-frequency analysis exist, but both attempt the same thing - to create a pseudo-stationary signal from a non-stationary one.[1] One approach filters different frequency bands, and then splits these bands into pieces and analyzes their energy content. The other approach splits the signal into equal length sections in the time domain and then examines these pieces individually for their frequency content. The DFT uses angular sampling to isolate frequency bands, while STFT uses windowed-time frequency analysis.

In short-time Fourier transformation, a non-stationary signal is divided into small windows in an attempt to achieve a locally stationary signal, as depicted in Figure 4. Each window is analyzed using a normal FT to obtain the local frequency composition of a signal. The window length selection is crucial because it ensures that the local signal is sufficiently stationary. Selection of the STFT window width involves a trade-off however: Narrowing the time window improves time/location resolution, but reduces clarity of the frequency information. Naturally, the STFT becomes an FT if the window length is taken as infinity. Although the STFT procedure manages to garner some signal position information via the location of the windows, it is still fundamentally hindered by the stationary nature of the sinusoidal basis of the Fourier transform method. A transformation method that uses basis functions that are *simultaneously* localized in both the time and frequency domains would neatly avoid this trade-off. The wavelet transform procedure does exactly this.

[FIGURE 4 HERE]

4. Wavelet Transform

The wavelet transform provides an entirely new perspective on traditional signal processing techniques (i.e., Fourier transformation methods) for breaking up a signal into its component parts. Literally, the term "wavelet" means little wave. More specifically, a wavelet is a function that satisfies the following two conditions: (1) It has a small concentrated finite burst of energy in the time domain, and (2) it exhibits oscillation in time.[2] The Daubechies 6 wavelet, illustrated in Figure 5, clearly exhibits these characteristics. The first condition makes the wavelet "little" in the sense that it is well localized in the time domain. The second condition makes it periodic, giving it some wave-like character.

[FIGURE 5 HERE]

Given a particular wavelet function (a particular basis function), the wavelet transform operates in a manner similar to the Fourier transform by using its basis function to convert a signal from one domain to another. The wavelet transform deconstructs a signal by using dilated (scaled) and translated (shifted) versions of the basis function. The attraction of wavelet transformation techniques over Fourier transform techniques is that they localize information effectively in both the time and frequency domains simultaneously.[6, 9] Because of this, the wavelet transform is an ideal tool for analyzing non-stationary signals – such as spectra or molecular property distributions.

Wavelet transform methods can be categorized into two main classes: (1) continuous wavelet transforms and (2) discrete wavelet transforms. Each is discussed below.

5. Continuous Wavelet Transform

The continuous wavelet transform is defined as:

$$CWT\left[f\left(x\right)\right] = \int_{-\infty}^{+\infty} f\left(x\right)\Psi^{*}_{a,b}\left(x\right)dx \qquad [4]$$

and the inverse transform is defined as:

$$f\left(x\right) = \frac{1}{C}\int_{0}^{+\infty}\int_{-\infty}^{+\infty} CWT\left[f\left(x\right)\right]\Psi_{a,b}\left(x\right)\frac{dadb}{a^{2}} \qquad [5]$$

where:

$$\Psi_{a,b}\left(x\right) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{x-b}{a}\right) \quad , \; a,b \in \mathbf{R}, \;\; a \neq 0 \qquad [6]$$

$$C = \int_{0}^{+\infty} \frac{\ddot{\psi}^{*}\left(\omega\right)\ddot{\psi}\left(\omega\right)}{\omega} d\omega \qquad [7]$$

$\Psi_{a,b}\left(x\right)$ is a dilated and translated version of the "mother wavelet" $\psi\left(x\right)$, where a is the scale, b is the translation, $1/\sqrt{|a|}$ is a normalization term, and * symbolizes complex conjugation. Equation [7] is the admissibility condition that gives C as a positive real number if the mother wavelet satisfies certain conditions (e.g., the integral of the mother wavelet equals zero).[1, 3-5, 9, 11, 18] There is no restriction on the choice of a and b in the continuous wavelet transform (other than $a \neq 0$), which means that the choice of $\Psi_{a,b}\left(x\right)$ is continuous in the time-scale domain. The wavelet can be positioned anywhere and scaled to any value for optimal fitting of the signal $f\left(x\right)$. The admissibility condition requires that only wavelets of certain character are capable of the *reverse* wavelet transform as described by Equation [5]. It does not restrict the wavelets capable of the forward wavelet transform. In fact, many useful wavelet applications use wavelets that make the reverse transform impossible.

While the Fourier transform separates the signal into a series of sine waves of different frequencies, the wavelet transform decomposes the signal into wavelets – dilated and translated versions of the "mother" wavelet. Compared to a smooth and infinite sinusoidal wave function, the wavelet function is irregular in shape and compactly supported (i.e., has a limited domain where the function is nonzero). These properties make wavelets an ideal tool for analyzing non-stationary signals of finite length or duration; their irregular shape enables them to characterize signals with discontinuities or sharp changes, and their compact support enables them to represent signals with temporal

or regional features. Figure 6 depicts how wavelets can fit and decompose a signal. By a variety of large and small dilations the wavelets can be fit into the various kinks, shoulders, arcs and spikes of a given signal. Scaling analysis allows us to process signals at different scales and resolution, elegantly revealing aspects of the signal that would be masked by the regularity of a sinusoidal wave. In fact, whereas the STFT attempts to force sinusoidal functions onto a time-localized section of the signal, wavelets scale naturally to represent a variety of different regions within a given signal.

[FIGURE 6 HERE]

The continuous wavelet transform is compactly described by Equations [4] and [6], but this definition allows for infinitely redundant transformations.[4, 19] There is no limit to the number of dilated and translated wavelets ($\Psi_{a,b}(x)$, where a and b are real numbers) used in the transform. This unrestricted and unguided use of wavelets to convert a signal into wavelet space often prevents the use of an inverse wavelet transformation because of violations of the conditions required by Equation [7]. Even though these transforms are redundant and non-reversible, they still reveal information about the character of a particular signal.

Techniques exist for reducing the redundancy in the CWT. These techniques isolate the dilated and translated wavelets to form the signal's "skeleton". The skeleton wavelets consist of important features found in the original signal, called ridges, which are the defining curves of the waveform. By focusing and restricting the wavelets to only these critical features redundancy is removed. Such small set of independent wavelets make these CWT adaptations practical, informative and manageable.[6, 19, 20] Additionally, placing very specific restraints on the selection of wavelet basis and on the dilation values (i.e., the a values) eliminates all redundancy in the transforms and gives rise to a fast and useful transformation method—the discrete wavelet transform (DWT).

6. Discrete Wavelet Transform

The main differences between the discrete wavelet transform (DWT) and the continuous wavelet transform are specific requirements for the mother wavelets and the allowable dilation and translation values. Explicitly, Equation [4] becomes:

$$DWT\left[f(x)\right] = \int_{-\infty}^{+\infty} f(x)\Psi^{*}_{a,b}(x)dx \quad , \quad a = 2^{i},\ i \in \bullet\ ,\ b \in \bullet \qquad [8]$$

where the dilation variable a is restricted to powers of two (i.e., $a = 2^{i}, i \in \{0,1,2,\mathsf{K}\}$), and our translation variable b is a whole number (i.e., $b \in \{0,1,2,\mathsf{K}\}$). These values are called dyadic dilations and translations. In principle, the CWT, with no restriction applied to the choice of these two coordinates, maps the entire "wavelet space" (i.e., the (a,b) plane). The DWT confines us to specific non-redundant regions of the wavelet space.

Rather than having a continuum of wavelet dilations as in the CWT, the discrete wavelet transform uses discrete dilations that can be thought of as filters of different scales. These act as cutoff frequencies to divide the signal into different frequency bands.[21] Wavelets are actually a pair of filters, called the wavelet and the scaling function, as depicted in Figure 7. To separate each frequency band, the wavelet is used to isolate the information pertaining to that band, and the scaling function separates out everything else. This process is part of a method called multi-resolution analysis.[7]

[FIGURE 7 HERE]

Multi-resolution analysis[3, 4, 7, 19] is the most commonly userd DWT method and uses a hierarchy of low- and high-pass filters to successively separate the finer details from the remainder of a signal. The term "DWT" is often used to refer specifically to multi-resolution analysis as implemented by the *pyramid algorithm*. Within the pyramid algorithm, the original signal is decomposed successively into components of lower frequency while the high frequency components are not analyzed further as illustrated in Figure 8.[7] The analysis begins with two complementary filters, one low-pass and one high-pass, to separate the high frequency details from the rest of the signal. Then, with each transformation step, increasingly coarse information is separated from the remaining portion of the signal (see Figure 9). The maximum number of dilations/separations that can be performed is determined by the input size of the data being analyzed. Specifically, given a signal represented by a series of 2^N values, the signal is divided into N discrete levels of detail. The transformation turns the original signal into a set of 2^N wavelet coefficients, which reflect the individual contributions of their associated scaled wavelet. The original signal is really a combination of the dilated and translated wavelets as prescribed by the wavelet coefficients.

[FIGURE 8 HERE]

[FIGURE 9 HERE]

The restrictions placed on the mother wavelets for multi-resolution analysis do not limit the variety of shapes that can be used as mother wavelets; different researchers have proposed a number of different wavelet functions, each with benefits and drawbacks.[3] The wavelet shape trade-off is between how compactly it can be localized in space and its level of smoothness. For example, the Haar wavelet, which is the simplest wavelet and was identified almost 100 years ago,[22] is well localized in space, but has an 'unnatural' square-wave oscillation (see Figure 10). There exist many related wavelets, collectively referred to as wavelet families;[6] some of these families include the Meyer wavelet, Coiflet wavelet, spline wavelet, orthogonal wavelet, symmlet wavelet, and local cosine basis. Figure 10 depicts several of these wavelets and illustrates some of their features. Depending on the shape of the original signal and how it is being analyzed, some wavelet functions will outperform others. Wavelet selection is, consequently, an important facet of wavelet analysis. For example, a wavelet basis with sharp, narrow peaks may be better suited to transform and characterize a particularly disjointed signal, whereas a simple,

gentle wavelet may be more appropriate for a smoother signal. Selecting an optimal wavelet adapted to a particular type of signal allows for finer, more economical separation of signal features. In a general sense, an optimal wavelet basis function would concentrate the signal features to a small number of large-valued wavelet coefficients. This leads to interesting and useful applications of the wavelet transform, such as signal compression and feature isolation, both of which are important in various chemical applications and are examined further in following sections.[15]

[FIGURE 10 HERE]

7. Wavelet Packet Transform

The wavelet packet transform (WPT) is a generalized version of the pyramid algorithm[8, 23] in which the signal is successively separated into the low-pass information and 'the rest' of the signal. In contrast to the pyramid algorithm used in DWT, both the low-pass and the high-pass information are iteratively transformed in WPT, creating a complete tree as compared to the single branch enumerated by the pyramid algorithm. Comparing Figures 11 and 8 illustrates this difference. As mentioned earlier, optimal wavelet function selection is important for obtaining an efficient representation of the signal in wavelet space. The wavelet packet transform takes this notion of efficiency even further by offering the flexibility of choosing the final signal representation, which is also known as the 'signal basis'.[24] With the full hierarchy evaluated, we have a choice of combining different levels of high-pass and low-pass filtering, so we can select a signal representation that is most suited to our needs.

[FIGURE 11 HERE]

Selection of the "best" basis or representation from the WPT hierarchy means choosing a combination of orthogonal, non-redundant coefficients from the different scale levels of the WPT. An entropy-based algorithm developed by Coifman and Wickerhauser[24] provides a quantitative criterion for selecting a basis for signal representation at the lowest information cost. While the pyramid algorithm blindly follows a single branch of the WPT tree and only computes a single combination of the many possible bases (see Figure 12), the WPT procedure affords the flexibility of choosing the optimal signal representation for any application.

[FIGURE 12 HERE]

Wavelets vs. Fourier Transforms: A Summary

The ability of wavelet transformations to resolve a signal into its component features makes it useable for many practical applications. The Fourier transform, while revealing the frequency characteristics of a signal, is limited by the assumption that all signals are stationary (periodic). This, in turn, provides no resolution in the time-domain. The power

of the wavelet transform is in its ability to succinctly distill the details of complex and irregular signals into meaningful and simple components – in essence, to encode both the time (or any domain) and frequency information simultaneously.

Application of Wavelets in Chemistry

1. Smoothing and Denoising

Experimental chemistry is full of situations where data needs to be cleaned of its noise, background or other confounding features. Experimental data always contains noise from various sources, and it must be isolated, understood, and usually removed before effective data analysis can proceed. The goal of smoothing and denoising techniques is to separate useful from useless information. These techniques can be tailored to remove the noise from spectra, improve the signal-to-noise ratio in analytical images, improve resolution of peaks by removing background signals, and to give clarity and focus to numerous types of data with different kinds of problems. While popular cleaning techniques such as Savitzky-Golay smoothing and Fourier filtering[25-28] could be employed for these tasks, the WT-based signal cleaning techniques have grown in popularity because of their efficacy and utility in handling noise.

In general terms, smoothing and denoising are used to isolate and remove noise or background signals, and to resolve features and peak shapes.[29] Smoothing and denoising are different, albeit related: *smoothing* is the process of removing the high frequency components of a signal, regardless of their amplitude, and *denoising* is the process of removing low-amplitude components, regardless of their frequency.[25] All types of signal noise exist in experimental data, the most common of which can be found in the high-frequency range, especially when considering either time-based data or spectral data. Background signals, another type of noise, are often found in the low frequency range. Though knowing the location (frequency and position) of noise in a signal is important, it is also important to understand the type or character of the noise. Heteroscedastic noise, or noise of changing variance, must be diagnosed and is usually handled differently from the more common homoscedastic noise. The treatment of heteroscedastic noise is a challenging problem for any signal smoothing technique, including WT methods. An example of heteroscedasticity is when the variance of noise increases as the overall signal strength increases.

