An analysis of Microarray data using Wavelet Power Spectrum

S. Prabakaran, R. Sahu, S.Verma.

Abstract-- **Microarray technique facilitates the generation of large amount of data useful for solving many biological problems. Analyzing this vast amount of data needs great efforts for analysis. Usually, statistical methods like clustering are used to extract the common features among existing informal groups in a microarray data. But, dimensionality reduction and denoising the data are required for effective utilization for these methods. Hence better exploratory techniques are required for quick and effective inferencing. The present paper studies the capability of transform oriented signal processing techniques especially wavelet transform and wavelet power spectrum for feature selection and classification in microarray data. The suitability of wavelet based technique has been demonstrated on such datasets and innovative methods based have been proposed. The methods have been validated for feature selection and classification of binary datasets. It was observed that the proposed technique is more efficient and also requires no extensive preprocessing of data.**

Index Terms—**Wavelet power spectrum, microarray, feature selection, classification**

I. Introduction

Microarray is a modern technique which facilitates the simultaneous analysis of vast amount of gene expression data needed for solving complex biological problems. One such problem domain where microarray analysis is used in recent years is disease classification. The disease classification is the problem of categorizing the given sample into one of the sub classes of a disease type. These sub classes have already been established. But, processing this vast amount of microarray data for classification needs more effort due to its huge dimension and complex relationships among various genes. Another important aspect in microarray processing is that extraction of a global picture of the data rather than the picture of individual genes is more preferred. Usually, statistical methods like clustering are used to extract the common features among the naturally existing informal groups in a microarray data. These methods generally need dimensionality reduction through filtering irrelevant data and denoising the data for effective utilization. But, in these processes, some of the informative genes may also be removed which causes the inaccurate results on processing further. Also, complexity of present statistical methods used for microarray application like disease classification are more which emphasis the need of more computationally fast and simple algorithms. Understanding microarray characteristics may help in moving ahead in this direction. Till today, transform oriented signal processing techniques which are capable of bringing out the hidden characteristics of a data set are not probed sufficiently to develop more effective tools. The aim of this paper is to study the characteristics of microarray data especially data used for disease diagnosis problems using wavelet transform power spectrum and to access the potential of wavelet power spectrum in classification problem. A feature selection method based wavelet power spectrum and a method for classifying the samples using these selected features have been discussed in this paper. The paper is arranged as follows: The following section discusses the back ground information about wavelet and microarrays and its visualization. Later sections discuss the suitability of a wavelet for microarray processing and analyze the different elements of behavioural characteristics of microarray data under wavelet power spectrum. New methods based on this analysis for feature selection and classification have been proposed and tested on binary datasets. The proposed techniques are simple but effective and the results are encouraging for further research in the domain of microarray analysis using wavelets.

Manuscript received April,30, 2006.

S. Prabakaran is a research scholar in Indian Institute of Information Technology and Management, Gwalior, India.(e-mail: pra_spin2@yahoo.com).

R. Sahu is with Indian Institute of Information Technology and Management, Gwalior, India.(e-mail: rsahu@iiitm.ac.in.

S. Verma is with Indian Institute of Information Technology and Management, Gwalior, India.(e-mail: sverma@iiitm.ac.in.)

II. Background

A. Wavelet

Wavelets are a family of basis functions that can be used to approximate other functions by expansion in orthonormal series. They combine such powerful properties as orthonormality, compact support, varying degrees of smoothness, localization both in time or space and scale (frequency), and fast implementation. One of the key advantages of wavelets is their ability to spatially adapt to features of a function such as discontinuities and varying frequency behaviour. The compact support means that each wavelet basis function is supported on a finite interval and it guarantees the localization of wavelets. That is, a region of the data can be processed with out affecting the data outside this

region.

A wavelet transform is a lossless linear transformation of a signal or data into coefficients on a basis of wavelet functions [1]. In signal processing, a transformation technique is used to project a data in one domain into another where hidden information can be extracted. A wavelet function can be viewed as a high pass filter which approximates a dataset. A wavelet transform decomposes a signal into several groups of coefficients. These co efficient vectors contain information about characteristics of the data at different scales. Fine scales capture local details of coefficients and coarse scales capture global features of a signal. Performing the discrete wavelet transform (DWT) of a signal x is passing it through low pass filters (scaling functions) and high pass filters simultaneously. The result at each pass of the filtering of the signal is a convolution of the impulse response g of the filter and the signal. Mathematically, this result can be represented as $y(n) = \sum_{k=-\infty}^{\infty} x[k].g[n-k]$. The frequency of the signal is halved after passing the signal through a filter. So, by Nyquist's rule, half of the samples can be discarded. This is achieved by down-sampling or decimation by a factor 2, that is, removing every alternative coefficient in *y(n)*. Hence, after simultaneously passing a signal through high pass and low pass filters and the subsequent down-sampling, the number of coefficients will be equal to half the length of the original input for each filter.. Therefore, the wavelet transform of a signal for both high pass filters and low pass filters can be represented by the following two equations

$$y_{low}(n) = \sum_{k=-\infty}^{\infty} x[k].g[2.n-k]$$

$$y_{high}(n) = \sum_{k=-\infty}^{\infty} x[k].h[2.n-k]$$

In matrix form, wt =WX where W =[L;H] where L and H are impulse responses of low pass and high pass filters and wt is wavelet transform of the input signal X.

