Comparative Analysis of Genomic Signal Processing for microarray data clustering

Robert S. H. Istepanian – *Senior MIEEE,* Ala Sungoor, Jean-Christophe Nebel – *Senior MIEEE*

***Abstract*—** Genomic Signal Processing is a new area of research that combines advanced digital signal processing methodologies for enhanced genetic data analysis. It has many promising applications in bioinformatics and next generation of healthcare systems, in particular, in the field of microarray data clustering. In this paper we present a comparative performance analysis of enhanced digital spectral analysis methods for robust clustering of gene expression across multiple microarray data samples. Three digital signal processing methods: linear predictive coding, wavelet decomposition and fractal dimension are studied to provide a comparative evaluation of the clustering performance of these methods on several microarray datasets. The results of this study show that the fractal approach provides the best clustering accuracy compared to other digital signal processing and well known statistical methods.

Index Terms—Microarray clustering, Discrete wavelet, Linear predictive coding, Vector quantisation, Fractal dimension, Genomic signal processing

1. Introduction:

In recent years, microarray data analysis has provided better insight on understanding and linkage of genetic disorders in diseases such as diabetes, cardiovascular diseases and some forms of cancer[1]. This process relies mainly on robust clustering, which aims at assigning observations defined in a high dimensional feature space, i.e. gene expression levels, into subsets sharing similar properties[2].

Genomic signal processing (GSP) is a new area of research that applies and develops advanced digital signal processing methodologies for genetic data processing and visualization [3]. In this

work, we are particularly interested in 'GSP clustering', i.e. clustering methods based on Digital Signal Processing (DSP) approaches applied to genomic signals. In recent years several clustering methods based on spectral analysis have been introduced for gene expression profiling [4, 5]. An autoregressive technique was proposed in [4] to evaluate the potential regulatory relationship between genes with dominant spectral components. Other methods presented the decomposition of expression profiles into spectral components to correlate profiles was shown to allow obtaining high accuracy expression values [6]. However, to-date no study has been reported on the comparative evaluation of the clustering performance of different methods designed for Digital Signal Processing against standard microarray clustering algorithms. In this paper we present such a detailed comparative analysis and select the best performance on different standard data sets. In particular we present the performance of Linear Predictive Coding (LPC), Discrete Wavelet Decomposition (DWD) and Fractal Dimension (FD), and compare the clustering performance of these applied on number of microarray datasets with standard clustering methods.

The structure of this paper is as follows. Within the context of microarray data analysis, section (2) reviews, first, conventional clustering methods and, secondly, techniques based on Digital Signal Processing approaches. Section (3) presents the general framework of GSP clustering. In section (4) the details of clustering methods (LPC, DWD or FD) combined with vector quantisation and cluster quality measures are introduced. In section (5), the comparative results are presented. Finally, the paper concludes with ongoing and future work in this area.

2. Related work

2.1- Microarray Clustering and Classification Methods

Although classification and clustering are different machine learning tasks, that depend, respectively, on supervised and unsupervised learning methods, both are relevant to the analysis of microarray data. In recent years many of these methods have been proposed to compare gene expression levels in samples drawn, in general, from two different conditions [7-26]. Table (1) shows a comprehensive summary of existing microarray clustering and classification methods. A brief description of these methods is presented here for completeness. Earlier techniques are based on statistical, deterministic, probabilistic and computational methods producing either distance or similarity measures to achieve dimensionality reduction. Earlier work of Golub et al. [7] used a statistical method, i.e. T-test measure. The method measured correlation that emphasizes the signal-

to-noise ratio by using a gene as a predictor that reflects the difference between the classes relative to the standard deviation within the classes. Large values indicate a strong correlation between the gene expression and the class distinction. The original method used one way clustering only, either for genes or samples, and was sensitive to the number of genes. Since two-way clustering methods are more powerful when dealing with highly dimensional data, Alon developed such a method based on a deterministic annealing algorithm [8], where a square-root barrier function was derived. It approximates a solution of the max-bisection problem allowing separation of a set of genes into two groups which leads to the arrangement of all genes in a binary tree. In order to improve standard annealing which relies on thresholds, another two-way clustering method was proposed using fuzzy C-means and entropy-based clustering [9]. Experiments showed misclassification errors depend on the number of iteration levels. Improved accuracy was achieved using a Biclustering algorithm [10] to identify local structures from gene expression dataset based on Singular Value Decomposition (SVD). The main limitation of all these methods is their dependence on the correct choice of the threshold level parameter that is used in the clustering estimation.

Another line of research investigated the use of Support Vector Machine (SVM) based clustering to microarray data, where the construction of an N-dimensional hyper plane allows the separation of data into two categories. The strength of SVM is it supports both regression and classification tasks and can handle multiple continuous and categorical variables. Furrey et al. [11] proposed an implementation of SVM applied to microarray data clustering where a kernel is initiated, starting with simple dot-product kernel, and then its diagonal factor is tuned using top ranked features to achieve the best performance. Iizuka et al. [12] introduced a Fisher's Linear Classifier to microarray analysis. They showed that this statistical method based on a linear combination that maximizes the ratio of samples between the class variances and the within class variance performs more accurately than SVM based systems. In order to improve these SVM schemes, a heuristic method was introduced for non-parametric clustering where SVM classifiers define support vectors describing portions of clusters and a model selection criterion is used to join these portions [13]. Hybrid models were also proposed to enhance accuracy of SVM based systems. SVM was combined with a Genetic Algorithm (GA) to select predictive genes [14]. An extension of that scheme integrated a specialized Size-Oriented Common Feature Crossover Operator in the GA to keep useful informative blocks and produce offsprings which have the same distribution as their parents [15]. Another hybrid model used metaheuristics consisting of a Particle Swarm Optimization to refine the SVM based approach [15].