Wavelet transform methods, particularly multi-resolution analysis (MRA) methods, decompose a signal into frequency bands. That separation allows for the targeting of frequencies of particular interest or, in the case of noise, disinterest to the researcher. A signal may be considered to be a combination of component wavelets, so, reducing the contribution of 'noise wavelets' will result in a clean, smooth signal. The algorithm for smoothing a signal (see Figure 13) consists of four steps: 1) transform the signal, 2) isolate the wavelet coefficients corresponding to the high frequency components, 3) 'zero-out' or reduce these coefficients, and 4) apply a reverse wavelet transform to the signal. The final smooth signal will have all the original features of interest without the

high frequency noise. Figure 13 illustrates the smoothing routine using the pyramid algorithm, but the routine is easily modified for use with the wavelet packet transform (WPT), often yielding better results.[23] WPT not only separates the signal into frequency bands for noise isolation, but also gives the optimal basis for signal representation. This means our signal is represented by the most sparse, compact basis of wavelet coefficients, focusing the signal features into a few large wavelet coefficients. This allows for confident and localized modification of frequency band information, with improved noise suppression and minimal cross-band effects when removing small wavelet coefficients.

The denoising algorithm is similar to the smoothing algorithm, except that in step 3, the small-amplitude coefficients are targeted for removal regardless of frequency (illustrated in Figure 14) – this is sometimes referred to as wavelet thresholding.[29] WT denoising methods are able to retain the interesting and often subtle features normally destroyed by other aggressive smoothing techniques, including maintaining edge sharpness and peak shapes.[30] Donoho formalized the process of denoising and showed that if *all* wavelet coefficients are moved towards zero by proportional amounts, the resulting signal will then be similar to the original signal.[31] This approach virtually eliminates the small wavelet coefficients (and their encoded noise information), while retaining the important wavelets that contribute most to the shape of the original signal. Although there exist variations of basic smoothing and denoising methods tailored to specific types of noise,[32] the core concept of wavelet smoothing and denoising techniques remains the same.

[FIGURE 13 HERE]

[FIGURE 14 HERE]

Variations of the wavelet smoothing and denoising routines described above have been thoroughly investigated and tested by Mittermayr et al. against more traditional methods, such as Savitzky-Golay smoothing and Fourier filtering.[27] Other investigators were rigorous in their evaluation of wavelet transform smoothing techniques; they used the best alternative smoothing techniques available, and evaluated optimally selected mother wavelets tailored to different types of data.[25, 26, 28, 33] Their analysis showed good results for wavelet denoising, outperforming the more traditional techniques on a variety of experimental data. There is an extensive list of publications within the chemistry literature, reviewed in references 9-10 where WT cleaning of data has improved data analysis. In all cases, the goal of smoothing or denoising a signal was to simplify and clarify experimental data. Other examples of wavelet-based smoothing and denoising may be found in a multitude of applications including chromatography,[17, 26, 34, 35] IR spectroscopy (with and without heteroscedastic noise),[33, 36-38] UV-VIS spectroscopy,[39] mass spectrometry,[30, 40-43] voltammetry,[39, 44-46] capillary electrophoresis,[28] molecular superposition methods,[47] and photoacoustic spectroscopy.[48]

Heteroscedastic backgrounds can confound signals in a way that simple wavelet thresholding routines become ineffective at removing their influences. The changing variance of this noise allows the noise to move from one frequency band to another. Simple wavelet thresholding of the DWT is not sufficient for handling noise that moves

between frequency bands, but WPT methods allow for intelligent isolation of the background. WPT denoising techniques were used to remove heteroscedastic noise from near-IR spectra[38] and from GC/MS spectra;[43] in both cases the resolution of overlapping peaks having low signal-to-noise ratio were improved. Although the continuous wavelet transform can isolate the heteroscedastic noise, it is a cumbersome approach and is not often implemented in practice.[38]

2. Signal Feature Isolation

Feature isolation is a general form of signal analysis. It concerns the broader problem of understanding and quantifying a signal and its component features rather than removing one particular type of feature from a signal (e.g., noise). These features often include the location and size of sharp spikes, critical points, smooth regions, discontinuities, and frequency composition.[49-54]

Multi-resolution analysis (MRA) techniques, which include the DWT and WPT methods, effectively divide a signal into frequency bands using wavelet functions of known position, which give the precise locations of all of the signal features in wavelet space. Knowing the locations of specific components of the signal allows them to be analyzed, enhanced, cleaned or removed. MRA methods have the additional advantage of giving an orthogonal representation of the original signal, which allows for local modifications or 'feature selection' without introducing the global changes that could otherwise affect the signal. Both the smoothing and denoising routines are examples of this application.

Although MRA methods are useful in many applications, the continuous wavelet transform (CWT) is often more accurate for signal characterization. The CWT does not enjoy the benefit of orthogonality, but, it is not restricted in its placement of wavelet functions during the transform. The CWT can use optimally dilated and translated wavelets to represent every crevice and region of a signal to locate features of interest precisely. This allows the CWT to recognize, with great precision, where important signal shape features occur, such as peaks, kinks, smooth regions, and edges. For example, signal critical points can be located, dramatic changes in frequency content can be identified, and the behavior of the derivative of a curve can be characterized. The penalty for not having an orthogonal representation of the signal is that CWT-based methods are not able to manipulate signals as conveniently as can MRA methods: nonetheless the redundant signal representation inherent in CWT methods increases the precision and resolution of the signal analysis. In general, MRA methods are used to clean, change or clarify a signal, while CWT methods are used to perform precise and thorough signal diagnostics.

Wavelet transform-based methods for signal isolation and analysis have improved significantly the deconstruction, quantitative analysis and cleaning of many types of experimental data[55, 56] ranging from chromatography,[57-64] IR spectroscopy,[65-72] UV-VIS spectroscopy, [67] mass spectrometry,[73] x-ray absorption,[74, 75] NMR spectroscopy,[76-80] and in other studies.[81, 82] Other interesting applications use methods of wavelet-based image

fusion[52, 83-87] to combine data obtained from different sources for better analysis and enhanced information extraction.[88-90]

3. Signal Compression

The purpose of data compression is to reduce storage space and to concentrate signal information. The central concept of wavelet compression is to represent data in an optimal manner, so that only a small number of wavelet coefficients are needed to capture a majority of the original signal.

Many specific compression routines using the wavelet transform exist, but all are similar in that they move a signal from a higher dimensional space to a lower dimensional subspace. Compression takes advantage of the wavelet transform because it is inherently able to focus the main features and overall shape of a signal into a relatively small number of wavelet coefficients. Data compression using the wavelet packet transform (WPT)[8] starts by separating a signal into its frequency bands using multi-resolution analysis. The best signal basis is then, selected such that it minimizes the number of wavelet coefficients, which best represents the original signal. The best-basis selection can be achieved using a variety of optimizing functions. Normally an information entropy metric is used to guide the selection of a maximally representing but minimum-length basis.[91] Compression is achieved when fewer variables or data values are required to represent the original signal. The selection of the best basis is usually optimized for each individual signal, but there is an advantage to selecting the best basis for a set of signals simultaneously,[8, 92] especially when creating a library of similar data, such as chemical spectra or molecular property information.

Data compression is a useful, if not novel, application of wavelets in chemistry. It has found great use in spectral compression and storage and stands as an important application for data archiving. WPT-based data compression is useful for archiving infrared spectral libraries,[93] specifically when a large amount of experimental data must be quickly and accurately processed and stored. IR spectra contain sharp peaks and features, making them well suited for compression by unsupervised DWT with basic wavelet-thresholding techniques. This is basically a denoising approach to compression, in that removing the noise coefficients saves storage space.[94]

A second approach to data compression is to compress IR spectra using a construct called a wavelet neural network (WNN).[95] The WNN approach stores large amounts of IR data for fast archiving of spectral data. This is achieved by modifying the machine learning technique of artificial neural networks (ANN)[96] to capture the shape of IR spectra using wavelet basis functions. The WNN approach is similar to another approach used to store UV-VIS spectral information that is presented in.[93] Though the ANN approach is very effective for storing spectral data, it is limited to operating on an expected type of data (i.e., IR spectra), and is not a general compression approach for any free form signal, image or otherwise.

4. Quantum Chemistry

Wavelet transform methods are flexible, robust and particularly useful for representing complex and intricate functions, making them an interesting alternative to other, more common basis sets for representing molecular wave functions. Quantum chemical applications typically involve the computational evaluation of many-electron molecular wave functions (or electron density distributions), through approximate solutions of the Schrödinger equation. Semi-empirical and *ab initio* quantum mechanical methods are commonly used to compute electronic structure and molecular properties derived from the wave function, as well as spectroscopic[97] properties arising from transitions among energy levels.[98-100]

The electronic wave functions of molecules vary much more dramatically near atomic nuclei than in the inter-atomic spaces.[101] This makes the computational cost of maintaining high-resolution wave functions near the nuclei much higher than when of representing valence regions further away from any nuclei. In traditional *ab initio* approaches for representing electronic structure, a molecular wave function is expanded in a series of basis functions, typically as a linear combination of Gaussian functions. Because the expansions are uniform, adjusting the approximation to give improved resolution near atom centers requires a dramatic increase in the number of terms through the introduction of double- and triple-zeta basis sets.[102] An alternative approach for increasing local resolution in highly non-linear wave functions is to use specialized basis sets that depend explicitly on their locations relative to nuclear positions. The spatially-localized property of wavelets can be useful for creating a consistent basis function for electronic distribution calculations.[103, 104] Wavelets are able to characterize the highly non-linear wave functions encountered in quantum chemistry because they are able to adjust themselves to fit widely varying non-stationary functions. [105] Wavelets have also helped to create well-behaved and consistent descriptions of the properties of electron density distributions.[106]

Iyengar and Frisch[107] have demonstrated the fundamental equivalence between the wavelet theory of multi-resolution analysis and the translation and dilation operations on the primitive Cartesian Gaussian basis functions used in electronic structure theory:

$$\chi^{R}_{l,m,n}(r) = (x - R_x)^l (y - R_y)^m (z - R_z)^n e^{-\alpha(r-R)^2} \quad , \quad l,m,n \in \bullet \qquad [9]$$

where

$$R = \{R_x, R_y, R_z\} \qquad [10]$$

The positive integers l,m and n determine the orbital angular momentum of the basis function, and R is the Gaussian center. Thus the Gaussian basis function χ has a translation property represented by its dependence on the position of atom center R as well as a dilatory property wherein for any given value of l,m,n and R, the exponent α may have multiple values; the primitive Gaussians with smaller α being simply dilated versions of the original – see Equation [6] and Figure 6. These authors have shown that primitive Gaussians are, in fact, multi-wavelets with non-integral scale factors. The Gaussian multi-wavelet basis is non-orthogonal on account of these non-integral scale

factors, giving rise to different levels of basis set completeness at different molecular geometries. Hence the quality of a Gaussian basis set changes as the nuclei move – a well-known artifact of *ab initio* quantum chemistry known as the basis set superposition error (BSSE).[108] This analysis provides new ways to ascertain and control the quality of a basis set during *ab initio* molecular dynamics.

Johnson and coworkers[109] have illustrated how wavelets can improve both resolution and accuracy over traditional *ab initio* methods. A three-dimension wavelet analysis was used for electronic structure calculations by Arias, Cho and coworkers.[97] They took advantage of the stable nature of the wavelet transform to provide a systematically improvable and tractable description of electronic wave functions, thereby overcoming some of the limitations of conventional basis set expansions. This was demonstrated by computing the 1s states for all the naturally occurring nuclei in the periodic table from hydrogen to uranium, as well as their interaction energies with the hydrogen molecule ion. Another study investigated position and momentum information from solutions of the Hartree-Fock equation and found wavelet-based analysis provided more information concerning the oscillatory nature of a time-dependent wave function than did traditional Fourier transform approaches.[110] These authors found that wavelets were valuable for improving current methods for total energy calculations.[111] Wavelets have also been used to improve empirical force field representations for studying the behavior of biological macromolecules.[112]

Harrison, Yanai and coworkers have reported an efficient, accurate multi-resolution solver for the Kohn-Sham[113] and Hartree-Fock[114, 115] self-consistent field methods for general polyatomic molecules. The Hartree-Fock exchange is a non-local operator, whose evaluation has been a computational bottleneck for electronic structure calculations, scaling as $\mathrm{O}\left(N^{3-4}\right)$ for small molecules and no better than $\mathrm{O}\left(N^2 \log N\right)$ for larger systems. While earlier applications of wavelets and multi-resolution analysis to quantum chemistry employed single-component smooth wavelets, these authors used sparse multi-wavelet bases and localized molecular orbitals to attain near linear scaling in electronic structure computations.

5. Classification, Regression and QSAR/QSPR

Understanding the relationship between the structure of a molecule and its physicochemical properties is vital for compound development and property optimization efforts. This is especially true for the expensive task of screening molecular databases for specific biological activities and developing new therapeutic lead compounds. Modern methods of rational drug design depend heavily on the use of computer models to better understand the relationships between compound structures and pharmacokinetic and biochemical behavior. The technique of quantitative structure-activity relationship (QSAR) modeling developed from this need, and seeks correlations between molecular structure and observable molecular properties. There is a logical disconnect in the way molecules need to be represented in order to be understandable to chemists, and the way they must be represented for machine learning applications. To appropriately represent

molecules for numerical analysis, the important features of each compound must be summarized by a concise set of descriptors.[116-118] The existing body of QSAR literature attests to the effectiveness of this technique. [119] Though the term 'QSAR' is normally associated with models developed to explain the properties of small drug-like molecules, it is often used to describe the broader field of Quantitative Structure-Property Relationship (QSPR) modeling that is utilized in chemometrics, cheminformatics and analytical chemistry.

Quantifying the relationship between a molecule and its properties is an important step towards understanding and predicting behavior, whether the models refer to experimental spectra, sensor responses or a set of molecular descriptors. Classification and regression are two main types of modeling approaches frequently used in QSAR/QSPR analysis: *classification* entails assigning data to a discrete category, or clustering it into similar classes, whereas *regression* forms a continuous model that estimates the magnitude of molecular responses. Both of these approaches utilize machine learning methods taken from the fields of statistics and computer science, and both seek to refine raw data into an understandable form. This usually means that a simplification of the raw data is required in order to reveal important discriminatory features within the data.