The two filters used at each stage of decomposition must be related to each other by $g[L-1-n] = (-1)^n.h[n]$ where g and h are the impulse responses of the two filters and L is such that $0 \le n < L$. These filters are known as quadrature mirror filters. The wavelet coefficients vector resulted from applying wavelet transform to a signal consists of both $y_{high}(n)$ (also called detailed coefficients) and $y_{low}(n)$ (also called approximation coefficients) coefficients in order. DWT proceeds further by recursively applying two convolution functions each producing an output stream that is half of the length of the original input, until the resolution (number of approximation coefficients) becomes one [1] or resolution level zero. Number of detailed coefficients at each level j is equal to $n/2^j$. The term 'scale' used in the context of wavelet transform at a level j is given by $\tau = 2^{j-1}$. Each detailed coefficient at a level tells us how much a weighted average of the data changes from a particular time period to next one. On the other hand, approximation coefficients are associated with averages of the data on scales $\tau_{J+1}\Delta t$ and higher where J is the largest level of wavelet decomposition for a signal and Δt is time interval between consecutive observations. The maximum level of decomposition depends on the wavelet function used for transformation. For example, the maximum level of decomposition of a signal x for Haar wavelet is given by $\log_2(x)$. Figure 1 depicts the entire process of DWT.

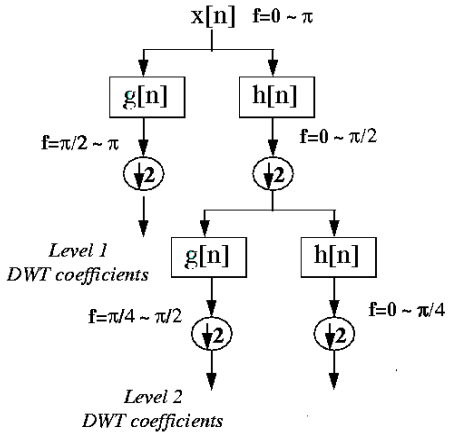

Figure 1 . A two level DWT for N data. The number of data is halved after every filtering and down sampling operation. A wavelet transform is applied on output of high pass filter (approximation coefficients) recursively keeping the output coefficients of each low pass filtering operation (detailed coefficients) at each stage. The wavelet transform of a data at any level n of decomposition consists of approximation coefficients only at nth level and all detailed coefficients up to nth level

A number of wavelet families like symlet, coiflet, daubechies and biorthogonal wavelets are already in use. They vary in various basic properties of wavelets like compactness. Among them, Haar wavelets belonging to daubechies wavelet family are most commonly used wavelets in database literature because they are easy to comprehend and fast to compute [2]. Haar transform can be viewed as a series of averaging and differentiating operations on a discrete function. The impulse response for high pass filter is given by $[-1/\sqrt{2}\ ,\ 1/\sqrt{2}\ ]$ and for low pass filter, the impulse response is $[1/\sqrt{2}\ ,1/\sqrt{2}\ ]$. That is, the minimum number of elements in input data should be 2 and the input data should always contain the number of elements 2^n where n is an integer. In matrix form, the Haar wavelet filter can be expressed as

$$\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$$

For a data having more than two elements, the Haar wavelet matrix of can be constructed by diagonally repeating this basic filter matrix to form a matrix of the size of input data.

Let an input data x = [5 3 2 1 -2 3 4 5]. The number of data points in the function is 8. Therefore, the maximum level of decomposition is $\log_2$ (8) which is equal to 3. The maximum scale is 2^{j-1} which is equal to 2.

The procedure to find Haar wavelet transform of x is shown in Table 1**.** The Haar wavelet transform of *f(x)* at the maximum level of decomposition H(*f(x)*)=[7.4246 0.3536 2.5000 -4.0000 1.4142 0.7071 -3.5355 -0.7071]. That is, the wavelet transform of a data at any level n of decomposition consists of approximation coefficients only at nth level and all detailed coefficients up to nth level In this example, the Haar wavelet transform at the 2^{nd} decomposition level (resolution level 1) is [5.5000 5.0000 2.5000 -4.0000 1.4142 0.7071 -3.5355 -0.7071]. The same result can be checked using matrix method also.

Resolution	Resolution level	Approximation coefficients	Detail coefficients
8	3	5, 3, 2, 1, -2, 3, 4, 5	-----
4	2	5.6569,2.1213, 0.7071, 6.3640	-1.4142, 0.7071, -3.5355, -0.7071,
2	1	5.5000 5.0000	2.5000 -4.0000
1	0	7.4246	0.3536

Table 1. **An example of one dimensional Haar wavelet transform**

The wavelet transform is a cumulative measure of the variations in the data over regions proportional to the wavelet scales or resolutions. Since a wavelet transform of a data at n level contains the approximation coefficients at nth level and all detailed coefficients, which incorporate the variations in information, up to nth level, analyzing the data at different subspaces formed by different scales is possible so that a proper subspace can be chosen for a specific application task to get a balance between accuracy and efficiency [3]. Hence, wavelet transform may be used to identify or to localize at what subspace of wavelet transform of the microarray data contain better details so as to use this information for analyzing microarray data for various biological purposes like classification. Alternatively, at what level of decomposition, the genes show different behaviour useful for classification or pathways may be identified and further probe may be continued based on this information.

The use of wavelet transforms provides economical and informative mathematical representations of many objects of interest [4] like images and audio signals. Also, the accessibility of wavelets has been made easier through many easily available software packages. Surveys of wavelet applications in biological data and in data mining are presented at [5, 6, 3] respectively. But, the spectral approach was not accounted or discussed in these papers. Mathematical details of wavelets may be referred at [7-9]. Applications of wavelet analysis have already been established in many critical areas like image processing and pattern recognition. In the field of study of biosignals, the use of wavelet analysis is in initial stage only.

Wavelet power spectrum

Wavelets facilitate multi level decompositions for a data set. Multi level decomposition refers to recursively applying wavelet decomposition to the data more than one time. The maximum level of decomposition depends on wavelet function selected. At each level of decomposition, as explained in the background section, variations in the data over regions are measured proportional to the scales. Hence, the cumulative measure of such variations in the data in various scales can be analyzed for extracting some useful information, if any, for finding a way to resolve the problem under scrutiny to be solved using the data. In case of microarrays gene expression data, one such general objective is to extract the differential genes which show some difference in different diagnostic category of the disease dataset. Since wavelet transforms deal with the variations in a dataset as explained above, analyzing wavelet transform at different decomposition levels may provide such an outlook to identify the differentially expressing genes. The reason for analyzing at various levels is that the variations among genes may be not be the same at all levels of decomposition and may be quite high enough at some level. So, a tool for such analysis is indeed needed and wavelet power spectrum is such a tool for this purpose.