Table (1): Summary of existing microarray clustering and classification studies

Techniques	Study	Datasets	Generation procedure	Group
T-test	Golub, 1999[7]	Leukaemia	T-statistics for gene selection Weighting voting for classification	Filtering
Two-way	Alone, 1999[8]	Colon	Correlation for gene selection Deterministic annealing algorithm for clustering	
Two-way	Chandra, 2006[9]	Leukaemia, and Colon	Preprocessing using entropy and correlation measure Clustering based on fuzzy *C*-means	
Biclustering, SVD	Yang, 2009[10]	Human Tissues, Lymphoma, and Leukemia	Preprocessing using statistics for gene selection Clustering based on Singular Value Decomposition	
FLC, SVM	Iizuka, 2003[12]	Hepatocellular	Classification using either Fisher Linear Classifier or Support Vector Machine	Statistical Learning
GA/SVM	Huerta, 2006[14]	Leukaemia, and Colon	Preprocessing using Genetic Algorithm Classification using Support Vector Machine	
PSO/GA-SVM	Jourdan, 2007[15]	Leukemia, Breast, Colon, Ovarian,Prostate	Particle Swarm Optimization (PSO) and a Genetic Algorithm (GA) (both augmented with Support Vector Machines SVM)	
kNN	Singh, 2002[16]	Prostate	K-Nearest Neighbour clustering	
kNN	Nutt, 2003[17]	Gliomas	K-Nearest Neighbour clustering	
PLSLD	Nguyen, 2002[23]	Leukaemia, and Colon	Dimension reduction using Partial Least Square,Classification using Logistic Discrimination and quadratic discriminant analysis	
PLSLD	Fort, 2005[24]	Leukemia, Colon and Prostate.	Combining partial least squares (PLS) and Ridge penalized logistic regression.	
SVM	Furey, 2000[11]	Leukaemia, and Colon	Classification using Support Vector Machine	Statistical Learning embedded with Feature Selection
FJC	Jong, 2003[13]	Leukaemia, and Colon	Preprocessing using support vector classifiers Clustering using Find and Join Clusters method.	
PAM	Tibshirani , 2002[19]	Leukaemia	Class prediction using Prediction Analysis of Microarrays - statistical technique using nearest shrunken centroid	
MARS, LGP, CART,RF	Mukamala , 2005[22]	Leukaemia, Prostate and Colon	Classification using either Linear Genetic Programs or Multivariate Regression Splines or Classification & Regression Tress or random forest	
KPCA	Liu, 2005[20]	Leukaemia, and Colon	Dimension reduction using Kernel Principal Component Analysis Classification with logistic regression (discrimination).	Feature Selection
P-ICR	Huang, 2006[21]	Leukaemia, Colon, Glioma, Hepatocellular	Regularizing gene expression data using Independent Component analysis Classification using Penalized discriminant method	
MRMR	Ding, 2004[25]	Leukaemia, and Colon	Minimum redundancy - maximum relevance (MRMR) feature selection	
MRMR-GA	El Akadi, 2009[26]	Leukaemia, and Colon	MRMR and Genetic algorithm	
FMG-K-mean	DeSouto, 2008[18]	Leukaemia, and Gliomas	finite mixture of Gaussians, followed closely by *k*-means,	Probabilistic

Although SVM has been successfully applied, it requires more training than the statistical and linear discriminant analysis; moreover the classification of data in more than two classes is difficult. In order to address this, distance-based clustering methods such as K-Nearest Neighbour clustering KNN [16, 17] have been used to select set of gene expression profiles. This simple approach assigns each point in data space to its nearest neighbour which forms clusters if distances are sufficiently small. However, this iterative process lacks robustness since it is very sensitive to the chosen number of neighbours. To tackle this weakness and the nonlinearity of the data, Nearest

Shrunken Centroid was successfully proposed for unsupervised gene clustering [19]. This method relies on using denoised versions of the centroids as prototypes for each class. However, due to the unstructured nature of gene data, their algorithm may fall into Local minimums which produce different partitions depending on initialization. Since k-means tend to cluster data within spherical regions of the Euclidean space, better clustering can be achieved using a Finite mixture of Gaussians (FMG-K), which is a curved summation of k multivariate Gaussian density functions or Gaussian components [18]. In a mixture model, each component in the mixture is assumed to model a group of samples. Based on density functions that produce mixing coefficients, one obtains the probabilities of a sample belonging to each cluster. Generally clustering approaches based on distance measures are ineffective to estimate multivariate functions in high dimensionality data. This can be addressed using dimension reduction as a preprocessing step within the cluster analysis pipeline so that not only high-dimensional data become manageable and computational cost is reduced, but also this provides users with possible visual examination of the data of interest. However, dimensionality reduction methods inevitably cause some loss of information which may damage the interpretability of the results. Principle component analysis (PCA) is one of the typical approaches that construct a linear combination of a set of vectors that can best describe the variance of data. Kernel Principle Component Analysis (KPCA) is an extension of PCA performing a nonlinear transformation using integral operator kernel functions [20]. Both processes view the profile vector as a point in this multi-dimensional space and use second-order statistical information of the data. However, since much of the microarray information may be contained in the high-order relationships between samples, these second-order methods are not ideal. Independent Component Analysis (ICA) has potential advantages over PCA [21] to overcome its limitations. ICA uses high-order statistics, not just the covariance matrix as PCA does, which is more suitable for the complexity of gene expression data. Moreover, it can handle a higher level of noise. The main drawback is that ICA ignores some of the spatial and temporal structure contained in the data.