Data refinement is important because it solves two problems that are frequently encountered while building quantitative structure-property relationship (QSPR) models. The first problem is the 'curse of dimensionality', and the second is 'data variance' within the raw data. These problems are seen while building two common types of QSPR models that appear in chemometrics: pattern recognition of spectroscopic and chromatographic data. The first problem – the 'curse of dimensionality' often arises when the number of features or dimensions used as input to a model greatly exceeds the number of cases or data points available for model development. Raw, continuous spectral data consists of large numbers of variables,[120] the use of which can result in a dimensionality problem, particularly when using high-capacity modeling methods such as artificial neural networks (ANN). This means that the quality and importance of each variable is reduced dramatically, which in turn can affect modeling adversely.[121] The large number of variables or dimensions can inundate a complex mathematical model with features, allowing the model to locate spurious, even false relationships between a molecule and a molecular property.[122] The second problem with using raw data is 'data variance', meaning that the model can suffer from unstable conditions such as changing noise variance or shifting relative peak positions.[123] This also arises as the classic problem of *data alignment* seen in both 2D and 3D molecular QSAR modeling. Such variability between related signals can render the data incomprehensible to classification schemes and regression methods. In general, models tend to increase in complexity with irrelevant sources of variance and noise,[124] making them less general, less robust and less accurate. An example where data variance harms pattern recognition is seen with infrared spectra: when taking spectra from two different spectrophotometers, the precise peak position or signal-to-noise ratio may be sufficiently different so as to confound pattern recognition routines – even when the IR spectra themselves are of good quality. Instead of using such raw data directly, preprocessing is often required to convert the data to a more useable form for effective model building.

Wavelet transform methods offer effective ways to address the two main problems of capturing the information contained within raw data. First, they facilitate feature/dimension reduction by converting raw data into a more succinct representation in wavelet space. The wavelets isolate and concentrate the shape and character of the raw signal into a relatively small number of wavelet coefficients. Using the wavelet coefficients themselves as the data features (descriptors) affords data representation with a dramatic reduction in the number of variables.[122, 125, 126] Second, the use of feature isolation and extraction methods removes variance and error from the raw data giving standardized and consistent data representations.[127] These methods often include basic signal smoothing and denoising, but more complicated data cleaning, such as removal of the variance in peak positions, is possible as well.[128] The removal of unnecessary information from the raw data further concentrates the desired chemical information in the remaining variables used to represent the signal. What remains is a highly compact, consistent and standardized representation of the important discriminatory features within the original signal. Using wavelet coefficients directly in QSAR/QSPR modeling provides a set of low-dimensional, information-rich descriptors that capture the shape and character of the raw data distribution,[129] and help to build more parsimonious models.[124]

Wavelet coefficient descriptors (WCD)[130, 131] exemplify how the wavelet transformation enhances the quality of current descriptor technology, and enables the development of improved models. WCDs are an adaptation of the Transferable Atom Equivalent (TAE) descriptors developed by Breneman.[132] TAE descriptors are derived by quantifying the distributions of multiple electronic properties computed on electronic van der Waals surfaces, defined as the 0.002 e-au^{-3} isosurfaces (Figure 15). TAE descriptors encode the distributions of electron density-based molecular properties such as electronic kinetic energy densities,[133] local average ionization potential,[134] electrostatic potential,[135-138] Fukui functions,[139-142] electron density gradients and electron density Laplacian,[133] in addition to the density itself. The term 'TAE descriptor' refers to a set of histograms with fixed-width bins that characterize surface property distributions (for one example, see Figure 16). Though the TAE descriptors are capable of generating high quality models, they are non-orthogonal, in that histogram representations of property distributions contain correlated information from the same property. Wavelet coefficient descriptors (WCDs) redefine and simplify TAE descriptor data into a stable, orthogonal representation.

[FIGURE 15 HERE]

Using multi-resolution analysis, each property distribution is transformed into wavelet space, separating the data into frequency bands. Because the low frequency features of the property density distributions contain most of the chemical information, a significant compression of the distributions is achieved by retaining only the wavelet coefficients in the very lowest frequency bands. These few wavelets are sufficient to describe the overall shape and character of the property distributions (see Figure 16) without carrying redundant information about their features. WCD descriptors have been shown to improve QSAR modeling because they are inherently an orthogonal representation that

separate and isolate features of each surface property distribution. This aspect of WCDs ensures the resulting models to be more robust, minimizing the risk of finding spurious relationships within the set of descriptors, thereby improving overall model parsimony, and leading to greater generalizability than when using numerous, non-orthogonal descriptors.

[FIGURE 16 HERE]

A demonstration of the performance of WCDs in a pharmaceutical setting can be illustrated by the development of a genetic algorithm/partial least squares (GA/PLS)[143-145] model of HIV reverse-transcriptase (HIVrt) inhibition.[146] In this example, a set of 64 molecules with assay (EC_{50}) values was used to train and evaluate a QSAR model using either TAE or WCD descriptors. Because the electronic surface properties being represented by each method are the same, this example serves to compare the benefits of the wavelet representation over the TAE surface histogram representation of these properties. The results shown in Figure 17 are taken from the cross-validated GA/PLS model predictions. It is significant to note that the number of descriptors required to produce the TAE-based model was nearly twice that of the WCD model. With its smaller number of features, the WCD-based model would be expected to be more stable and robust. This was, in fact, demonstrated during the process of model building, where WCD-based models were found to be less sensitive to GA/PLS tuning parameters than models built using TAE descriptors.

[FIGURE 17 HERE]

Other chemistry-related modeling applications have also seen improvement through the use of wavelet coefficient representations of molecular or spectral properties. An extensive list of examples of this phenomenon is available in the literature. For instance, wavelet transform methods were shown to improve spectral classification models,[147] artificial neural network (ANN)-based chromatography methods[148] on flame ionization data.[149] Although other descriptor-based methods of representing IR spectral features have been used with some success in this application,[150-152] wavelet transform methods were found to be superior for building classification or QSAR regression models. Pattern recognition and regression using wavelet coefficients were also found to improve other IR spectral analysis studies.[122, 125, 126, 128, 153] Due to their ability to concentrate chemical information and reduce unwanted features, a rich literature is developing around the use of wavelets as descriptors of chemical data.

Examples of the improvement possible in classification and regression analysis on experimental data using wavelet transform methods can be found in a number of areas: Finite impulse response (FIR) models built on impulse response improved dramatically, for example, when the signals were significantly down-sampled by converting data with wavelet transforms.[154] Classifications by ANN of HPLC data of trace organic impurities[155] and thermally modulated sensor signals for various gas types[156] were improved dramatically by wavelet transform preprocessing over previous approaches.[157] The wavelet neural network (WNN)[95] has been used for predicting retention times in

programmed-temperature gas chromatography (PTGC)[158] and to build more generalizable models for predicting association constant values (K_a) for benzene derivatives.[159] An interesting application of wavelet transform methods to improve the ANN classification of chromatography data is presented in the work of Yiyu, Minjun and Welsh,[160] where the wavelet transform was used to decompose the chromatography data and fractal analysis was used to analyze the wavelet components. The ANN categorized the compound using the 'chromatographic fingerprint' or fractal dimension of the wavelet coefficients. Other studies that used wavelet transform methods to generate data descriptors giving enhanced modeling results are found in references 120, 161-163.

Summary

We have illustrated the utility of wavelet transformations throughout this chapter, for cleaning, smoothing and denoising data, as well as the benefits of their direct use as molecular property descriptors for extracting information from these signals and the utility of multi-resolution. It is clear from the examples cited that this versatile technology can identify and quantify important features within spectra or property distributions of chemical interest for use in both classification and regression models, to achieve near linear scaling in electronic structure calculations and also serve to control the quality of a basis set in *ab initio* molecular dynamics simulations. The evolution of wavelets in chemistry parallels the development of ever more sophisticated computational methods and hardware performance. In a relatively short period of time, wavelet methods have grown in importance from a noise filter and baseline correction tool to a fundamental component of modern data analysis, computational chemistry and knowledge discovery.

References

1. G. Bachman, L. Narici and E. Beckenstein, *Fourier and Wavelet Analysis,* Springer-Verlag, New York, 2000.
2. A. Teolis, *Computational Signal Processing with Wavelets,* Birkhäuser, Boston, 1998.
3. I. Daubechies, *Ten Lectures on Wavelets,* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1992.
4. R. M. Rao and A. S. Bopardikar, *Wavelet Transforms: Introduction to Theory and Applications,* Addison-Wesley, Reading, MA, 1998.
5. A. Grossman and J. Morlet, *SIAM J. Math. Anal.*, **15**, 4 (1984), 723-736. Decompositions of Hardy Functions into Square Integrable Wavelets of Constant Shape.
6. Y. Meyer, *Wavelets: Algorithms and Applications,* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1993.
7. Y. Mallet, *IEEE T. Pattern Anal.,* **11** (1989), 674-693. A Theory for Multiresolution Signal Decomposition - the Wavelet Representation.
8. R. R. Coifman, Y. Meyer, S. Quake and M. V. Wickerhauser, In *Wavelets and Their Applications*, J. S. Byrnes, J. L. Byrnes, K. A. Hargreaves and K. Berry, Eds. Kluwer Academic Publishers, Norwell, MA, 1994; Vol. 442, pp 363-379. Signal Processing and Compression with Wavelet Packets.
9. B. Walczak, *Wavelets in Chemistry*. Elsevier Science: New York, 2000; Vol. 22.
10. A. Felinger, *Data Analysis and Signal Processing in Chromatography,* Elsevier Health Sciences, Amsterdam, 1998; Vol. 21.
11. B. K. Alsberg, A. M. Woodward and D. B. Kell, *Chemom. Intell. Lab. Syst.,* **37** (1997), 215-239. An Introduction to Wavelet Transforms for Chemometricians: A Time-Frequency Approach - A Tutorial.
12. U. Depczynski, K. Jetter, K. Molt and A. Niemöller, *Chemom. Intell. Lab. Syst.,* **39** (1997), 19-27. The Fast Wavelet Transform on Compact Intervals as a Tool in Chemometrics. I. Mathematical Background.
13. J. J. Workman, D. J. Veltkamp, S. Doherty, B. B. Anderson, K. E. Creasy, M. Koch, J. F. Tatera, A. L. Robinson, L. Bond, L. W. Burgess, G. N. Bokerman, A. H. Ullman, G. P. Darsey, F. Mozayeni, J. A. Bamberger and M. S. Greenwood, *Anal. Chem.,* **71** (1999), 121R-180R. Process Analytical Chemistry.
14. X. Shao, A. K.-M. Leung and F.-T. Chau, *Acc. Chem. Res.,* **36** (2003), 276-283. Wavelet: A New Trend in Chemistry.
15. A. K.-M. Leung, F.-T. Chau and J.-B. Gao, *Chemom. Intell. Lab. Syst.,* **43** (1998), 165-184. A Review on Applications of Wavelet Transform Techniques in Chemical Analysis: 1989-1997.
16. S. Wold and M. Sjöström, *Chemom. Intell. Lab. Syst.,* **44** (1998), 3-14. Chemometrics, Present and Future Success.
17. H. C. Smit and E. J. Van Den Heuvel, In *Chemometrics and Species Identification*, C. Armanino, M. Forina, P. H. E. Gardiner, E. J. Van Den Heuvel, G. Kateman, S. Lanteri, H. C. Smit and B. G. M. Vandeginste, Eds. Springer-

Verlag, New York, 1987; Vol. 141, pp 63-89. Signal and Data Analysis in Chromatography.
18. J. Trygg, Wavelets in Chemometrics - Compression, Denoising, and Feature Extraction. http://www.acc.umu.se/~tnkjtg/Chemometrics/Editorial (October 2003),
19. B. B. Hubbard, *The World According to Wavelets*, A K Peters, Ltd., Natick, MA, 1998.
20. C. Gonnet and B. Torrésani, *Signal Process.*, **37** (1994), 389-404. Local Frequency Analysis with the Two-Dimensional Wavelet Transform.
21. J. Blanc-Talon and D. C. Popescu, *Imaging and Vision Systems: Theory, Assessment and Applications*, Nova Science Publishers, Inc., Huntington, 2001.
22. M. Misiti, Y. Misiti, G. Oppenheim and J.-M. Poggi *Wavelet Toolbox User's Guide Version 3*, The MathWorks, Inc.: Natick, MA, 2004.
23. B. Walczak and D. L. Massart, *Chemom. Intell. Lab. Syst.*, **36** (1997), 81-94. Noise Suppression and Signal Compression Using the Wavelet Packet Transform - A Tutorial.
24. R. R. Coifman and M. V. Wickerhauser, *IEEE T. Inform. Theory*, **38**, 2 (1992), 713-718. Entropy-Based Algorithms for Best Basis Selection.
25. V. J. Barclay and R. F. Bonner, *Anal. Chem.*, **69** (1997), 78-90. Application of Wavelet Transforms to Experimental Spectra: Smoothing, Denoising, and Data Set Compression.
26. C. Cai and P. D. B. Harrington, *J. Chem. Inf. Comput. Sci.*, **38** (1998), 1161-1170. Different Discrete Wavelet Transforms Applied to Denoising Analytical Data.
27. C. R. Mittermayr, S. G. Nikolov, H. Hutter and M. Grasserbauer, *Chemom. Intell. Lab. Syst.*, **34** (1996), 187-202. Wavelet Denoising of Gaussian Peaks: A Comparative Study.
28. C. Perrin, B. Walczak and D. L. Massart, *Anal. Chem.*, **73** (2001), 4903-4917. The Use of Wavelets for Signal Denoising in Capillary Electrophoresis.
29. D. L. Donoho, In *Different Perspectives on Wavelets*, I. Daubechies, Ed. American Mathematical Society, Providence, R.I., 1993; Vol. 47, pp 173-205. Nonlinear Wavelet Methods for Recovery of Signals, Densities, and Spectra from Indirect and Noisy Data.
30. M. Wolkenstein, H. Hutter and M. Grasserbauer, *Fresenius J. Anal. Chem.*, **358** (1997), 165-169. Wavelet Filtering for Analytical Data.
31. D. L. Donoho, *IEEE T. Inform. Theory*, **41**, 3 (1995), 613-627. De-Noising by Soft-Thresholding.
32. S. Sakakibara, In *Wavelets: Theory, Algorithms, and Applications*, C. K. Chui, L. Montefusco and L. Puccio, Eds. Academic Press, New York, 1994; pp 179-196. A Practice of Data Smoothing by B-spline Wavelets.
33. B. K. Alsberg, A. M. Woodward, M. K. Winson, J. Rowland and D. B. Kell, *Analyst*, **122** (1997), 645-652. Wavelet Denoising of Infrared Spectra.
34. C. R. Mittermayr, H. Frischenschlager, E. Rosenberg and M. Grasserbauer, *Fresenius J. Anal. Chem.*, **358** (1997), 456-464. Filtering and Integration of Chromatographic Data: A Tool to Improve Calibration?