Wavelet power spectrum is a graphical representation useful for analyzing the result of application of a wavelet transform on a data. It can be used to analyze the information variation of a data under various decomposition levels since each level incorporates some variations among data points. This feature of a power spectrum may be useful in identifying the characteristics of a microarray data. In addition to provide a visualization of features of a microarray data, the global information variation of gene expression in a given sample can be consolidated by plotting wavelet power spectrum and it may reveal many other hidden structures of the data on further analysis.

Local wavelet power spectrum at a particular decomposition level is calculated by summing up the squares of wavelet coefficients at that level [10]. For a set of wavelet coefficients $C_{j,k}$, where j is level of decomposition and k is the order of the coefficient,

$$\text{spectrum}[j]=\sum_{k=0}^{2^j-1} C_{j,k}^{\ 2}$$

If there are N elements in an array, there will be $\log_2(N)$

coefficient bands or levels of decomposition . That is, the power spectrum can be referred as a graphical representation of cumulative information variation at each scale of decomposition.

The present paper especially concentrates on various types of cancer microarrays in the context of classification problem which is very important in clinical diagnosis and prognosis. The scope of this paper is to analyze the capabilities of wavelet power spectrum to reveal more useful information, in addition to visualization like heat map, regarding the characteristics of microarrays generated for disease classification problems and to explore the possibilities of developing simpler tools in this problem domain. The technique is very simple and proved effective in many data analysis issues and is easy to be implemented and requires no special software.

Microarrays and its visualization

The use of microarrays to study gene expression profiles in biologic samples started in 1995 [11]. Microarray technology facilitates the monitoring of expressions of thousands of genes across different conditions, times or tissue samples simultaneously. Microarrays are based on the principle of the hybridization technique was developed by Southern [12].

A microarray is typically organized as m x n matrix spots [13] of sizes of about 300 microns or larger, on a nylon filter, on a glass slide or on a silicon chip and can be easily imaged by existing gel and blot scanners. In each array spot, single stranded pieces of known (c) DNA are attached. When a mixture of labeled copies of mRNA purified from the query tissue is applied on these spots, the label matching that in the spot is localized after hybridization. The hybridization results are imaged and the image thus formed is converted into physical matrix using appropriate scanners.

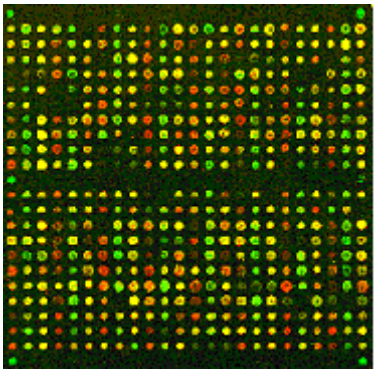

Figure 2. A sample Microarray Image.

As shown in Figure 2, a microarray image has n x n matrix spots where known sequences of DNA are immobilized. When a mixture of labeled copies of mRNA purified from the query tissue is applied on these spots, the label matching that in the spot is localized after hybridization Various intensities of the spots illustrates the various intensities resulted by hybridization .

In most of the microarray experiments, irrespective of the objective of the experimentation, the major aim is to identify the differentially expressed genes to use them for further processing and to filter irrelevant data. That is, the techniques are generally sensitive to noise and are computationally intensive.

The most common tools used for visualizing the differentially expressed genes in a microarray data are heat maps introduced by Eisen et al [14]. In a heat map, expression values of a gene in different samples are depicted as colored rectangle. The scope of heat maps is limited to just illustration of differential expression of genes (see Figure 3). They do not provide any detailed information such as characteristics of samples. A visualization technique which may bring out the complex hidden structures is desirable for a better analysis.

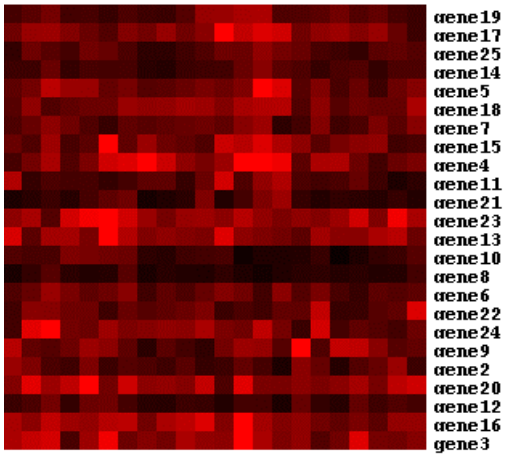

Figure3. Heat map of first 25 genes in first 2 samples of Hedenfalk breast cancer data. The genes are rearranged by hierarchical clustering method. Intensity from black to red indicates the increasing order of expression level.

III. SELECTION OF SUITABLE WAVELETS

Selection of correct parameters is critical for the success of a method.. Statistical methods used in microarrays need various parameters like correct number of clusters in partition based methods. But the selection of such input parameters needs priori knowledge about the data for successful application [15]. For example, to select the correct number of clusters, one should have the knowledge of natural clusters present in the data. But, in the case of wavelet, selection of a suitable wavelet doesn't need such complex details.

The main criterion is that the wavelet chosen has to approximate the data more closely. Through wavelet power spectrum itself, it can be identified. No additional information about the nature of the data is needed as in the case of statistical methods. A wavelet approximating a data well will produce minimum value of the spectrum in the farthest possible band, that is, at the maximum level of decomposition. For example, refer Figure 4.