In [22], authors applied t-test to extract different dimensional genes, then applied Multivariate Adaptive Regression Splines (MARS), Classification and Regression Tress (CART), Random Forests (RF) and Linear Genetic Programs (LGP) to classify microarray data. MARS is a nonparametric regression procedure that makes no assumption about the underlying functional relationship between the dependent and independent variables. Instead, MARS constructs this relation from a set of coefficients and basis functions that are entirely “driven” from the data. The method is based on the “divide and conquer” strategy, which partitions the input space into regions, each with its own regression equation. This makes MARS particularly suitable for problems with

higher input dimensions. CART was built to predict continuous dependent variables (regression) and categorical predictor variables (classification). A Random Forest is a classifier consisting of a collection of tree structured classifiers with independent identically distributed random vectors, where each tree casts a unit vote for the most popular class of input. Linear Genetic Programming is a variant of the genetic programming technique that acts on linear genomes. Comparative evaluation shows LGP achieved consistently better results than other methods. However, the underlying problem of this iterative method is that it becomes computationally expensive when dealing with highly dimensional feature vectors.

Another analysis procedure combined dimension reduction using Partial Least Squares (PLS) and classification using either Logistic Discrimination (LD) or Quadratic Discriminant Analysis (QDA) based on the classical multivariate normal model for each class [23]. Experiments show that LDA yields better classification performances than QDA. Although PLS proves more appropriate than PCA for gene feature extraction, it has limitations. First, it is designed to handle continuous responses while the variance of the error in the models differs across gene expression observations. Secondly, this algorithm does not always converge. In order to deal with this, the PLS method was extended to binary response variables to be able to handle the high-dimensional gene expression space [24]. However, this limits its usage to two-class problems. Moreover, experiments show its performance is very sensitive to the choice of iteration and regression parameters.

As a preliminary step in the clustering process, features can be selected using the maximum relevance/minimum redundancy (MRMR) algorithm, which is based on solid multivariate filter procedures [25]. This method addresses data redundancies by selecting genes which have high mutual information (maximum relevance) and simultaneously are mutually maximally independent (minimum redundancy). This process can be further improved with the help of genetic algorithm combined with multi-class SVM. A new fitness function for MRMR-GA with GA-SVM [26] was proposed which always selects the smallest set of genes that provides maximum accuracy.

Most classification and clustering methods require a predefined gene sample similarity or distance metric which has a major impact on their performance depending on how that metric reflects the real relationship among samples. Generally, data-dependent metrics are used; they include Euclidean distance, Manhattan distance and Pearson-correlation. However, in practice, it is desirable to use an adaptive scheme which can estimate the best metric according to input data, i.e. the local features of the gene sample data in this study. In order to address this important challenge, new approaches based on digital signal processing methods have been recently proposed. The next section will introduce such methods and examples.

2.2- Clustering Methods based on Digital Signal Processing methods

To overcome the clustering disadvantage of the standard approaches, several methods based on DSP principles have been proposed for the clustering of genomics data in recent years. In general these methods provide superior characteristics compared to the traditional methods outlined earlier. The gene expression samples can be seen as a signal profile that involves some episodic waveform transitions within time samples. Processing gene expression as time series produces ranges of frequencies that allow finding targets that are expressed periodically with specific correlations between both genes and samples. These characteristics can be analysed further in the frequency domain using different methods designed for processing digital signals to predict and identify samples pattern using techniques such as autocorrelation, trend analysis and autoregressive models. In addition, highly dimensional data, which are a combination of observed and latent variables, could be modelled for marginal inference under multi-biological conditions in a probabilistic system.

The functional nonlinear relations between genes have motivated research in developing nonlinear DSP based techniques for modelling gene expression data samples. The fundamental DSP algorithms which were investigated for analysis of microarray clustering are linear predictive coding, wavelets and fractal dimension. In all cases, the main objective has been to represent a gene expression signal with a set of predictive coefficients which could be processed by spectral clustering using measures such as spectral difference and spectral distortion [2].

Method based on wavelet transform was introduced for identification of microarray features and exploration of their relationship with phenotypic outcomes [27]. This approach allows decomposing a gene signal into components of different length scales, providing a convenient basis for exploring gene behaviour and their clustering according to their expression signal. A hybrid analysis method combining wavelet and GA was proposed to find significant genes [28]. Multilevel wavelet decomposition was performed to reduce the dimensionality of microarray features by breaking gene profiles into approximation and detail coefficients. Approximation coefficients were reconstructed to build the approximation, whereas the genetic algorithm selected the optimal features from approximation coefficients. Experiments, where 15 GAs at 2^{nd} level of wavelet decomposition were used, showed the method achieved more accurate results than statistical methods. A comparative study of multidimensional dataset clustering methods showed that, not only, the Wavelet method is more computationally efficient and accurate than statistical methods, i.e. classical K-means and hierarchical clustering, but it is more sensitive to detect sudden changes in input data [29].

A systematic determination of cluster boundaries using the ratio of within-class and between-class variances was introduced in [30]. Moreover, in order to reduce the noise content in the expression data, discrete wavelet transform with a threshold value was used before the clustering procedure. They tested three different types of mother wavelet functions, i.e. Daubechies, Haar and Symlet, and showed that Daubechies wavelets are the most accurate. Moreover, they discovered that data enhancement by wavelet transforms yielded better results for time series data which have periodicity. The multi-resolution property of wavelet transforms also inspired researchers to consider algorithms that could identify clusters at different scales [31]. This was applied recently to microarray data analysis where feature extraction was based on multilevel wavelet decomposition[32].

Fractals analysis is an effective and relatively recent scientific paradigm that has been used successfully in many areas including biomedical and biological sciences. In particular, it has been recognised as a useful method in quantifying the complexity of dynamical data and signals [33]. The fractal concepts of self-similarity and scaling invariance have been applied to many biological systems, from branching patterns of bronchial and circulatory vessels, to cardiac rhythms, to the geometry of shells and trees [34]. Applications also include genomics where multifractal spectrum analysis was performed on DNA sequences [35].