35. X. Shao and W. Cai, *J. Chemom.,* **12** (1998), 85-93. Resolution of Multicomponent Chromatograms by Window Factor Analysis with Wavelet Transform Preprocessing.
36. D. Jouan-Rimbaud, B. Walczak, R. J. Poppi, O. E. De Noord and D. L. Massart, *Anal. Chem.,* **69** (1997), 4317-4323. Application of Wavelet Transform To Extract the Relevant Component from Spectral Data for Multivariate Calibration.
37. A. M. Woodward, B. K. Alsberg and D. B. Kell, *Chemom. Intell. Lab. Syst.,* **40** (1998), 101-107. The Effect of Heteroscedastic Noise on the Chemometric Modelling of Frequency Domain Data.
38. H.-W. Tan and S. D. Brown, *J. Chemom.,* **16** (2002), 228-240. Wavelet Analysis Applied to Removing Non-Constant, Varying Spectroscopic Background in Multivariate Calibration.
39. X.-Q. Lu and J.-Y. Mo, *Analyst,* **121** (1996), 1019-1024. Spline Wavelet Multi-Resolution Analysis for High-Noise Digital Signal Processing in Ultraviolet-Visible Spectrophotometry.
40. M. Wolkenstein, T. Stubbings and H. Hutter, *Fresenius J. Anal. Chem.,* **365** (1999), 63-69. Robust Automated Three-Dimensional Segmentation of Secondary Ion Mass Spectrometry Image Sets.
41. M. Wolkenstein, H. Hutter, S. G. Nikolov and M. Grasserbauer, *Fresenius J. Anal. Chem.,* **357** (1997), 783-788. Improvement of SIMS Image Classification by Means of Wavelet De-Noising.
42. S. G. Nikolov, H. Hutter and M. Grasserbauer, *Chemom. Intell. Lab. Syst.,* **34** (1996), 263-273. De-Noising of SIMS Images via Wavelet Shrinkage.
43. X.-N. Li, Y.-Z. Liang and F.-T. Chau, *Chemom. Intell. Lab. Syst.,* **63** (2002), 139-153. Smoothing Methods Applied to Dealing with Heteroscedastic Noise in GC/MS.
44. H. Fang and H.-Y. Chen, *Anal. Chim. Acta,* **346** (1997), 319-325. Wavelet Analyses of Electroanalytical Chemistry Responses and an Adaptive Wavelet Filter.
45. L. Bao, J. Mo and Z. Tang, *Anal. Chem.,* **69** (1997), 3053-3057. The Application in Processing Analytical Chemistry Signals of a Cardinal Spline Approach to Wavelets.
46. D. N. S. Permann and H. Teitelbaum, *J. Phys. Chem.,* **97** (1993), 12670-12673. Wavelet Fast Fourier Transform (WFFT) Analysis of a Mullivolt Signal for a Transient Oscillating Chemical Reaction.
47. B. B. Goldman and W. T. Wipke, *J. Chem. Inf. Comput. Sci.,* **40** (2000), 644-658. Quadratic Shape Descriptors. 1. Rapid Superposition of Dissimilar Molecules Using Geometrically Invariant Surface Descriptors.
48. J. Mao, P. Sun, Z. Pan, Q. Su and Z. Maosen, *Fresenius J. Anal. Chem.,* **361** (1998), 140-142. Wavelet Analysis on Photoacoustic Spectra of Degraded PVC.
49. G. Beylkin, *SIAM J. Numer. Anal.,* **29**, 6 (1992), 1716-1740. On the Representation of Operators in Bases of Compactly Supported Wavelets.
50. W. Dahmen and C. A. Micchelli, *SIAM J. Numer. Anal.,* **30**, 2 (1993), 507-537. Using the Refinement Equation for Evaluating Integrals of Wavelets.
51. E. J. Stollnitz, T. D. Derose and D. H. Salesin, *IEEE Comput. Graph.,* **15**, 3 (1995), 76-84. Wavelets for Computer Graphics: A Primer, Part 1 - A Tutorial.

52. E. J. Stollnitz, T. D. Derose and D. H. Salesin, *IEEE Comput. Graph.*, **15**, 4 (1995), 75-85. Wavelets for Computer Graphics: A Primer, Part 2 - A Tutorial.
53. D. Marr and E. Hildreth, *P. Roy. Soc. Lond. B. Bio.*, **207**, 1167 (1980), 187-217. Theory of Edge Detection.
54. J.-P. Antoine, R. Murenzi and B. Piette, In *Wavelet and Applications*, Y. Meyer, Ed. Masson and Springer-Verlag, Paris, 1992; Vol. 20, pp 144-159. Image Analysis with 2D Continuous Wavelet Transform: Detection of Position, Orientation and Visual Contrast of Simple Objects.
55. R. R. Klevecz, *Funct. Integr. Genomics*, **1** (2000), 186-192. Dynamic Architecture of the Yeast Cell Cycle Uncovered by Wavelet Decomposition of Expression Microarray Data.
56. P. Teppola and P. Minkkinen, *J. Chemom.*, **14** (2000), 383-399. Wavelet-PLS Regression Models for Both Exploratory Data Analysis and Process Monitoring.
57. F. Ehrentreich, *Anal. Bioanal. Chem.*, **372** (2002), 115-121. Wavelet Transform Applications in Analytical Chemistry.
58. X. Shao, W. Cai, P. Sun, M. Zhang and G. Zhao, *Anal. Chem.*, **69** (1997), 1722-1725. Quantitative Determination of the Components in Overlapping Chromatographic Peaks Using Wavelet Transform.
59. L. Sun, W. Cai and X. Shao, *Fresenius J. Anal. Chem.*, **370** (2001), 16-21. A Two-Dimensional Immune Algorithm for Resolution of Overlapping Two-Way Chromatograms.
60. S. Hai-Lin, W. Ji-Hong, L. Yi-Zeng and C. Wen-Can, *Chem. J. Chinese. U.*, **18**, 4 (1997), 530-534. Multiresolution Analysis of Hyphenated Chromatographic Data.
61. H. Shen, J. Wang, Y. Liang, K. Pettersson, M. Josefson, J. Gottfries and F. Lee, *Chemom. Intell. Lab. Syst.*, **37** (1997), 261-269. Chemical Rank Estimation by Multiresolution Analysis for Two-way Data in the Presence of Background.
62. X. Shao, W. Cai and P. Sun, *Chemom. Intell. Lab. Syst.*, **43** (1998), 147-155. Determination of the Component Number in Overlapping Multicomponent Chromatogram Using Wavelet Transform.
63. X. Shao and S. Hou, *Anal. Sci.*, **15** (1999), 681-684. On-Line Resolution of Overlapping Chromatograms Using the Wavelet Transform.
64. X. Zhang, J. Jin, J. Zheng and H. Gao, *Anal. Bioanal. Chem.*, **377**, 7-8 (2003), 1153-1158. Genetic Algorithms Based on Wavelet Transform for Resolving Simulated Overlapped Spectra.
65. N. Lu, F. Wang and F. Gao, *Ind. Eng. Chem. Res.*, **42** (2003), 4198-4207. Combination Method of Principal Component and Wavelet Analysis for Multivariate Process Montoring and Fault Diagnosis.
66. C. L. Stork, D. J. Veltkamp and B. R. Kowalski, *Appl. Spectrosc.*, **52**, 10 (1998), 1348-1352. Detecting and Identifying Spectral Anomalies Using Wavelet Processing.
67. L. Nie, S. Wu, X. Lin, L. Zheng and L. Rui, *J. Chem. Inf. Comput. Sci.*, **42** (2002), 274-283. Approximate Derivative Calculated by Using Continuous Wavelet Transform.
68. J. Chen and X. Z. Wang, *J. Chem. Inf. Comput. Sci.*, **41** (2001), 992-1001. A New Approach to Near-Infrared Spectral Data Analysis Using Independent Component Analysis.

69. D. Broadhurst, R. Goodacre, A. Jones, J. J. Rowland and D. B. Kell, *Anal. Chim. Acta,* **348** (1997), 71-86. Genetic Algorithms as a Method for Variable Selection in Multiple Linear Regression and Partial Least Squares Regression, with Applications to Pyrolysis Mass Spectrometry.
70. H. Martens, M. Høy, B. M. Wise, R. Bro and P. B. Brockhoff, *J. Chemom.,* **17** (2003), 153-165. Pre-Whitening of Data by Covariance-Weighted Pre-Processing.
71. J. A. Hageman, M. Streppel, R. Wehrens and L. M. C. Buydens, *J. Chemom.,* **17** (2003), 427-437. Wavelength Selection with Tabu Search.
72. J. M. Brenchley, U. Hörchner and J. H. Kalivas, *Appl. Spectrosc.,* **51**, 5 (1997), 689-699. Wavelength Selection Characterization for NIR Spectra.
73. S. L. Shew, Method and Apparatus for Determining Relative Ion Abundances in Mass Spectrometry Utilizing Wavelet Transforms. Patent #5,436,447, 1995.
74. X. Shao, L. Shao and G. Zhao, *Anal. Commun.,* **35** (1998), 135-137. Extraction of Extended X-ray Absorption Fine Structure Information from the Experimental Data Using the Wavelet Transform.
75. Y. Ding, T. Nanba and Y. Miura, *Phys. Rev. B: Condens. Matter,* **58**, 21 (1998), 14279-14287. Wavelet Analysis of X-Ray Diffraction Pattern for Glass Structures.
76. D. Barache, J.-P. Antoine and J. M. Dereppe, *J. Magn. Reson.,* **128** (1997), 1-11. The Continuous Wavelet Transform, An Analysis Tool for NMR Spectroscopy.
77. P. Guillemain, R. Kronland-Martinet and B. Martens, In *Wavelet and Applications*, Y. Meyer, Ed. Masson and Springer-Verlag, Paris, 1992; Vol. 20, pp 38-60. Estimation of Spectral Lines with the Help of the Wavelet Transform - Applications in NMR Spectroscopy.
78. G. Neue, *Solid State Nucl. Magn. Reson.,* **5** (1996), 305-314. Simplification of Dynamic NMR Spectroscopy by Wavelet Transforms.
79. F. G. Meyer and G. Mccarthy, In *Information Processing in Medical Imaging: 17th International Conference, IPMI 2001, Davis, California, USA, June 18-22, 2001, Proceedings*, M. F. Insana and R. M. Leahy, Eds. Springer-Verlag, New York, 2001; pp 232-238. Estimation of Baseline Drifts in fMRI.
80. H. Serrai, L. Senhadji, J. D. De Certaines and J. L. Coatrieux, *J. Magn. Reson.,* **124** (1997), 20-34. Time-Domain Quantification of Amplitude, Chemical Shift, Apparent Relaxation Time T*2, and Phase by Wavelet-Transform Analysis. Application to Biomedical Magnetic Resonance Spectroscopy.
81. U. L. Günther, C. Ludwig and H. Rüterjans, *J. Magn. Reson.,* **156** (2002), 19-25. WAVEWAT - Improved Solvent Suppression in NMR Spectra Employing Wavelet Transforms.
82. V. L. Martins, L. F. De Almeida, S. L. De Castro, R. K. H. Galvão, M. C. U. De Araújo and E. C. Da Silva, *J. Chem. Inf. Comput. Sci.,* **43**, 6 (2003), 1725-1732. A Multiscale Wavelet Data Treatment for Reliable Localization of Inflection Points for Analytical Purposes.
83. R. A. Devore, B. Jawerth and B. J. Lucier, *IEEE T. Inform. Theory,* **38**, 2 (1992), 719-746. Image Compression Through Wavelet Transform Coding.
84. J. Le Moigne, Multi-Sensor Image Registration, Fusion and Dimension Reduction. http://satjournal.tcom.ohiou.edu/issue03/applications.html