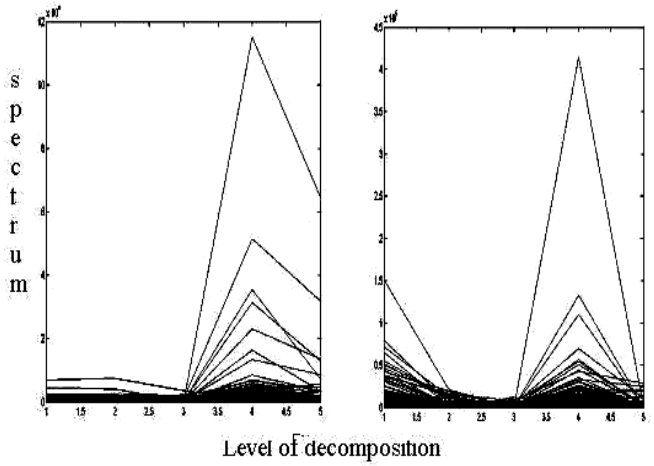

Figure 4. Wavelet power spectrum of SRBCT dataset from Khan et al using db2 and Biorthogonal wavelet 1.3 respectively. Biorthogonal wavelet 1.3 is more suitable to anlayse the given data than db2 since the spectral values at the farthest level of decomposition (band) is comparatively lesser than those for db2.

The spectra generated using db2 and biorthogonal wavelet 1.3 [16] for SRBCT dataset are depicted in Figure 3. It has minimum value for most of the genes in the spectrum generated by biorthogonal wavelet1.3 than that in spectrum generated using db2 at farthest band along x-axis. So, first wavelet is more suitable than the latter.

IV. BEHAVIOUR OF SAMPLES OF VARIOUS CANCER TYPES

A variety of cancer types have been identified in men and women. Each type of cancer affects some part of the human body. While traditional methods like Immunohistochemistry can examine the proteins one by one and take long time to complete a diagnostic test which may become very late to save the patient, microarray gene expression profiling permits a simultaneous analysis of multiple markers. The nature of pattern present in gene expression profiling of different cancer types may not be the same. To study the behaviour of different cancer types under wavelet power spectrum, gene expression profiles of various cancer types like Small round blue cancer tumor samples (SRBCT) from Khan et al [17], breast cancer samples from Bhattacharya et al [18], Diffuse large B-cell lymphoma (DLBCL) samples from Alizadeh et al [19] and Leukemia cancer samples from Golub et al [20] were subjected to a wavelet transform and their corresponding spectra were plotted (See Figure 5). For a comparative study, the Biorthogonal wavelet 3.5 was utilized to make the results comparable. This does not mean that this wavelet is best suitable for all datasets. The best suitable wavelet for each dataset may actually be different. Among these datasets, Golub's Leukemia dataset is a binary class dataset and others are multi class datasets. The visualization of the microarray data using wavelet power spectrum is better than heat map visualization since heat maps just illustrate the differential expression while wavelet power spectrum is helpful to infer many other factors like distinct global common trends present in the wavelet power spectrum of different cancer datasets which is not possible through heat maps. These trends are obvious from Figure4. Almost all the samples belonging to a particular dataset follow the same trend. Little variations observed may be due to the fact that the datasets were used without denoising. There are least chances to expect the variability in these trends since whole the wavelet spectrum of each sample wise data (column in microarray data) is plotted in Figure 5.

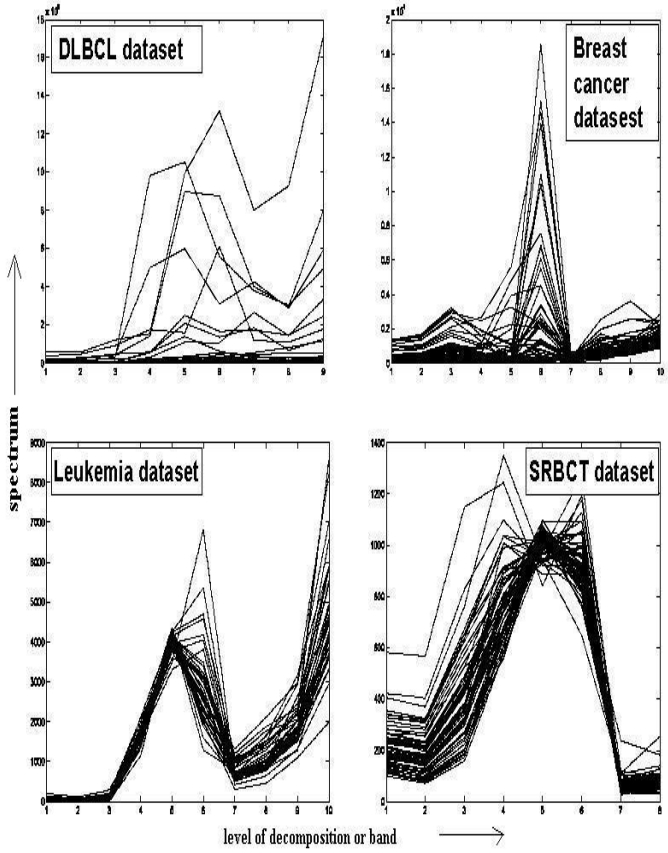

Figure 5. Trends in wavelet power spectrums of different cancer datasets for Bi-orthogonal wavelet 3.5. 'Level of decomposition' refers to the number of times the wavelet transform is applied recursively. It is observed that almost samples of each type of cancer class follow the different trend in a wavelet spectrum and sample belonging to the same class follow the same trend even though variation in magnitudes are observed

Apart from the heterogenic nature of different samples, some characteristic peaks are also exhibited in each datasets. The number of characteristic peaks in each dataset also is observed differently. Probing further at these peaks may reveal many new factors. Among multiple characteristic peaks one appears to be dominant than all. Also, the spectrum in different datasets spreads over different range of bands.