The determination of fractal dimension (FD) can be used for the characterisation of microarray datasets to measure the similarity of gene expression samples. It can be considered as a relative measure of the number of basic building blocks that form a gene sample pattern. Applications of FD in biomedical and signal processing include two types of approaches: (i) time domain where the original signal is considered as geometric and (ii) phase space domain which estimates the FD in state-space domain [36]. Clustering using FD is a type of grid-based clustering, where the data space is divided in cells by a grid. Some of the well-known techniques that use grid-based clustering are STING [37], WaveCluster [31] and Hierarchical grid clustering [38]. Generally, the effects of these techniques are influenced by the size of the predefined grid and the threshold of the significant cells. Moreover, the technique cannot be scaled to high dimensional datasets due to the computational complexity in number of cells. This was addressed in [39] where they proposed a technique of adaptive grids in subspace whose determination was based on data structure and distribution.

From these studies, it can be noted that although several methods designed for processing digital signals has been used successfully for microarray clustering, no comparative analysis and detailed correlated study of their clustering performance with traditional statistical methods has been carried

out to-date. Furthermore, these studies have also been evaluated using two performance metrics the Davies-Bouldin and Silhouette width methods, widely used in clustering performance analysis studies. In this work we will focus on DSP based extraction methods namely; LPC, DWD and FD for microarray clustering using the same evaluation metrics.

3. GSP Clustering Method principle

The processing blocks of clustering methods based on DSP methods applied to microarray data are shown in Fig. (1). They are summarized in the following steps:

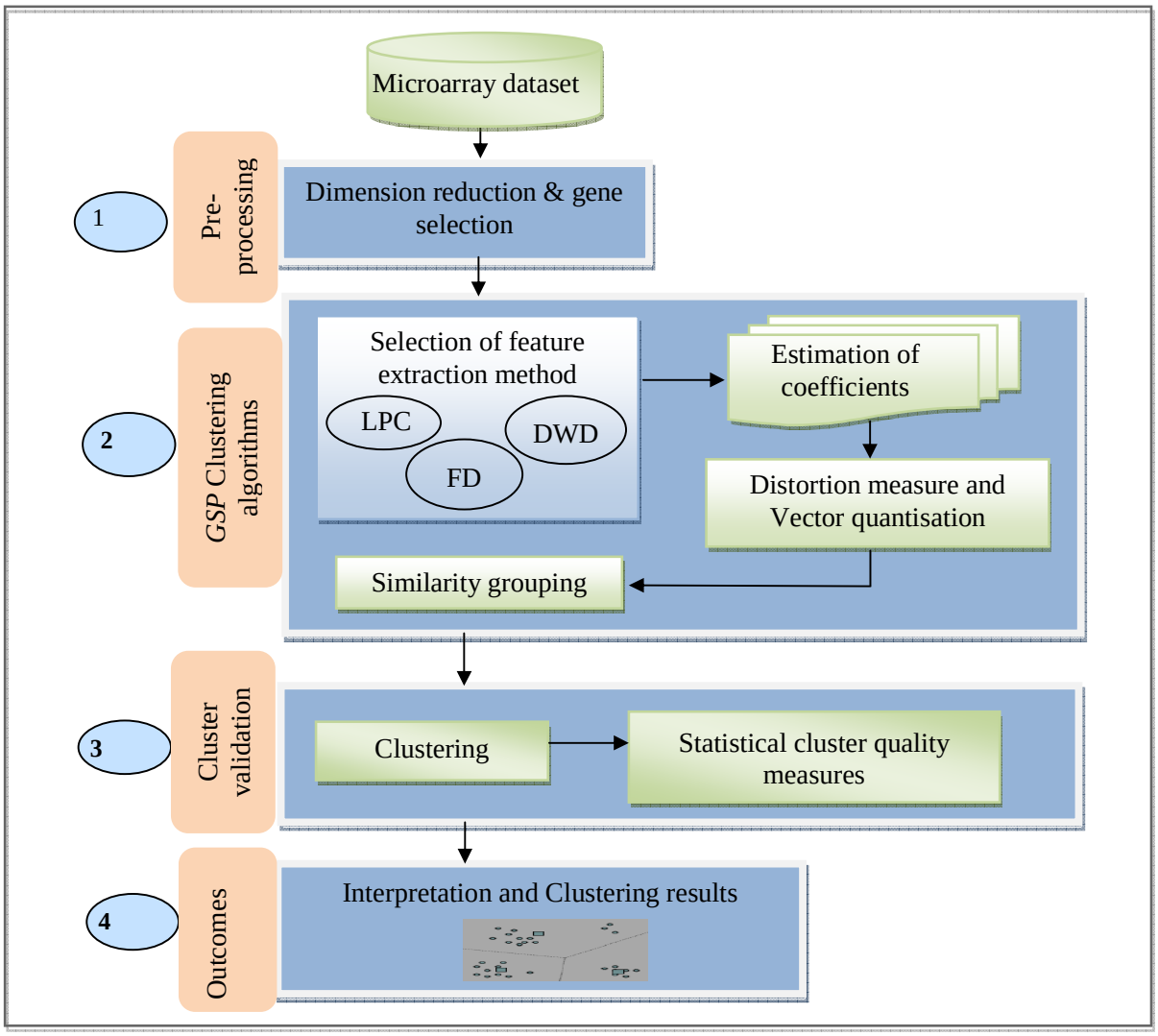


Figure 1 GSP analysis for microarray clustering

1) ***Pre-processing*.** Since gene expression data are high dimensional and contain short multivariate time series, the reduction of the dimensionality of the gene expression variables is required. This can be achieved by either statistically selecting the most expressed genes or specifying a number of genes in the profiles.
2) ***Clustering algorithms*.** This refers to the selection of relevant algorithms to produce informative clusters. For GSP clustering, the approach is divided into two stages:

   i) Selection of feature extraction method: In this, a method based on DSP approach is selected to translate the signal into a representation relevant to the vector of expression profile and to find the best predictive coefficients for the microarray model. This step also determines the proximity measure relative to the similarity-quantified measurement between two vectors of the coefficients measure
   ii) Vector quantisation allows the clustering of the resultant coefficients of the transformed data model into the relevant class partitions. This step determines the distortion measure between vectors of coefficients to quantise into the closest group.