85. J. Le Moigne and I. Zavorin In *Use of Wavelets for Image Registration*, Wavelet Applications VII, SPIE Aerosense 2000, Orlando, FL, USA, 2000; Orlando, FL, USA, 2000.
86. J. M. Shapiro, *IEEE T. Signal Proces.*, **41**, 12 (1993), 3445-3462. Embedded Image Coding Using Zerotrees of Wavelet Coefficients.
87. L. J. Chipman, T. M. Orr and L. N. Graham, In *Wavelet Applications in Signal and Image Processing III, SPIE, 12-14 July, 1995, San Diego, California, Proceedings*, A. F. Laine, M. A. Unser and M. V. Wickerhauser, Eds. SPIE, the International Society for Optical Engineering, Bellingham, 1995; pp 208-219. Wavelets and Image Fusion.
88. F. Peyrin, M. Zaim and R. Goutte, *J Math. Imaging Vis.*, **3** (1993), 105-121. Construction of Wavelet Decompositions for Tomographic Images.
89. H. Li, S. Manjunath and S. K. Mitra, *Graph. Model. Im. Proc.*, **57**, 3 (1995), 235-245. Multisensor Image Fusion Using the Wavelet Transform.
90. Z. Zhang and R. S. Blum In *Multisensor Image Fusion Using a Region-Based Wavelet Transform Approach*, DARPA Image Understanding Workshop 1997, New Orleans, LA, 1997; New Orleans, LA, 1997; pp 1447-1451.
91. A. K.-M. Leung, F.-T. Chau, J.-B. Gao and T.-M. Shih, *Chemom. Intell. Lab. Syst.*, **43** (1998), 69-88. Application of Wavelet Transform in Infrared Spectrometry: Spectral Compression and Library Search.
92. B. Walczak and D. L. Massart, *Chemom. Intell. Lab. Syst.*, **38** (1997), 39-50. Wavelet Packet Transform Applied to a Set of Signals: A New Approach to the Best-Basis Selection.
93. K.-K. Li, F.-T. Chau and A. K.-M. Leung, *Chemom. Intell. Lab. Syst.*, **52** (2000), 135-143. Compression of Ultraviolet-Visible Spectrum with Recurrent Neural Network.
94. F.-T. Chau, J.-B. Gao, T.-M. Shih and J. Wang, *Appl. Spectrosc.*, **51**, 5 (1997), 649-659. Compression of Infrared Spectral Data Using the Fast Wavelet Transform Method.
95. L. Wei and L. Jinping, *Chin. Sci. Bull.*, **42**, 10 (1997), 822-826. The Compression of IR Spectra by Using Wavelet Neural Network.
96. K. L. Peterson, In *Reviews in Computational Chemistry*, K. B. Lipkowitz and D. B. Boyd, Eds. Wiley-VCH, New York, 2000; Vol. 16, pp 53-140. Artificial Neural Networks and Their Use in Chemistry.
97. K. J. Cho, T. A. Arias, J. D. Joannopoulos and P. K. Lam, *Phys. Rev. Lett.*, **71**, 12 (1993), 1808-1811. Wavelets in Electronic Structure Calculations.
98. M. F. Herman, *Annu. Rev. Phys. Chem.*, **45** (1994), 83-111. Dynamics by Semiclassical Methods.
99. R. Kosloff, *Annu. Rev. Phys. Chem.*, **45** (1994), 145-178. Propagation Methods for Quantum Molecular Dynamics.
100. H. Nakamura, *Annu. Rev. Phys. Chem.*, **48** (1997), 299-328. Theoretical Studies of Chemical Dunamics: Overview of Some Functional Mechanisms.
101. T. A. Arias, *Rev. Mod. Phys.*, **71**, 1 (1999), 267-311. Multiresolution Analysis of Electronic Structure: Semicardinal and Wavelet Bases.
102. D. Feller and E. R. Davidson, In *Reviews in Computational Chemistry*, K. B. Lipkowitz and D. B. Boyd, Eds. VCH Publishers, Weinheim, 1990; Vol. 1, pp 1-

43. Basis Sets for Ab Initio Molecular Orbital Calculations and Intermolecular Interactions.
103. T. A. Arias, K. J. Cho, J. D. Joannopoulos, P. K. Lam and M. P. Teter, In *Toward Teraflop Computing and New Grand Challenge Applications*, R. K. Kalia and P. Vashishta, Eds. Nova Science Publishers, Inc., Commack, N.Y., 1995; pp 25-36. Wavelet-Transform Representation of the Electronic Structure of Materials.
104. S. Wei and M. Y. Chou, *Phys. Rev. Lett.*, **76**, 15 (1996), 2650-2653. Wavelets in Self-Consistent Electronic Structure Calculations.
105. P. Fischer and M. Defranceschi, In *Conceptual Trends in Quantum Chemistry*, E. S. Kryachko and J. L. Calais, Eds. Kluwer Academic, Dordrecht, 1994; pp 227-247. The Wavelet Transform: A New Mathematical Tool for Quantum Chemistry.
106. S. Nagy and P. János, *Int. J. Quantum Chem.*, **84** (2001), 523-529. Multiresolution Analysis of Density Operators, Electron Density, and Energy Functionals.
107. S. S. Iyengar and M. J. Frisch, *J. Chem. Phys.*, **121**, 11 (2004), 5061-5070. Effect of Time-Dependent Basis Functions and Their Superposition Error on Atom-Centered Density Matrix Propagation (ADMP): Connections to Wavelet Theory of Multiresolution Analysis.
108. N. R. Kestner and J. E. Combariza, In *Reviews in Computational Chemistry*, K. B. Lipkowitz and D. B. Boyd, Eds. Wiley-VCH, New York, 1999; Vol. 13, pp 99-132. Basis Set Superposition Errors: Theory and Practice.
109. J. P. Modisette, P. Nordlander, J. L. Kinsey and B. R. Johnson, *Chem. Phys. Lett.*, **250** (1996), 485-494. Wavelet Bases in Eigenvalue Problems in Quantum Mechanics.
110. P. Fischer and M. Defranceschi, *Int. J. Quantum Chem.*, **45** (1993), 619-636. Looking at Atomic Orbitals Through Fourier and Wavelet Transforms.
111. P. Fischer and M. Defranceschi, In *Wavelets: Theory, Algorithms and Applications*, C. K. Chui, L. Montefusco and L. Puccio, Eds. Academic Press, New York, 1994; pp 495-506. Representation of the Atomic Hartree-Fock Equations in a Wavelet Basis by Means of the BCR Algorithm.
112. E. Von Kitzing and E. Schmitt, *THEOCHEM*, **336** (1995), 245-259. Configurational Space of Biological Macromolecules as Seen by Semi-empirical Force Fields: Inherent Problems for Molecular Design and Strategies to Solve Them by Means of Hierarchical Force Fields.
113. R. J. Harrison, G. I. Fann, T. Yanai, Z. Gan and G. Beylkin, *J. Chem. Phys.*, **121**, 23 (2004), 11587-11598. Multiresolution Quantum Chemistry: Basic Theory and Initial Applications.
114. T. Yanai, G. I. Fann, Z. Gan, R. J. Harrison and G. Beylkin, *J. Chem. Phys.*, **121**, 14 (2004), 6680-6688. Multiresolution Quantum Chemistry in Multiwavelet Bases: Hartree-Fock Exchange.
115. T. Yanai, G. I. Fann, Z. Gan, R. J. Harrison and G. Beylkin, *J. Chem. Phys.*, **121**, 7 (2004), 2866-2876. Multiresolution Quantum Chemistry in Multiwavelet Bases: Analytic Derivatives for Hatree-Fock and Density Functional Theory.
116. L. B. Kier and L. H. Hall, *Molecular Connectivity In Chemistry and Drug Research*, Academic Press, New York, 1976.

117. L. B. Kier and L. H. Hall, *Molecular Connectivity in Structure-Activity Analysis*, John Wiley, London, 1986.
118. L. B. Kier, L. H. Hall, W. J. Murray and M. Randic, *J. Pharm. Sci.*, **64**, 12 (1975), 1971-1974. Molecular connectivity. I: Relationship to Nonspecific Local Anesthesia.
119. C. Hansch and A. Leo, *Exploring QSAR*, American Chemical Society, Washington, D.C., 1995.
120. M. Cocchi, R. Seeber and A. Ulrici, *Chemom. Intell. Lab. Syst.*, **57** (2001), 97-119. WPTER: Wavelet Packet Transform for Efficient Pattern Recognition of Signals.
121. L. I. Nord and S. P. Jacobsson, *Chemom. Intell. Lab. Syst.*, **44** (1998), 153-160. A Novel Method for Examination of the Variable Contribution to Computational Neural Network Models.
122. Y. Mallet, D. Coomans, J. Kautsky and O. D. Vel, *IEEE T. Pattern Anal.*, **19**, 10 (1997), 1058-1066. Classification Using Adaptive Wavelets for Feature Extraction.
123. P. Wunsch and A. F. Laine, *Pattern Recogn.*, **28**, 8 (1995), 1249-1249. Wavelet Descriptors for Multiresolution Recognition of Handprinted Characters.
124. O. E. D. Noord, *Chemom. Intell. Lab. Syst.*, **23** (1994), 65-70. The Influence of Data Preprocessing on the Robustness and Parsimony of Multivariate Calibration Models.
125. Y. Mallet, D. Coomans and O. De Vel, *Chemom. Intell. Lab. Syst.*, **35** (1996), 157-173. Recent Development in Discriminant Analysis on High Dimensional Spectral Data.
126. M. Bos and J. A. M. Vrielink, *Chemom. Intell. Lab. Syst.*, **23** (1994), 115-122. The Wavelet Transform for Pre-processing IR Spectra in the Identification of Mono- and Di-substituted Benzenes.
127. B. Walczak, E. Bouveresse and D. L. Massart, *Chemom. Intell. Lab. Syst.*, **36** (1997), 41-51. Standardization of Near-Infrared Spectra in the Wavelet Domain.
128. B. Walczak, B. Van Den Bogaert and D. L. Massart, *Anal. Chem.*, **68** (1996), 1742-1747. Application of Wavelet Packet Transform in Pattern Recognition of Near-IR Data.
129. P. D. B. Harrington, A. Urbas and C. Wan, *Anal. Chem.*, **72** (2000), 5004-5013. Evaluation of Neural Network Models with Generalized Sensitivity Analysis.
130. C. M. Breneman, N. Sukumar, K. P. Bennett, M. J. Embrechts, C. M. Sundling and L. Lockwood, Wavelet Representations of Molecular Electronic Properties: Applications in ADME, QSPR, and QSAR. In *QSAR In Vivo: American Chemical Society National Meeting*, American Chemical Society: Washington, D.C., 2000.
131. L. Lockwood. 1. Effects of a Trimethylsilyl Substituent on the Photochemical Rearrangement of 2,5-cyclohexadienones. 2. Development and Implementation of Novel Electron Density Derived Molecular Surface Descriptors. Ph.D. Dissertation, Rensselaer Polytechnic Institute, Troy, NY, 2002.
132. C. M. Breneman and M. Rhem, *J. Comput. Chem.*, **18**, 2 (1997), 182-197. A QSPR Analysis of HPLC Column Capacity Factors for a set of High-Energy Materials Using Electronic Van der Waals Surface Property Descriptors Computed by the Transferable Atom Equivalent Method.

133. R. F. W. Bader, *Atoms in Molecules: A Quantum Theory*, Oxford Press, Oxford, 1990.
134. J. S. Murray and P. Politzer, In *Theoretical Organic Chemistry*, C. Párkányi, Ed. Elsevier Science Publishers BV, Amsterdam, The Netherlands, 1998; Vol. 5, pp 189-202. Average Local Ionization Energies: Significance and Applications.
135. P. Politzer and D. G. Truhlar, *Chemical Applications of Atomic and Molecular Electrostatic Potential*. Plenum Press: New York, 1981.
136. P. Politzer, N. Sukumar, K. Jayasuriya and S. Ranganathan, *J. Am. Chem. Soc.*, **110** (1988), 3425-3430. A Computational Evaluation and Comparison of Nitramine Properties.
137. J. S. Murray, N. Sukumar, S. Ranganathan and P. Politzer, *Int. J. Quantum Chem.*, **37**, 5 (1990), 611-629. A Computational Analysis of the Electrostatic Potentials and Relative Bond Strengths of Hydrazine and Some of its 1,1-dimethyl Derivatives.
138. C. M. Breneman and M. Martinov, In *Molecular Electrostatic Potentials: Concepts and Applications*, J. S. Murray and K. Sen, Eds. Elsevier, Amsterdam, 1996; Vol. 3, pp 143-179. The Use of Electrostatic Potential Fields in QSAR and QSPR.
139. K. Fukui, T. Yonezawa and C. Nagata, *J. Chem. Phys.*, **27**, 6 (1957), 1247-1259. MO-Theoretical Approach to the Mechanism of Charge Transfer in the Process of Aromatic Substitutions.
140. K. Fukui, *Theory of Orientation and Stereoselection*, Springer-Verlag, Berlin, 1975.
141. R. G. Parr and W. Yang, *J. Am. Chem. Soc.*, **106** (1984), 4049-4050. Density-Functional Approach to the Frontier-Electron Theory of Chemical Rectivity.
142. R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules*, Oxford University Press, New York, 1989.
143. K. Hasegawa and K. Funatsu, *SAR QSAR Environ. Res.*, **11**, 3-4 (2000), 189-209. Partial Least-Squares Modeling and Genetic Algorithm Optimization in Quantitative Structure-Activity Relationships.
144. K. Hasegawa, Y. Miyashita and K. Funatsu, *J. Chem. Inf. Comput. Sci.*, **37**, 2 (1997), 306-10. GA Strategy for Variable Selection in QSAR Studies: GA-Based PLS Analysis of Calcium Channel Antagonists.
145. M. Ozdemir. Evolutionary Computing for Feature Selection and Predictive Data Mining. Ph.D. Dissertation, Rensselaer Polytechnic Institute, Troy, NY, 2002.
146. R. Garg, S. P. Gupta, H. Gao, M. S. Babu, A. K. Debnath and C. Hansch, *Chemical Reviews*, **99** (1999), 3525-3691. Comparative Quantitative Structure−Activity Relationship Studies on Anti-HIV Drugs.
147. B. Debska, B. Guzowska-Swider and D. Carbol-Bass, *J. Chem. Inf. Comput. Sci.*, **40** (2000), 330-338. Automatic Generation of Knowledge Base from Infrared Spectral Database for Substructure Recognition.
148. C. Cai and P. D. B. Harrington, *Anal. Chem.*, **71** (1999), 4134-4141. Prediction of Substructure and Toxicity of Pesticides with Temperature Constrained-Cascade Correlation Network from Low-Resolution Mass Spectra.