In some datasets, the characteristic peaks of almost all

samples converge at a particular band and those of the samples in some datasets, like found in DLBCL dataset, are observed to be not convergent at a single band. Instead, they spread over in different peaks. So, we may analyse the DLBCL dataset using some other wavelets for better approximation. In breast cancer dataset, the cumulative information of samples is amplified at the levels of decomposition 3, 6 and 9. Among them, in band 6, the information variation is observed well amplified. So, one may analyse the wavelet transform of this data at this subspace for extracting better accurate information variation about various samples necessary for sample classification.. Similarly the spectra of other datasets may also be analysed to get guidelines for further research. Since the aim is this paper is to demonstrate the capabilities of wavelet spectral technique and to provide a guidelines, we do not do further research and do not present any conclusion.

V. Behavioural characteristics of genes in wavelet power spectrum

The global analysis of samples through visualization of wavelet power spectrum explores the possibilities of developing new techniques based on wavelet power spectrum and provides a guide point to select proper subspace for effective analysis as seen in previous section. A closer look on the behavioural characteristics of genes in a given dataset may further enhance this idea. Plotting the power spectra of all genes in a dataset together explores the presence of common characteristics and some distinct trends among genes of various datasets. Genes having common characteristics may be grouped together and analyzed separately. Distinct trends can be probed further for further applications. This may help in clustering genes .For illustrating this view, the power spectra of all genes in Golub leukemia dataset were plotted and presented in Figure 6.

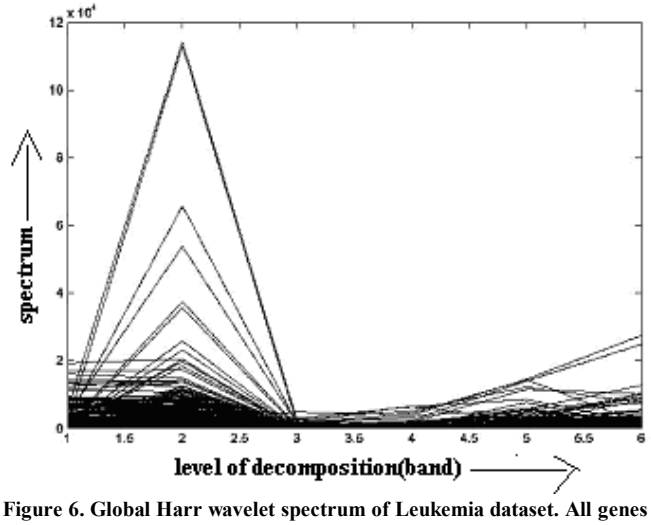

Figure 6. Global Harr wavelet spectrum of Leukemia dataset. All genes have a common trend of having minimum at band 3. But, from band 1 to 2, two different trends are observed.

We used the simple Haar wavelet transform for calculating the spectra of various genes. Since wavelet power spectra of all 7129 genes were plotted together, the global nature of the genes were able to be analysed. It is observed from Figure 5 that all genes behave similarly at the subspace defined by band3 in that all genes have a minimum at band 3. Another similar behaviour of these genes is evident at band 2. Most of them have a maximum in this band and the information variance in this subspace is comparably higher than in other subspaces. Hence further probe may be focused at this subspace of wavelet transform. At the same time, all genes so two types of behaviour in the subspaces defined by band 1 and band 2. A group of these genes exhibit comparatively very lesser spectral values at band 1 than that at band2 while another group of genes exhibit very little difference or almost equal in their spectral values at band 1 and band 2. Since the natural clusters in this dataset is two, this alternative trend between band 1 and band 2 is notable one and further analysis in these sub spaces may provide a better insight in clustering these genes.

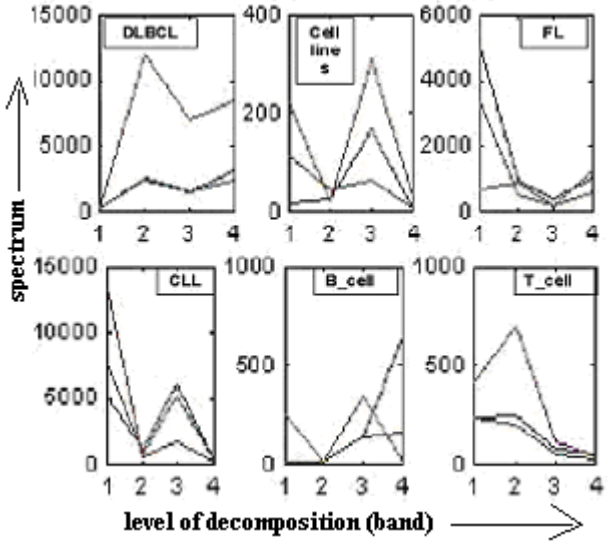

Figure 7 .Power spectrums of first four genes of DLBCL datasets in different diagnostic category which exhibits the different trends among genes. The genes may be clustered into various clusters based on these trends

Another illustration to explore the behavioural patterns of genes depicted in Figure 7 uses DLBCL dataset. But, a little difference between this illustration and the previous one is that the spectral behaviour of genes in various diagnostic categories was studied instead of that in whole dataset. The wavelet power spectrum of first four genes of DLBCL datasets in different diagnostic category is shown in Figure 6. It is observed that the spectral behaviour of the genes showed similar trend within the same diagnostic category even though some difference in magnitude of the spectra was observed. This observation along with observations concerned to information variation in different bands may be utilized for classification of diagnostic categories. For instance, two diagnostic categories of DLBCL samples DLBCL and T_cell show a maximum at subspace defined by band 2 This property can be used as a guide point to distinguish theses categories against all other categories.