3) **Cluster validation**. Since the clustering process requires no a priori knowledge, its output needs to be evaluated using specific criteria. Statistical comparative approaches are used in most applications to benchmark microarray data clustering methods.
4) **Interpretation and Results.** This final step transforms the cluster validation into a meaningful biological interpretation of the GSP clustering process.

4. GSP Clustering Method

In this section we detail the three DSP based methods mentioned earlier.

4.1- Linear predictive method

In this approach, we tailor the LPC method for microarray data clustering. The gene expression data contain rich information based on a set of a finite number of expression sample values for a set of genes can be represented as (v_d) where d is the dimension of the gene expression array. In considering these characteristics the following gene expression data vector and their relationship can be given by:

$$G_{exp} = \{v_d\}_{(d=1,\ldots,g)} \quad (1)$$

In the LPC method, a predictor is built to estimate the expression variation component as a model coefficient. This is performed by applying a speech analysis method on the microarray expression data. This tailored approach, called *miLPC*, builds on LPC which is a well known and a predominant method for estimating basic speech parameters and which can extract the spectral features of microarray data due to its ability to model multidimensional non linear data. LPC is a method for signal source modelling through observation of input and output sample sequences. The basic concept of LPC analysis is to estimate a functional set of component coefficients which describe the behaviour of a system where each expression sample is approximated as a combination of past samples [40]. A conceptual framework of the miLPC method is illustrated in Fig. (2) where inputs are represented by gene waveforms $v_{(g,n)}$. LPC coding generates a series of coefficient models that involves spectra of the original gene samples signal variation. The computation is based on the principle that the estimated value of a particular gene expression data in microarray at sample *x*, denoted as $\hat{v}_{(g,x)}$, can be predicted approximately by linear combination of the past *p* gene expressions data defined as:

$$\hat{v}_{(g,x)} = \sum_{j=1}^{p} a_j \; v_{(g,x-j)} \quad (2)$$

The prediction variation in expression value $\Delta v_{(g)}$ is the difference between the original data expressions and the predicted as follows:

$$\Delta v_{(g)} = \{(v_{(g)} - \hat{v}_{(g)}), x\} \quad (3)$$

The goal of the LPC analysis is to estimate the best prediction coefficients a_j over *n* gene expression data samples and set the order *p* of the required predictor (usually $n>>p$), so that the predicted expression sample is a good approximation of the original expression sample. This optimization process used to calculate the predictor coefficients is based on minimizing the mean energy in the expression variation over *n* expression samples of the dataset by *least-squares* minimization method. This process leads to a system of *p* equations with *p* unknowns which are solved to find the best fitting predictor coefficients.

There are a number of methods to solve those linear equations. The most common one is the covariance method which is an efficient linear prediction for spectral estimation techniques and is appropriate when estimating coefficients from a sample of a non stationary signal. The covariance

method windows the gene expression variation $\Delta v_{(g)}$ instead of the individual gene expression sample $v_{(g)}$. This prevents the introduction of distortion into the spectral estimation procedure.

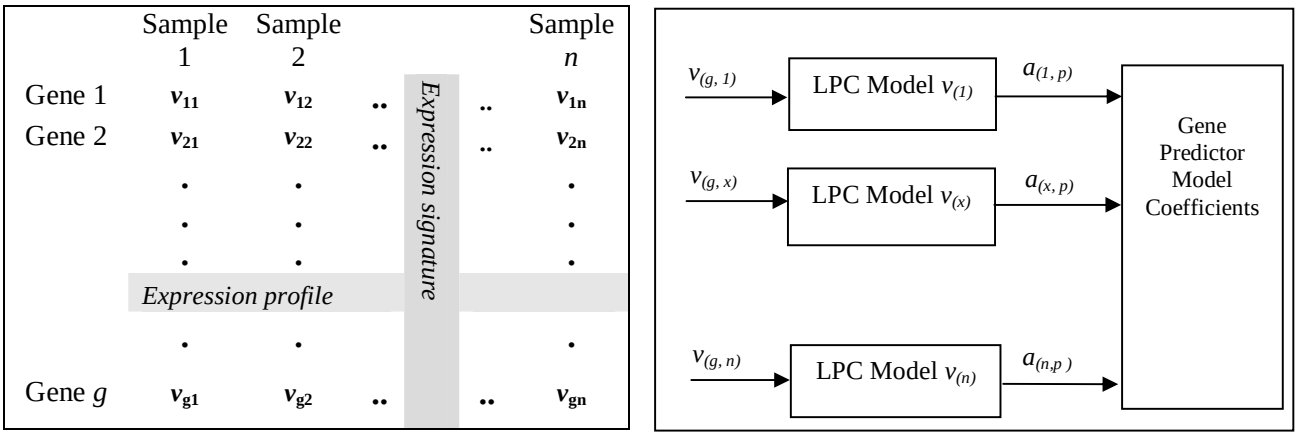


Figure 2 Conceptual framework of the miLPC method

However, direct quantisation of the coefficients a_j is not advisable because of their relatively large dynamic range and possible filter instability problem: small changes due to quantisation error could result in the internal digital filter pole becoming unstable and producing large spectral errors. Thus, other superior parametric representations have been formulated to replace the coefficients a_j [41]. In this work we chose the Line Spectral Frequency (LSF) representation to produce Gene Expression Spectral Frequency (GESF) to capture the spectral expression of information sequence. Since LSF is independent of the characteristics of the source of the sequence, it has been shown to be a particularly efficient for quantisation of information [42]. Moreover, it does not distort the spectrum, varies smoothly across the sequence and offers a better coding in relation to spectral peaks. These GESF coefficients are used subsequently to determine distortion between samples. Fig. (3) describes the processing steps of the miLPC algorithm.