149. M. Jalali-Heravi and M. H. Fatemi, *J. Chromatogr. A*, **825** (1998), 161-169. Prediction of Flame Ionization Detector Response Factors Using an Artificial Neural Network.
150. C. M. R. Ginn, D. B. Turner and P. Willett, *J. Chem. Inf. Comput. Sci.*, **37** (1997), 23-37. Similarity Searching in Files of Three-Dimensional Chemical Structures: Evaluation of the EVA Descriptor and Combination of Rankings Using Data Fusion.
151. A. M. Ferguson, T. W. Heritage, P. Jonathon, S. E. Pack, L. Phillips, J. Rogan and P. J. Snaith, *J. Comput. Aided Mol. Des.*, **11** (1997), 143-152. EVA: A New Theoretically Based Molecular Descriptor for Use in QSAR/QSPR Analysis.
152. T. W. Heritage, A. M. Ferguson, D. B. Turner and P. Willett, *Perspect. Drug Discov.*, **9-11** (1998), 381-398. EVA: A Novel Theoretical Descriptor for QSAR Studies.
153. J. Trygg and S. Wold, *Chemom. Intell. Lab. Syst.*, **42** (1998), 209-220. PLS Regression on Wavelet Compressed NIR Spectra.
154. M. Nikolaou and P. Vuthandam, *AIChE J.*, **44**, 1 (1998), 141-150. FIR Model Identification: Parsimony Through Kernel Compression with Wavelets.
155. E. R. Collantes, R. Duta, W. J. Welsh, W. L. Zielinski and J. Brower, *Anal. Chem.*, **69** (1997), 1392-1397. Preprocessing of HPLC Trace Impurity Patterns by Wavelet Packets for Pharmaceutical Fingerprinting Using Artificial Neural Networks.
156. S. Maldonado-Bascón, S. Al-Khalifa and F. López-Ferreras, In *Computational Methods in Neural Modeling: Seventh International Conference, IWANN 2003, Mao, Spain, June 3-6, 2003, Proceedings*, J. Méra and J. R. Alvarez, Eds. Springer-Verlag, New York, 2003; Vol. 2687, pp 798-805. Feature Reduction Using Support Vector Machines for Binary Gas Detection.
157. P. Althainz, J. Goschnick, S. Ehrmann and H. J. Ache, *Sensor. Actuat. B*, **33** (1996), 72-76. Multisensor Microsystem for Contaminants in Air.
158. X. Zhang, J. Qi, R. Zhang, M. Liu, Z. Hu, H. Xue and B. Fan, *Comput. Chem. (Oxford)*, **25** (2001), 125-133. Prediction of Programmed-Temperature Retention Values of Naphthas by Wavelet Neural Networks.
159. L. Liu and Q.-X. Guo, *J. Chem. Inf. Comput. Sci.*, **39** (1999), 133-138. Wavelet Neural Network and Its Application to the Inclusion of beta-Cyclodextrin with Benzene Derivatives.
160. C. Yiyu, C. Minjun and W. J. Welsh, *J. Chem. Inf. Comput. Sci.*, **43**, 6 (2003), 1959-1965. Fractal Fingerprinting of Chromatographic Profiles Based on Wavelet Analysis and Its Application to Characterize the Quality Grade of Medicinal Herbs.
161. M. Köküer, F. Murtagh, N. D. Mcmillan, S. Riedel, B. O'rourke, K. Beverly, A. T. Augousti and J. Mason, *J. Chem. Inf. Comput. Sci.*, **43** (2003), 587-594. A Wavelet, Fourier and PCA Data Analysis Pipeline: Application to Distinguishing Mixtures of Liquids.
162. J. Zhao, X. W. Yang, J. P. Li and Y. Y. Tang, In *Wavelet Analysis and Its Applications: Second International Conference, WAA 2001, Hong Kong, China, December 18-20, 2001, Proceedings*, Y. Y. Tang, M. V. Wickerhauser, P. C.

Yuen and C. H. Li, Eds. Springer-Verlag, New York, 2002; Vol. 2251, pp 424-429. DNA Sequences Classification Based on Wavelet Packet Analysis.
163. G. V. Gkoutos, H. Rzepa, R. M. Clark, O. Adjei and H. Johal, *J. Chem. Inf. Comput. Sci.*, **43**, 5 (2003), 1342-1355. Chemical Machine Vision: Automated Extraction of Chemical Metadata from Raster Images.

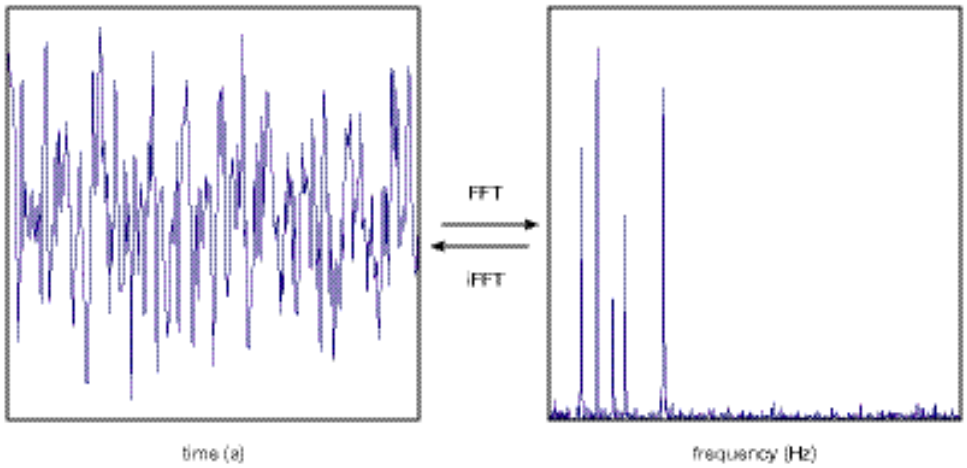

Figure 1 – This illustration shows a signal converted from the time domain to the frequency domain using a Fourier transform technique. The fast Fourier transformation (FFT) is a discrete and computationally efficient version of the general Fourier transformation.

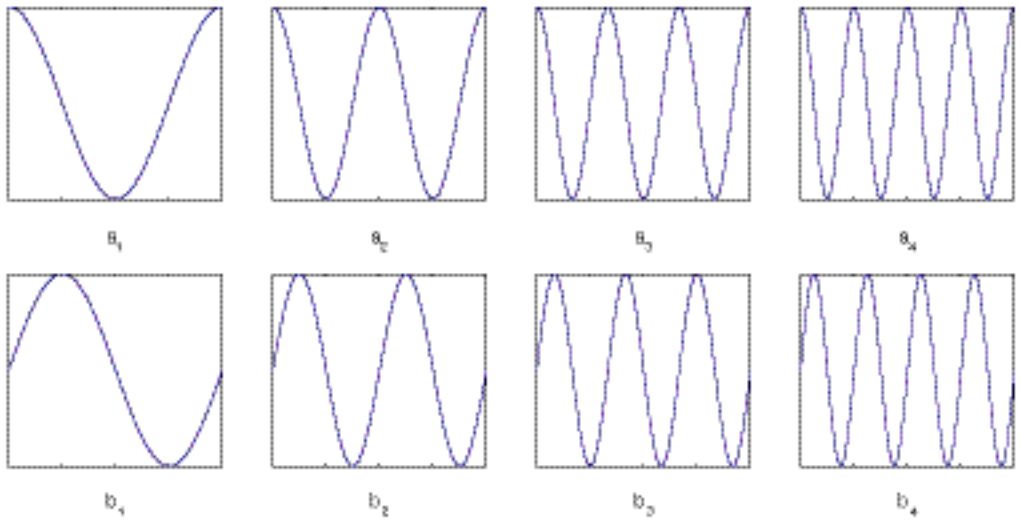

Figure 2 – Illustrated are four example pairs of sinusoidal Fourier basis functions that constitute a portion of a Fourier series. Each pair of functions *a* and *b* have the same frequency, but are 90˚ out of phase. They combine to give a series of terms of infinite domain, at particular frequencies, but of arbitrary phases.

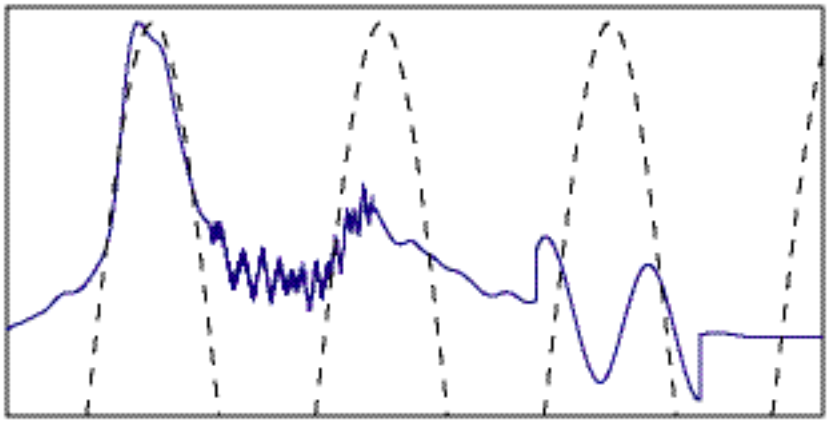

Figure 3 – A non-stationary signal (solid line) is fitted with a Fourier transform basis function (dashed line). The basis function is fitted well against a single signal peak, but outside of the immediate region of the peak the signal approximation suffers. This illustrates the ability of a Fourier basis function to isolate frequency information, but not time-domain information.

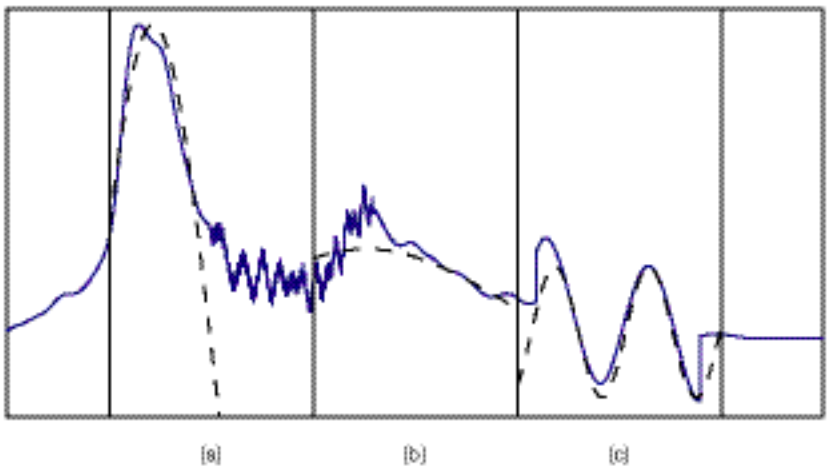

Figure 4 – The windowing (windows a, b, and c shown here) of a non-stationary signal (solid line) in short-time Fourier transform analysis gives some locality of time information to the Fourier transform (dashed line is Fourier basis function). Even so, this method still suffers from a trade-off of knowledge between time-domain and frequency-domain information.

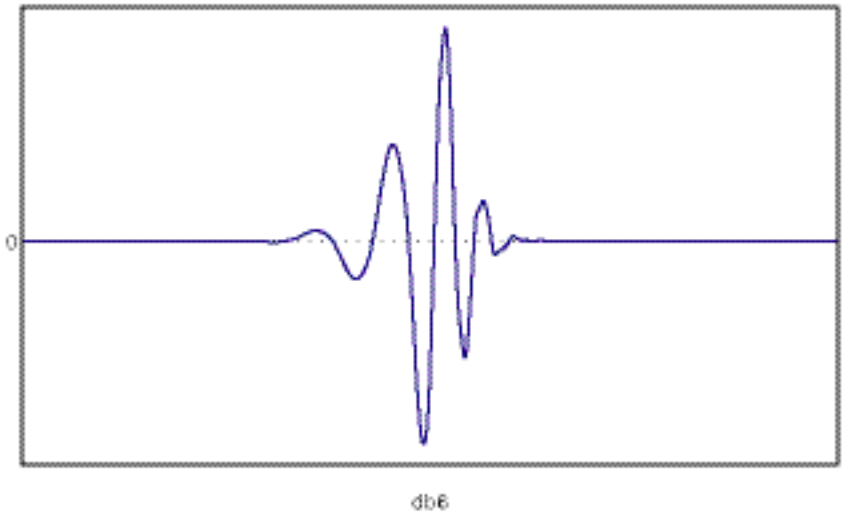

Figure 5 – This is an example of a Daubechies wavelet (Daubechies 6) that illustrates the two interesting wavelet properties: localized in time and oscillation.

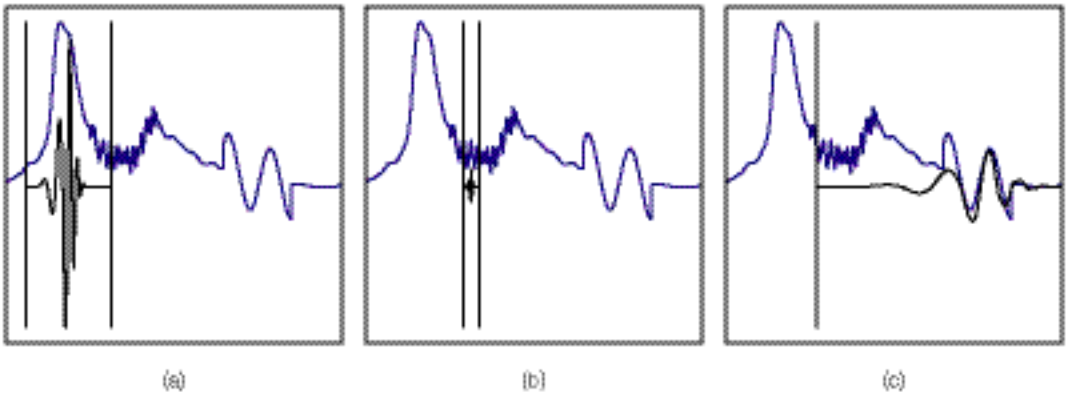

Figure 6 – Shown here are three wavelet “windows” over a non-stationary signal that illustrate a partial wavelet scaling analysis (a, b, c) and how wavelets simultaneously identify component frequency and position information within the signal. The continuous wavelet transform performs an exhaustive fitting of the different features of the signal at different scales and positions of the wavelet function.