Combining this property with the another fact that DLBCL has lesser spectral value at band 3 than that at band 4 and T_cell exhibits just opposite behaviour both these diagnostic categories can be classified. Also, clustering of genes useful for classification of different diagnostic categories may be performed on the basis of the various spectral trends observed in these spectra. Thus, the wavelet power spectra of genes appear to have potential to provide more information and guidelines to develop simpler and efficient tools for solving clustering as well as classification problems which is not possible with the traditional heat map visualization.

VI. CLUSTERING OF GENES USING WAVELET POWER SPECTRUM

The wavelet power spectrum of all genes projected together displays the global trends existing among various genes in a dataset. The global trend thus observed is helpful in separating the group of genes exhibiting similar behaviour. The genes in a dataset exhibit more information variation at some subspaces. These subspaces may be identified through visual inspection of the power spectra of all expressing genes projected together. On the basis of the information variation at the subspaces thus identified, a strategy to cluster the genes with similar behaviour can be developed.

The binary class datasets were considered in the present work to analyse the usefulness of the wavelet power spectrum in clustering the genes. Our proposed strategy of the clustering of genes was found efficient but simple. Two consecutive subspaces, where the different trends in information variation were the most, were identified. Some genes had more spectral values in one subspace than in another. The reverse to this trend was also true. The genes were separated according to these trends into two clusters. That is, genes having higher spectral values in the first subspace in comparison to the second subspace were grouped together in to a cluster. The genes having lesser spectral values in the first subspace in comparison to the second subspace were put in the second cluster. The subspaces are determined by the levels of decomposition.

The next task was to identify the labels of the clusters formed. To accompany this task, the expressions of the genes in a cluster were divided into two groups. First group contained the expressions of the genes in one category of the samples and another group contained the expressions of the genes in another category of samples in the dataset under consideration. The average spectral value of the genes per sample was calculated for these two groups. The number of genes having the higher average spectral value per sample was calculated for each group separately. The label of the group for which more number of genes has higher average spectral value per sample was assigned to the cluster. The same procedure was repeated to the genes in the second cluster also. It was observed that the procedure presented here exactly labeled the two clusters with alternate labels present in the binary class dataset under consideration. The method was tested on the datasets used in previous works [20] [21]. The clusters thus formed had been observed to contain the respective genes identified as important for classifying the samples of the category for which the cluster was assigned.

VII. AN ANALYSIS OF FEATURE SELECTION BASED ON WAVELET POWER SPECTRUM

Feature selection is the problem of selecting a small number of genes which are capable of classifying the samples into different categories. This is a necessary precursor for classification problem. The vast amount of irrelevant data in view of the classification task is filtered by the feature selection. This is important to increase the speed and efficiency of a classification method. The statistical methods are in wide use in this domain [22-24]. But, they require extensive preprocessing and the complexity of such statistical algorithm is also higher. As an alternative, the capability of the wavelet power spectrum in feature selection has been analysed in the present work and a simple feature selection method has been proposed.

The proposed method uses the clusters of genes formed for a dataset as described in the previous section. The method is quite simple. It ranked the genes and the top genes were tested for classification. The expressions of the genes in a cluster were divided into two sub groups. One group contained the expressions of the genes in the samples of a diagnostic category. Another group contained the gene expressions in another diagnostic category present in the binary class dataset under consideration. The average spectral values per sample for each gene in two diagnostic categories were calculated. The ratio of the average spectral values of a gene in a diagnostic category to another diagnostic category was calculated. On the basis of the ratio thus calculated the genes were ordered in descending order. The top ranked genes occupied the top slots. The method was tested on such binary cancer datasets used in earlier works [20] [21]. Most of the genes selected as informative in the earlier works were observed to be at the top slots of the list of ranked genes. That is, more quality genes useful for classification were observed to be at the top slots of the rank list. This shows the possibility of developing simple but effective feature selection methods based on wavelet power spectrum.

VIII. A CLASSIFICATION ANALYSIS BASED ON WAVELET POWER SPECTRUM

The classification of gene expression data samples involve generally two steps: feature selection which is a necessary precursor for classification to identify potentially relevant genes for classification and a classifier design to classify the data into known classes using the features selected. The recognition process of classes is achieved in two phases: training phase and testing phase. In training phase, a classification model is built using a subset of microarray data and their labels. In testing phase, the remaining microarray data unseen in training phase are used to test the validity of the classifier for its ability

to classify. The accuracy of the classification has strong impact on feature selection for knowledge discovery applications [25].

Statistical methods of classification are complex and needs more preprocessing of the data. Also they do not fit for a wide variety of datasets. They also coupled with a feature selection method. Many input parameters and priori knowledge of the dataset to choose a better classification method is required. The classification error is also some what higher. Artificial intelligence methods such as ANN [26] [27] perform better classification. But, they require properly defined architecture and learning parameters and enough training samples. Also, the presence of noise in the data affects their performance. The present method is the effect of the search to access the capability of wavelet power spectrum in developing more robust and simple classification methods. More sophisticated methods may be developed on further research in this domain.

The present method of classification is the follow up of the feature selection method described in the previous section. The list of ranked features prepared in feature selection step favourable for classifying a diagnostic category using training dataset was used. The average spectral values of the ranked genes in two diagnostic categories in the training dataset were calculated. The spectral values of the genes in each sample also were calculated. To test a sample if it belongs to the diagnostic category of the cluster of the genes used, for each ranked genes in the list, the absolute difference of the spectral value in the sample and its average spectral value per sample in a diagnostic category was calculated. A merit spectral distance was calculated using this difference. The spectral distance was defined as the difference of the absolute difference of the spectral value with one. It was assigned as D_{cl}. Similarly, the spectral distance of a gene in the samples of another diagnostic category was calculated and was assigned as D_{ot}. The number of genes having D_{cl} lesser than D_{ot} was calculated for top five genes. If this number was greater than 50% of the genes used, that is 3 or more, the sample was allocated to the diagnostic category for which the cluster belongs to. This procedure was tested for all samples in the training dataset and the number of correctly classified samples was counted. If it was not equal to the actual number of samples present in each diagnostic category in the training test, the training step was over. Otherwise, the procedure was repeated with top 10 genes and so on. Finally, a number of top ranked genes that could more correctly classify the samples in the training dataset was identified.