The magnitude of the power spectrum depends on the spacing of the GESF parameters. Closely positioned parameters correspond to the peaks of the spectrum, while widely positioned ones correspond to the spectrum valleys. Since the power spectrum information is more important to the gene expression samples, finer quantisation of the GESF parameters in these regions is desired. This can be achieved by finer quantisation of closely positioned parameters.

Input: Predictor Coeff order *p*, Gene Expression data $v_{g,x}$,
x: Size of Gene expression samples, *g:* Gene Number
Output: Codeword vector
Processing:
1-Compute the gene expression predictive coefficients $\{a_p\}$
2-Compute Mean Square Error, *MSE*
3-Translate coefficients $\{a_p\}$ to Gene Expression Spectral Frequency
4-Drive codebook based on VQ method
5-Explore using the codebook the sample label of the computing clustering of microarray
End

Figure 3 miLPC algorithm

4.2- DWD method

In general, wavelets tend to be irregular, asymmetric and are capable of revealing aspects of data that other analysis techniques disregard. DWD is a method allowing the decomposition of a signal onto a set of basis functions and its analysis by transforming its input time domain into a time-frequency domain. The main advantages of DWD are that it provides resolution optimality in both time and frequency domains, and it does not require a stationary signal [43]. In this work we tailor DWD for microarray gene expression data processing. The method, called *miDWD*, is based on two major sub operations: scaling captures the gene profile information at different frequencies by successive low pass/ high pass filtering and down sampling, whilst translation captures information at different locations. The miDWD method decomposes expression data into several groups of coefficients which contain information regarding the sampled signal at different scales. Coarse scale coefficients represent gross and global features of the signal while fine scale coefficients contain local details. The higher is the number of correlated coefficients between the localized sections of two samples, the more similar the sections are. The wavelet detail coefficients at different levels disclose the fully statistical information contained in the gene expression vector's derivatives.

The goal of the miDWD method is to start from scale-oriented decomposition, and then to analyse the obtained signals on frequency subbands. Using these decomposition coefficients, microarray data clustering can be achieved by measuring similarities between datasets using the vector quantisation method in order to obtain precise discrimination between features of microarray samples and perform robust clustering.

The conceptual framework of the miDWD algorithm is shown in Fig. (4). The method starts by applying recursively two convolution functions, i.e. a low and high pass filters on the given data signal $v_{(g,n)}$. Each function produces an output stream that is half the length of the original input in a specific resolution level. As a result, two sets of coefficients are calculated: the *cA(n)* coefficients are generated by the low pass filter and the *cD(n)* coefficients are produced by the high pass filter. This representation provides information about microarray gene expression sample approximation coefficients and detail coefficients at different scales. Detail and approximation at level *j* are expressed respectively by Eq. (4) and Eq. (5) as follows:

$$D_{j+1}(n) = \sum_t a_j(t)\, g(2n - t) \tag{4}$$

$$A_{j+1}(n) = \sum_t a_j(t)\, h(2n - t) \tag{5}$$

where *h(2n-t)* and *g(2n-t)* are the low-pass filters and high-pass filters. The coefficient vectors are produced by down sampling and are only half the signal length of the coefficient vector at the previous level. The processing steps of the miDWD algorithm are shown in Fig. (5).

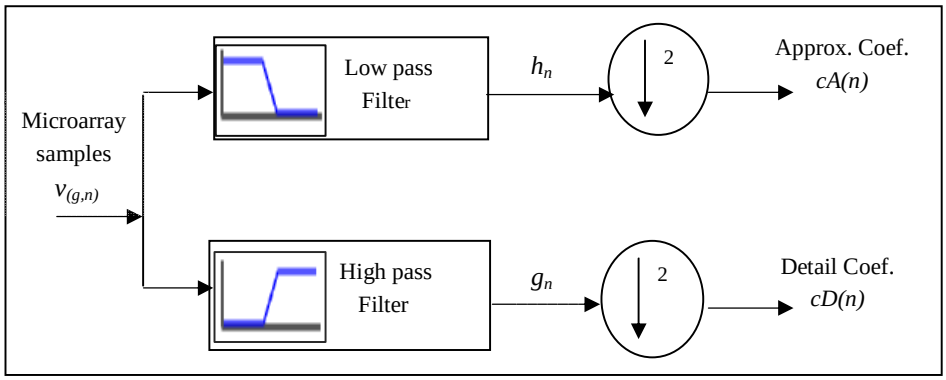


Figure 4 Conceptual framework of the miDWD method

Input: Level of DWD decomposition *lv*, Gene Expression data $v_{g,x}$,
x: Size of Gene expression samples, *g:* Gene Number
Output: Codeword vector
Processing:
1-Compute the gene expression miDWD coefficients { *cA(lv), cD(lv)* }
2-Compute Mean Square Error, *MSE*
3-Drive codebook based on VQ method
4-Explore using the codebook the sample label of the computing clustering of microarray
End

Figure 5 miDWD processing algorithm

4.3- Vector Quantisation for miLPC and miDWD methods

The two methods described earlier miLPC and miDWD require Vector Quantisation (VQ) for their clustering process as explained in section 3. In the clustering application, VQ has two main advantages [44]. First, it allows capturing meaningful classes in the microarray gene expression data samples, represented by the centres of their samples, and second, it makes subsequent classification decisions more robust to the inherent noise of the gene data samples. The principle of VQ is to map *P*-dimensional input vectors $x=[x_1; \ldots ; x_p]^T$ by a finite set of *L* code vectors called *codebook* : $Y = \{y_i; 1 \leq i \leq L\}$. To design a codebook, the *P*-dimensional space is partitioned into *L* cells $\{C_i; 1 \leq i \leq L\}$, then the *quantisation* process assigns one code-vector y_i to each *x* according to which cell, C_i, they belong to: $q(x)=y_i; \ if \ x \in C_i$. The average quantisation error between input source and their reproduction codeword is called the distortion of the vector quantiser. A major aspect of the design of a vector quantiser codebook is to find the best trade-off between distortion and rate. Once the number of quantisation levels is defined, the rate is set. Then the focus is on data quantisation as a means of removing noise from data. The centres of the groups of data corresponding to different quantisation levels should be selected so that distortion is minimized.