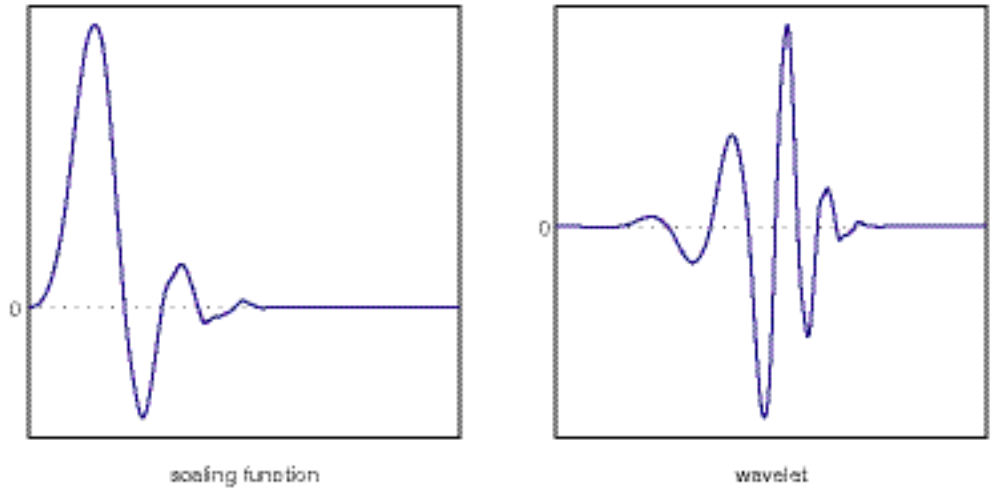

Figure 7 – Each wavelet basis is actually a pair of functions: the wavelet and its scaling function. These are the two functions of the Daubechies 6 wavelet.

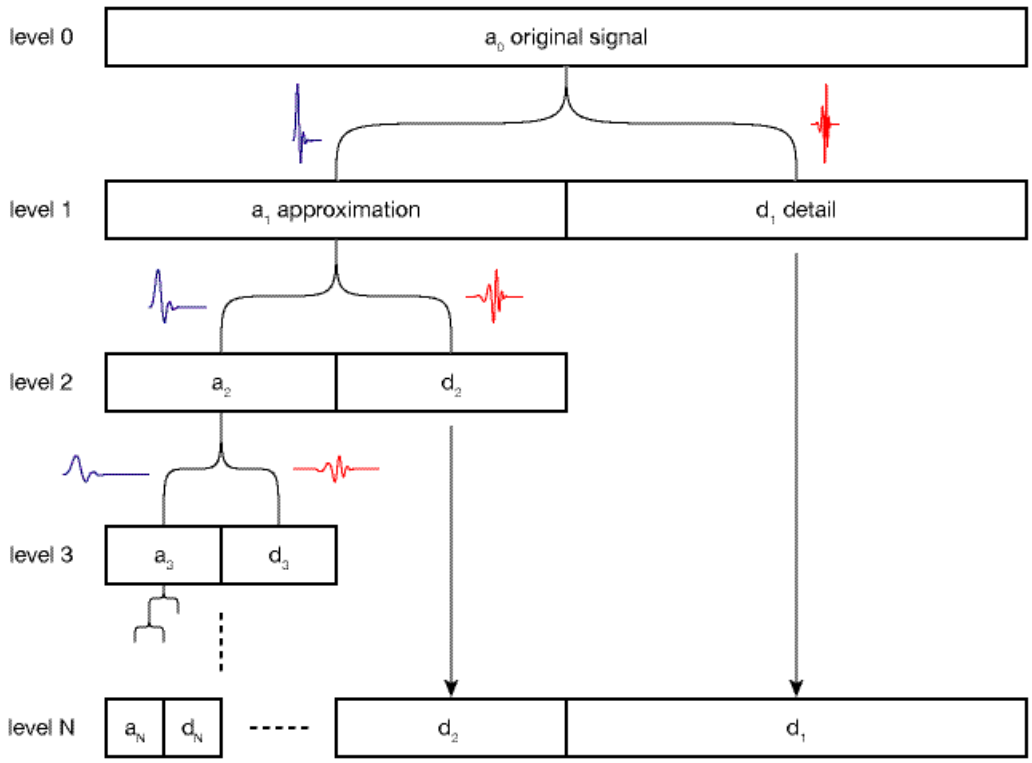

Figure 8 - Overview of the discrete wavelet transform and multi-resolution analysis scheme known as the *pyramid algorithm*.[7] The original signal is separated into low frequency and high frequency components, which comprise the signal *approximation* and *detail* information respectively. Each level decomposes the approximation information further, making each level of detail (d_i) a separate frequency band.

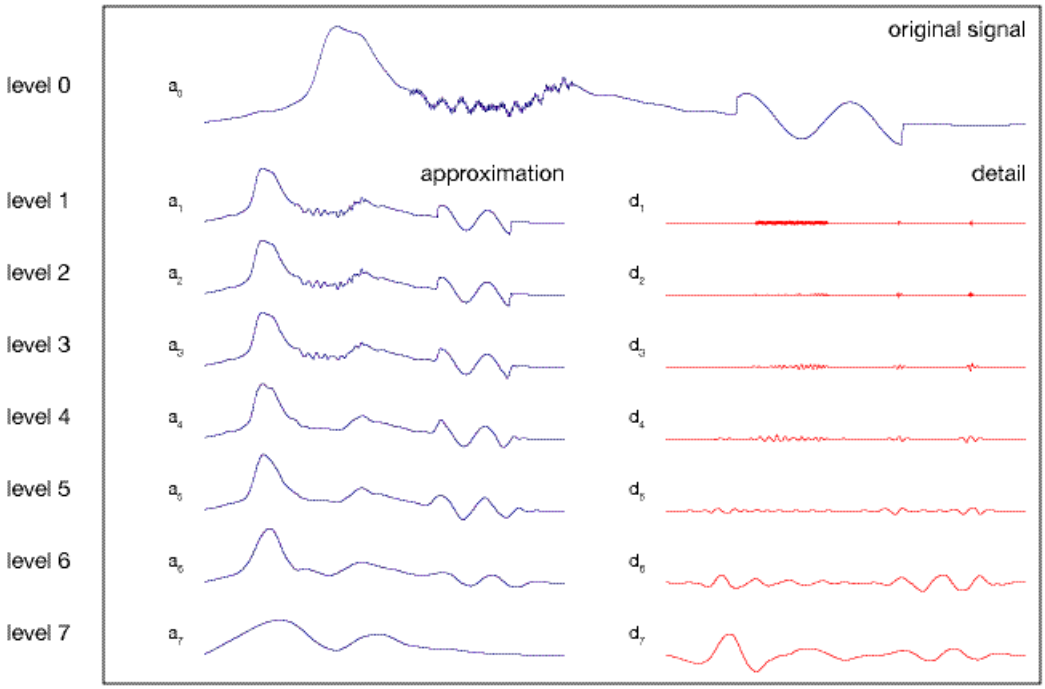

Figure 9 – This series illustrates multi-resolution analysis, separating out the high frequency information at each level of transformation in the pyramid algorithm (illustrated in Figure 8). Note that the approximation (a_i) signal in higher-level iterations contain much less detailed information, since this has been removed and encoded into wavelet detail coefficients at each discrete wavelet transform (DWT) deconstruction step. To reconstruct the original signal, the inverse DWT needs the wavelet coefficients of a given approximation level *i* (a_i) and all detail information leading to that level ($d_{1\text{-}i}$).

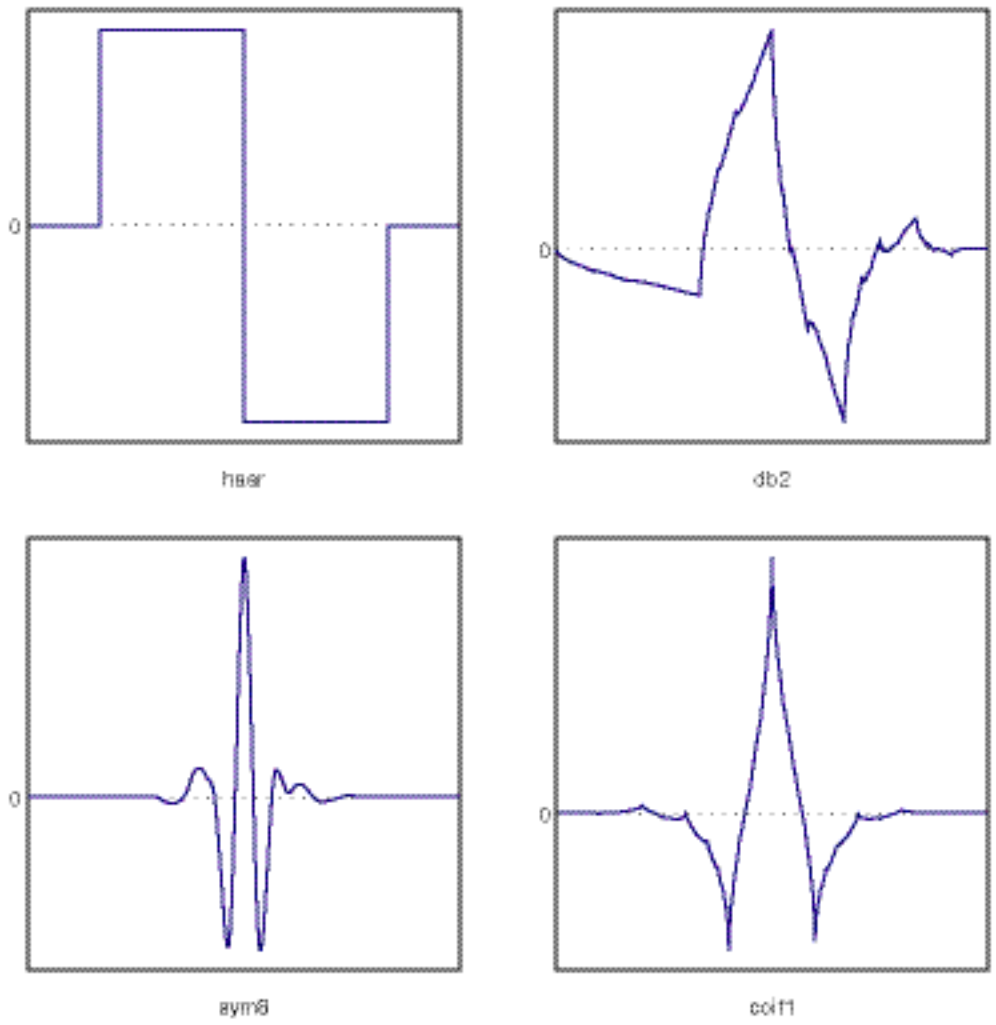

Figure 10 – Some families of wavelets used for multi-resolution analysis: (a) Haar, (b) Daubechies 2, (c) Symmlet 8 and (d) Coiflet.

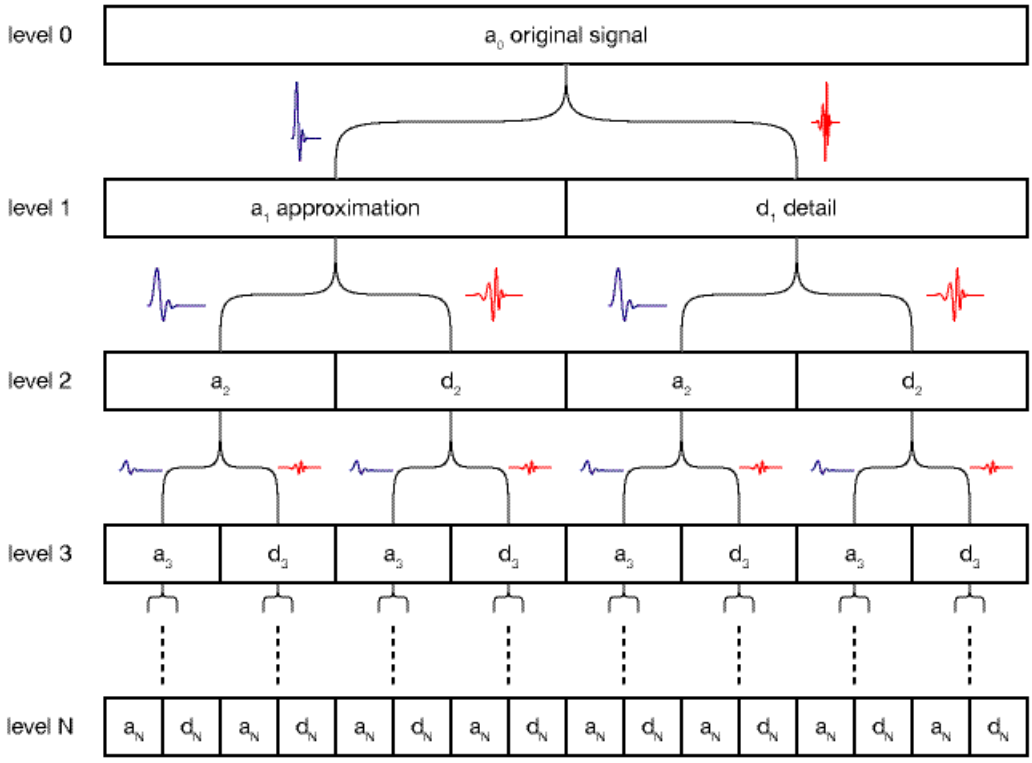

Figure 11 – Illustrated here is the wavelet packet transform (WPT). The pyramid algorithm enumerates a single branch (see Figure 8), while the WPT enumerates the complete tree of iterative decompositions. At each level, both the approximation and detail information are separated into low frequency and high frequency components.