To test the validity of the method, the classification test as discussed here was performed on samples in the test dataset with the top genes established in training step. It was observed that the classification accuracy for datasets used in the present experiment was quite high despite the simplicity of the present method. The method was observed quite effective. Also, the robust identification of number of top ranked genes useful for classification is an added advantage of the present in comparison with the most of the other classification methods where no such clear idea of number genes to be used was established.

IX. Experimental results

A. Acute Leukemia dataset

The first dataset used to test the capabilities of the wavelet power spectrum based techniques discussed above was Acute Leukemia dataset from Golub et al []. This dataset consists of 7129 genes. The training dataset consists of 27 ALL and 11 AML samples. The test dataset consists of 20 ALL and 14 AML samples. More details about the dataset and downloading of the dataset is available at [20].

Rank	Gene index	ratio	Rank	Gene index	ratio
1	4342	454.3571	16	1909	29.5720
2	5290	146.325	17	5772	27.4925
3	2642	124.4613	18	1630	27.1150
4	6510	119.6607	19	5062	25.9463
5	4050	98.01723	20	4535	25.1394
6	1010	90.429	21	3562	23.4597
7	2733	71.252	22	2215	23.1003
8	5442	53.63562	23	3688	22.6295
9	6573	50.13584	24	3334	21.8121
10	4375	42.81756	25	1144	21.1314
11	6575	39.93943	26	6167	18.9714
12	1604	36.46125	27	6855	18.7640
13	524	35.31253	28	6895	18.6301
14	5146	35.24931	29	2348	18.2397
15	3056	29.66779	30	6623	17.9830

Table 2 . The list of 30 top ranked genes and their rank for Golub et al's dataset. The 'ratio' in the table means the ratio of the average spectral value per sample in ALL samples to average spectral value in AML samples in the training dataset.

The clustering, feature selection and classification were performed as described in the previous sections. The cluster used for analysis was the cluster labeled as ALL cluster. The

identification of the cluster label was described in the previous sections. In the original work of Golub et al, the number of genes used to predict the samples was 50. This 50 genes predictor was able to 36 of 38 samples in the training set correctly. The number of samples correctly predicted in test set was 29 out of 34. But, the present method based on the wavelet power spectrum approach was able to predict better than the original method described in Golub et al. Also, the present algorithm does not need any parameters like weight factor used in Golub's method and of less complexity.

Two samples were wrongly classified in the test set while using top 20 genes. The minimum number of genes satisfying the classification criteria, as described in previous sections, in the training dataset to separate the samples into ALL in the training dataset was 15. The same number was 18 while used top 25 genes. While top 30 genes were used, this number was 21. With top 25 genes, samples in training dataset were classified with out any error and only one sample was misclassified. But, with top 30 genes, all samples in both training and test datasets were classified correctly. So, the present method performs better. The top 30 genes selected are listed in Table 2**.**

B. DLBCL dataset

Next binary dataset used in the present work was DLBCL binary dataset from Shipp et al[21]. This dataset consists of 7129 genes and 77 samples. Out of these 77 samples, 58 samples belong to DLBCL diagnostic category and 19 samples belong to Fl diagnostic category. The dataset can also be downloaded from www.bme.jhu.edu/~actan/KTSP/Gene_ expression_data. To test the present methods, the dataset was roughly divided into two subsets: training and test datasets. The training dataset was created using first 32 samples from DLBCL category and 16 samples from FL category. The remaining samples were assigned to test dataset. The clustering, feature selection and classification were performed as described in the previous section. The cluster used for analysis was the cluster labeled as DLBCL cluster.

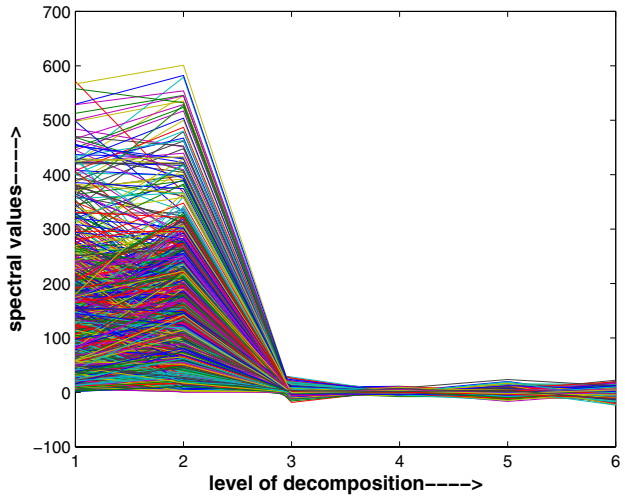

Figure 8. Global Harr wavelet spectrum of DLBCL dataset.. But, from band 1 to 2, two different trends are observed.

The behaviour of total genes for the training dataset is depicted in Figure 7. It is obvious from Figure 7 that the information variation among the expressing genes of DLBCL is prominent from sector 1 to sector 2along x axis. So, further analysis may be preceded in these sectors. The minimum number of genes satisfying the classification criteria, as described in previous sections, in the training dataset to separate the samples into DLBCL and FL in the training dataset was 16 while top 20 genes were used. For a lesser number of top ranked genes, the error was more. Similarly, use of more than 20 genes did not improve classification further. The list of top 20 genes selected was presented in Table 3.