In this work, we use a 'nearest neighbour' vector quantiser in the microarray data space, i.e. a vector *z* is represented as a vector of gene expression samples which is mapped to a code vector q_m of expressions in microarray. Implementation of vector quantisation in clustering microarray gene expression samples is achieved as follows:

1- Selecting the expression vector q_m that is nearest to a vector *z*, as defined by

$$c = \arg\min_i (d(z, q_i)) \tag{6}$$

where *d* is a suitable distortion measure. The gain-normalized log spectral distortion is used since it is widely accepted as a good quality measure of signals [45]. It evaluates the similarity of two auto-regressive envelopes of gene expression samples and produces a microarray code book.

The distance *df* between consecutive GESF vectors can be calculated according to the following expression:

$$df(LF_i, LF_k) = \sum_{j=1}^{p} \left[w_j \left(lf_{ij} - lf_{kj} \right) \right]^2 \qquad \text{with} \qquad w_j = P(lf_j) \tag{7}$$

where LF_i and LF_k are vectors of GESFs, lf_{ij} is the j^{th} frequency of LF_i and w_j is the power

spectral distortion measure. Here, the gain-normalized log spectral distortion is used since it is a popular quality measure of coded speech spectra, which evaluates the similarity of two auto-regressive envelopes. It is expressed in the frequency domain by the following equation:

$$d(z, q_i) = \int_{-\pi}^{\pi} \left(log\ P_z(w) - log\ P_{qi}(w) \right)^2 \frac{dw}{2\pi} \quad (8)$$

where $P(w)$ is the auto-regressive envelope that is defined as:

$$P(w) = \frac{1}{\left| 1 + \sum_{k=1}^{p} a_k e^{-jwk} \right|^2} \quad (9)$$

2- Assigning the resultant microarray codebook C_q as cluster label to the data grouped in q.

The design of codebooks is usually accomplished by an iterative algorithm called the Lloyd algorithm. This algorithm generates a set of representative vectors of the source data and optimizes the codebook using the distortion measure method as shown in Fig. (6). Finally, once the codebook has been defined, GESF coefficient vectors are extracted and compared to all codewords of C and mapped to a single codeword that represents the different genes mapped on the tested microarray data.

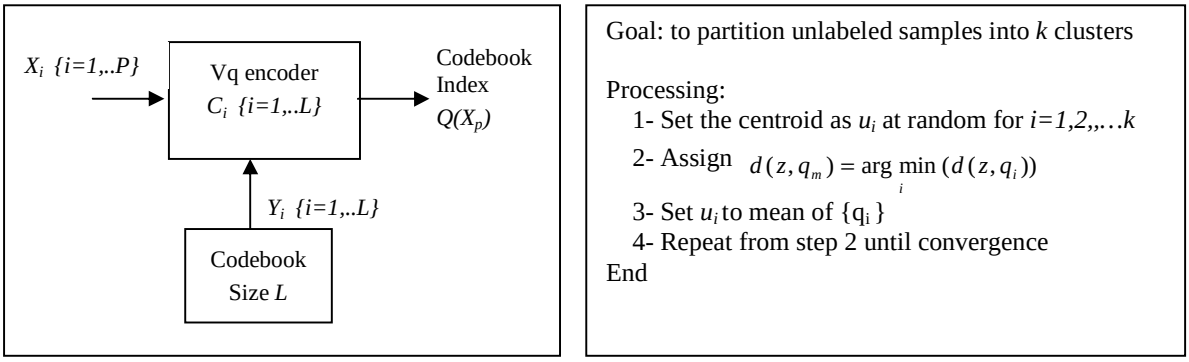


Figure 6 VQ algorithm for the miLPC & miDWD methods

4.4- Fractals Dimension method (FD)

Since a microarray dataset can be represented with columns as attributes (features) and rows as different data objects. Within this framework, the embedding dimension E of a microarray dataset is the dimension of its address space which represents the number of attributes of the dataset, whilst the intrinsic dimension D is the dimension of the spatial object represented by the dataset, regardless of the space where it is embedded. By embedding the dataset in an E-dimensional grid whose cell sides are of size r, the frequency of data points falling into the i^{th} cell can be calculated:

$$FD = \frac{\log(\sum_i C_{r,i}^2)}{\log(r)} \quad (10)$$

where r is the grid size, $C_{r,i}$ is the number of objects in the i^{th} cell under grid size r. Eq. (10) expresses the correlation fractal dimension which measures the probability that two points chosen at random are within a certain distance of each other. Changes in the correlation dimension mean changes in the distribution of points in the dataset. The use of correlation FD as the intrinsic dimension of a dataset allows identifying the correlated attributes and discarding those uncorrelated. We call clustering microarray data in a *D*-dimensional space using fractal dimension method *miFD.*

miFD is based on the box-counting and correlation fractal dimension algorithms [45]. The basic concept can be illustrated as a composition of multi resolution levels describing, for a given object, structures having a self-similarity on varying scales of magnification. The method starts by partitioning the structure of the signal data space dimension into pieces of equal size in a grid of magnification factor size τ. Then, the number of pieces that contain information of the original signal is counted. The process is repeated by iterative partitioning. FD can be calculated by taking the limit of the quotient of the log of the change in object size divided by the log of the change in the measurement scale. Fig. (7) describes the processing steps of the miFD algorithm.