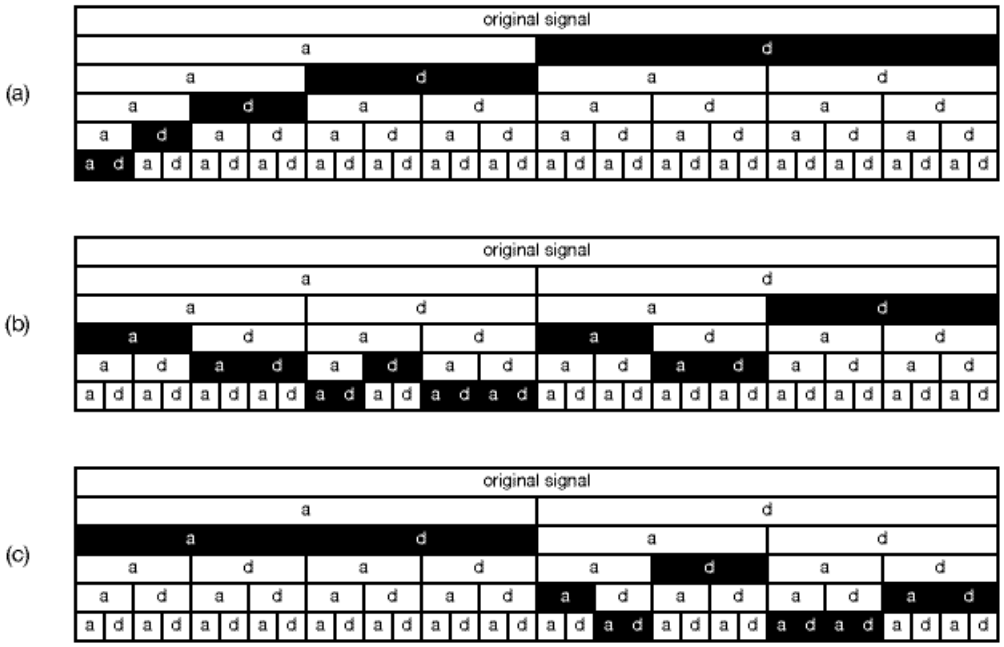

Figure 12 – Above is shown the signal representation or basis of the (a) pyramid algorithm, and two examples (b, c) of selected representations from the entire hierarchy computed by the wavelet packet transform (WPT). The WPT allows for optimal signal basis selection by enumerating the complete decomposition tree, so all signal representations can be evaluated.

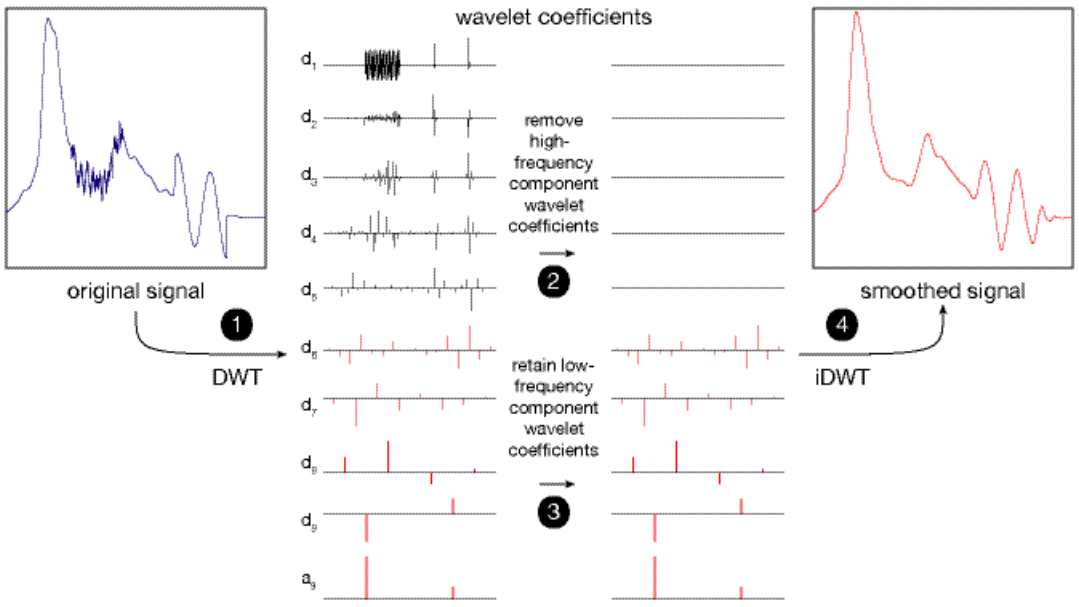

Figure 13 – Wavelet transform based smoothing has four steps: 1) transform the signal, 2) isolate the wavelet coefficients corresponding to the high frequency components (black wavelet coefficients), 3) "zero-out" or reduce these coefficients, and 4) apply a reverse wavelet transform to the signal. Compare this to the denoising routine illustrated in Figure 14.

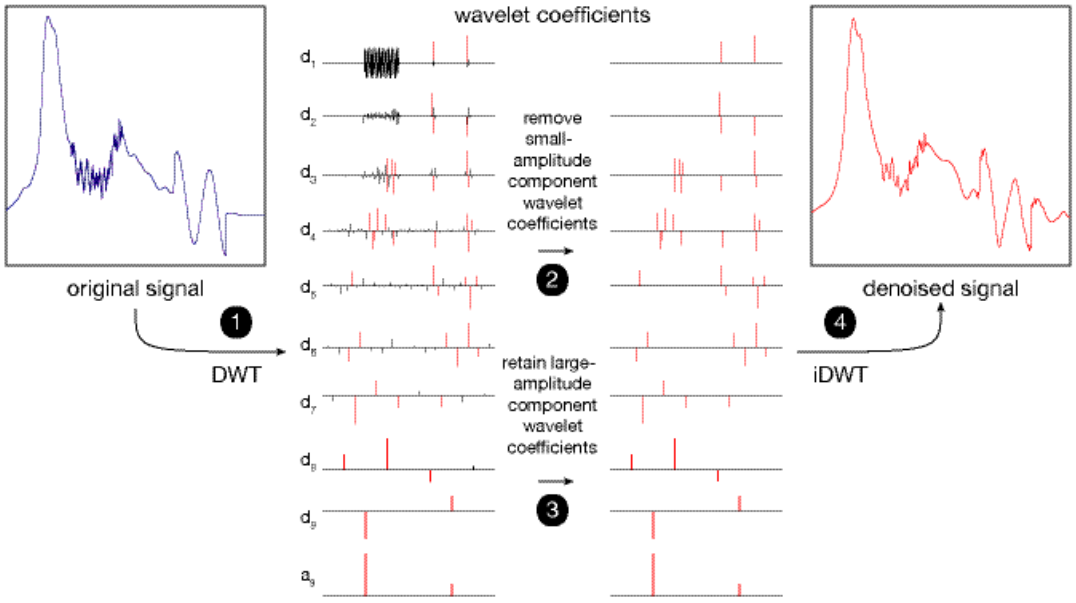

Figure 14 – Wavelet transform based denoising has four steps: 1) transform the signal, 2) isolate the small-amplitude wavelet coefficients corresponding to the noise components (black wavelet coefficients), 3) "zero-out" or reduce these coefficients, and 4) apply a reverse wavelet transform to the signal. Compare this to the smoothing routine illustrated in Figure 13. The isolation of small-amplitude coefficients in step 2 was achieved by using a progressive reduction hard-thresholding approach, which reduces the elimination threshold for each lower frequency band.

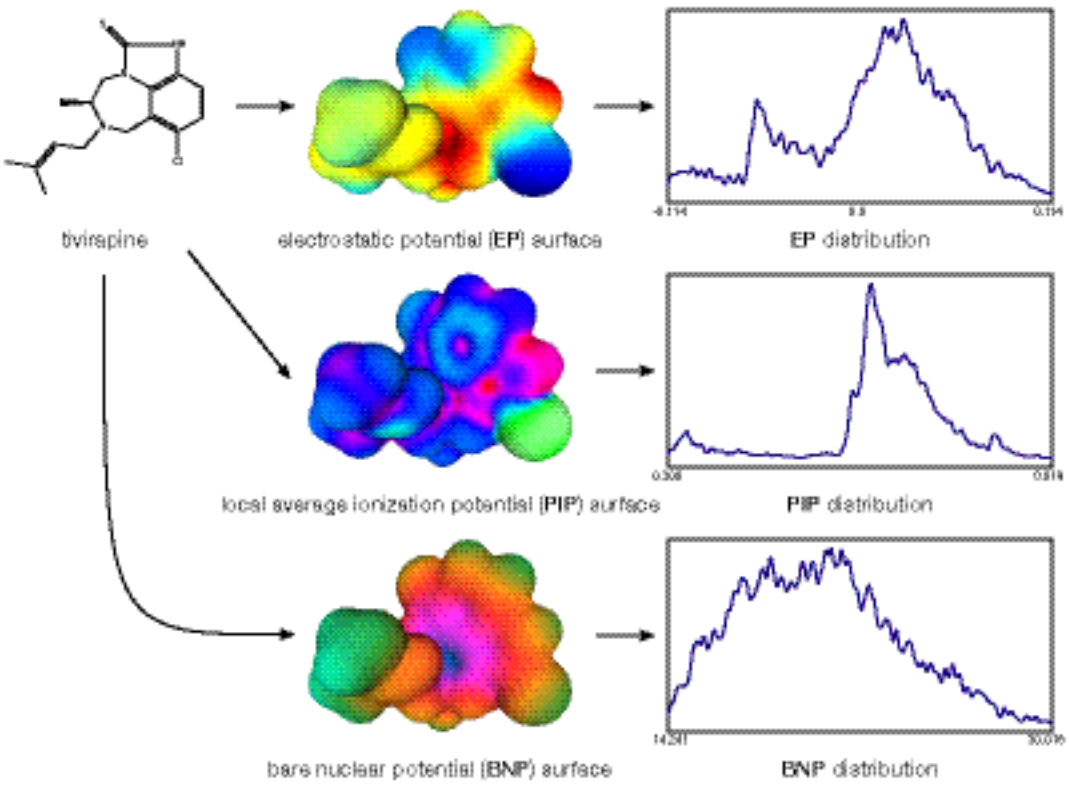

Figure 15 – Illustrated is the HIV reverse transcriptase inhibitor tivirapine, with its electronic van der Waals surface encoded with three electron density-derived properties and their respective property distributions.

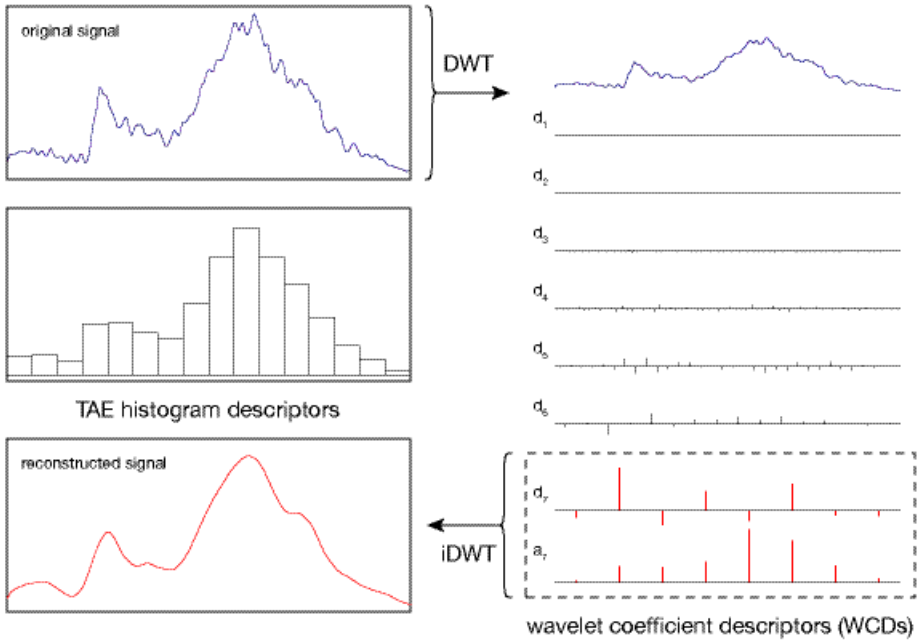

Figure 16 –Wavelet coefficient descriptors (WCDs) are generated as illustrated for each electron density-derived property. The property distribution is deconstructed using the discrete wavelet transform (*pyramid algorithm*), allowing the isolation of the lowest frequency and coarsest approximation coefficients (a_7 and d_7). These few coefficients are sufficient to reconstruct the majority of the original signal (via the inverse DWT), and contain the vital molecular property information needed for modeling. The WCDs replace the original TAE histogram descriptors and are orthogonal, consistent and representative.

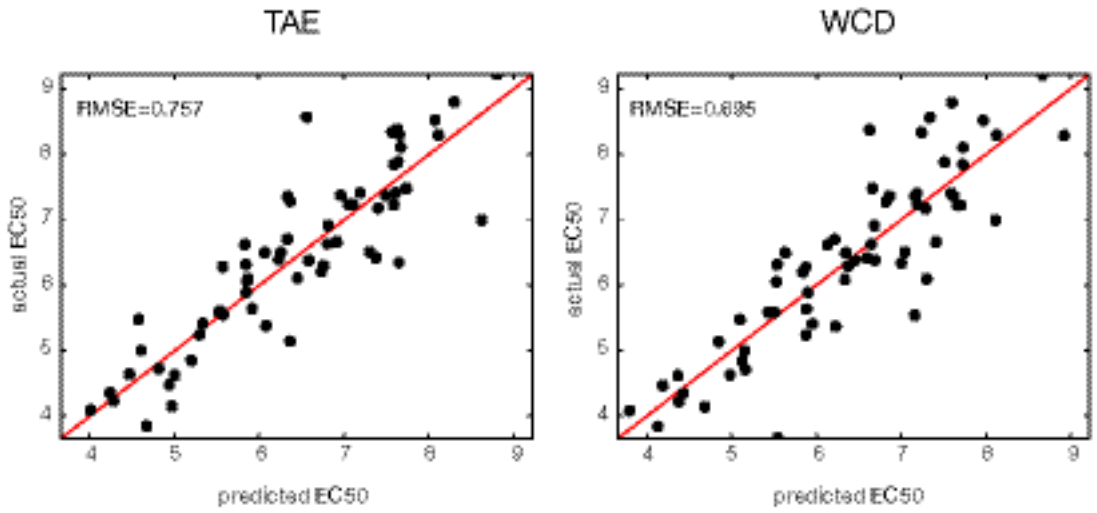

Figure 17 – Comparison of HIV reverse-transcriptase inhibitor (HIVrt) EC_{50} model parsimony using TAE and WCD descriptors. In each case, cross-validated GA/PLS models were constructed for a set of 64 HIVrt inhibitors using five PLS dimensions. The TAE model required 13 descriptors, while the WCD model needed only seven descriptors for similar performance.

View publication stats