Rank	Index no	ratio	Rank	Index no	ratio
1	203	34.7705	11	6322	12.7541
2	6656	33.0930	12	1430	12.6254
3	4153	21.4702	13	6659	12.5339
4	2043	20.6158	14	1704	12.1709
5	658	17.6826	15	1173	11.8012
6	4024	16.4915	16	1092	11.4246
7	5458	15.2856	17	2380	10.9998
8	6815	15.2437	18	5935	10.7911
9	3005	14.9182	19	4028	10.6505
10	1800	14.1341	20	4340	10.4755

Table 3 . The list of 20 top ranked genes and their rank for Shipp et al's dataset. The 'ratio' in the table means the ratio of the average spectral value per sample in DLBCL samples to average spectral value in FL samples in the training dataset.

Two samples in the training dataset, one from DLBCL category and one from Fl category were misclassified. In the test dataset, only one sample was misclassified. Remaining 74 samples were correctly classified.

X. CONCLUSION

Wavelet power spectra of microarray gene expression data exhibit the potential of wavelet transform and particularly its power spectrum technique in solving various biological problems like classification of data. The present paper demonstrates the capabilities of wavelet power spectrum in microarray data analysis. Visualisation of microarray data through wavelet power spectrum at various levels exhibits greater information regarding the nature of the data which is not inferred with use of heat maps. The different trends and magnitudes observed at various levels of resolution or bands are helpful in selecting proper subspaces of wavelet transform which may be researched further for extracting simple guidelines useful in developing efficient, fast, and simplistic tools for further applications in gene expression data. The present work demonstrates such simplistic methods based on wavelet power spectrum for clustering, feature selection, and classification of microarray dataset for binary datasets. The results are encouraging, effective, and exhibit the possibility of

developing more advanced tools for using wavelets for the analysis of microarray datasets.

REFERENCES

[1] S. Mallat,"A Wavelet Tour of Signal Processing", Academic Press, San Diego, January 1998.

[2] M.Garofalakis and P.B. Gibbons, "wavelet synopses with error guarantee", Proceedings of 2002 ACMSIGMOD, Madison ,Wisconsin, USA, June 2002, ACM Press.

[3] T. Li et al, "A survey on wavelet applications in data mining", SIGKDD explorations, Vol.4, Issue 2: 2002, pp.49-68.

[4] F. Abramovich, T. Bailey and T. Sapatinas, "Wavelet analysis and its statistical applications", JRSSD, Vol.48, 2000, pp. 1–30.

[5] A. Aldroubi and M. Unser editions, "Wavelets in Medicine and Biology", CRC Press, Boca Raton, 1996.

[6] P. Lio, "Wavelets in bioinformatics and computational biology: State of art and perspectives", Bioinformatics, Vol. 19, 2003, pp.2–9..

[7] G. Strang, "Wavelets and dilation equations: A brief introduction", SIAM Review, Vol.31 (4), 1989, pp. 614–627.

[8] I. Daubechies, "Ten Lectures on Wavelets", Capital City Press, Montpelier, Vermont, 1992.

[9] C.K. Chui, "An Introduction to Wavelets", Academic Press, Boston, 1992.

[10] Ian Kaplan, "Spectral Analysis and Filtering with the Wavelet Transform" Available: http://www.bearcave.com/misl//misl_tech/wavelets/freq/index.html.

[11] E.K. Lobenhofer et al, "Progress in the application of DNA microarrays", Environ Health Prospect, Vol. 109(9), 2001, pp. 881–89.

[12] E. Southern, "Detection of specific sequences among DNA fragments separated by gel electrophoresis", J. Mol. Biol., Vol. 98, 1975, pp. 503–517.

[13] A. Jagota, "Microarray data analysis and visualization" (e-book), Bioinformatics, The Bay Press, 2001.

[14] M.B. Eisen, P.T. Spellman, P.O. Brown and D. Botstein, "Cluster analysis and display of genome-wide expression pattern"', Proc. Natl Acad. Sci., USA, Vol. 95, 1998, pp. 14863-14868.

[15] F. Valafar,"Pattern Recognition Techniques in Microarray Data Analysis: A survey", Ann. N.Y. Acad. Sci. Vol. 980, 2002, pp. 41-64.

[16] A. Aldroubi, P. Abry and M. Unser, "Construction of Biorthogonal Wavelets Starting from Any Two Multiresolutions", IEEE Transactions on Signal Processing, vol. 46, no. 4, April,1998, pp. 1130-1133.

[17] Khan J et al, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks", Nature medicine, Vol.7 (6), 2001, pp. 673-679.

[18] Bhattacharyya C et al, "Simultaneous relevant feature identification and classification in high-dimensional spaces: application to molecular profiling data", Signal Processing, Vol. 83, 2003, pp. 729-743.

[19] A.A. Alizadeh et al, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling", Nature, Vol.403 Issue 6769, 2000, pp.503-11.

[20] T.R. Golub et al, "Molecular classification of cancer class discovery and class prediction by gene expression monitoring", Science, Vol. 286, 1999, pp. 531-537.

[21] M.A.Shipp et al, "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning", Nat. Med, Vol. 8, Issue 1, 2002 Jan, pp. 68-74.

[22] R. Kohavi and G. John,"Wrappers for feature selection ", Artificial Intelligence, Vol.97 Issue 12, 1997 December, pp: 273-324.

[23] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning", Artificial Intelligence, Vol.97, Issue 12, December 1997, 245-271.

[24] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, "Gene selection for cancer classification using support vector machines", Machine Learning, Vol.46, 2002, pp.389.422.

[25] E.R. Dougherty, "Small sample issues for microarray-based classification", Comp. Funct.Genomics, vol. 2, no. 1, 2001, pp. 28—34.

[26] F. Azuaje, "A Computational neural approach to support the discovery of gene function and classes of cancer", IEEE transactions on Biomedical Engineering, Vol.48, 2001, pp.332-339.

[27] J. Khan et al, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural network", Nature Medicine Vol. 7, 2001, pp. 673-679.