Input: Gene Expression data $v_{g,x}$ and τ

x: Size of Gene expression samples, *g:* Gene Number, *τ:* magnification factor.

Output: FD value

Processing:

1- Select the genes expression sample data and represent them as a signal.

2- Estimate the range of signal space according to starting, ending, minimum and maximum values.

3- Perform regularization of the sample signal into unit squares, τk, which form a mesh grid of dimension τ.

4- Count the number of grid squares, Nτk, which intersect with the signal.

5- Plot log Nτk vs –log and find the slope of the regression line. The slope represents the FD.

End

Figure 7 miFD processing algorithm

5. Comparative performance analysis

In this section we present the details of the comparative analysis of the GSP methods explained earlier. We first describe the microarray datasets and the performance metrics used in this study.

5.1 Microarray datasets

In this paper, we used five well known microarray datasets. These datasets are used universally in microarray data clustering research and considered as benchmark datasets for such studies.

Table (2) shows a summary of each dataset and a brief description of their particular associated diseases. Each dataset has two subsets namely training and test sets. However, since the GSP methods do not depend on any form of training, we combined both sets to produce a unique test sets.

5.2- Performance metrics

The evaluation of spectral clustering depends on two validity indices based on statistical measures, i.e. Davies-Bouldin (DB) and Silhouette Width (SW). These have been widely used in earlier clustering studies [46]. DB is based on the maximization of the distances between clusters while minimizing the distances within a cluster itself. A DB-index is determined as a function of the ratio of the sum of the distances within a cluster to the distance between clusters: the smaller the DB- index, the greater the quality of the achieved clustering.

Table (2): Summary of the tested microarray datasets

Study	Type of disease	No. of genes	Training set			Test set			Total no. of samples	Goal
			Total	Class1	Class2	Total	Class1	Class2		
Golub, 1999 [7]	Leukaemia	7129	38	11 AML	27 ALL	34	14 AML	20 ALL	72	47 ALL 25 AML
Alone, 1999 [8]	Colon cancer	2000	40	14 normal	26 tumor	22	8 normal	14 tumor	62	40 tumor 22 normal
Iizuka, 2003 [12]	Hepatocellular carcinoma	7129	33	12 sick	21 healthy	27	8 sick	19 healthy	60	20 sick 40 healthy
Singh, 2002 [16]	Prostate cancer	12600	102	52 tumor	50 normal	34	25 tumor	9 normal	136	77 tumor 59 normal
Nutt, 2003 [17]	Gliomas	12625	21	14 glio	7 oligo	29	14 glio	15 oligo	50	28 glio 22 oligo

SW exploits inherent features of clusters to assess the validity of results and select the optimal partitioning of the data of interest. This method is based on cluster compactness (in terms of intra-cluster variance) and density between clusters (in terms of inter-cluster density): a good cluster should display an intra density which is much higher than its inter density. To determine SW, firstly the SW of each sample (SW_i) is calculated using Eq. (11). Then the average SW for each cluster is computed. Finally, the overall average SW for all samples is calculated:

$$SW_i = \frac{sc(i) + sd(i)}{\max\{sc(i), sd(i)\}} \tag{11}$$

where *sc(i)* is the average distance between the sample *i* to other samples in the same cluster, *sd(i)* is the average distance between the sample *i* and other samples which belong to the nearest cluster.

The average of Silhouette score for SW_i class *C* across all genes reflects the overall quality of the clustering result as expressed by Eq. 12:

$$\mathrm{ASW(c)} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{SW_i} \tag{12}$$

To measure the global goodness of clustering using the Silhouette index, two parameters are required to be calculated. They are the Silhouette Width range, which is between 1 and -1, and the Average Silhouette Width (ASW). If the value of the Average Silhouette Width is greater than 0.5 it indicates that clusters achieved a reasonable partition of the data. However, if its value is lower than 0.2, it expresses that the data do not exhibit cluster structure.

5.3 Results analysis and Discussion

In order to validate the GSP methods described earlier, a MATLAB® simulation model was implemented. In all GSP methods, we followed the same designed signal processing procedures. First, after pooling together the training and test samples to generate the unique test set for each dataset, we applied the most common gene selection approach called gene ranking [7] to microarray data to select an appropriate number of genes. A univariate analysis approach was used to evaluate each gene individually with respect to a criterion that represents class discrimination ability. Procedures of gene selection are based on computed rank value of each gene according to its signal-to-noise ratio. These selected expression data were then processed to predict microarray coefficients using either the miLPC, miDWD or miFD method. miLPC and miDWD based clustering were performed by vector quantisation method using two codewords which corresponds to the number of

classes in all the tested datasets. For miFD, clustering was carried out by the estimation of fractal dimension for each sample and then by finding cluster based correlation between these dimensions.

We first provide a more detailed analysis of the three GSP methods for Leukaemia. Then, we present general performances obtained on all 5 datasets. Since each set was captured in very different contexts associated to specific medical conditions, they vary in terms of gene and sample sizes. Therefore, for each set, each GSP method had to automatically evaluate the parameters to achieve best performance in the clustering process. Values of these parameters are shown in table 3.

miLPC algorithm iteratively calculates the LPC order *p*, which is the main influencing parameter, by minimising the Mean Square Error (MSE) between the original gene sample signal and the prediction signal. Fig. (8) shows the impact of the number of genes for a variety of orders on the between a signal and its prediction. Experimentally, it was found that a minimum MSE of 0.838 provides accurate clustering analysis of the test Leukaemia dataset for g={75,125} genes using an order *p*={34,35}. Higher order selection would lead to an increased complexity of the analysis without providing better accuracy.

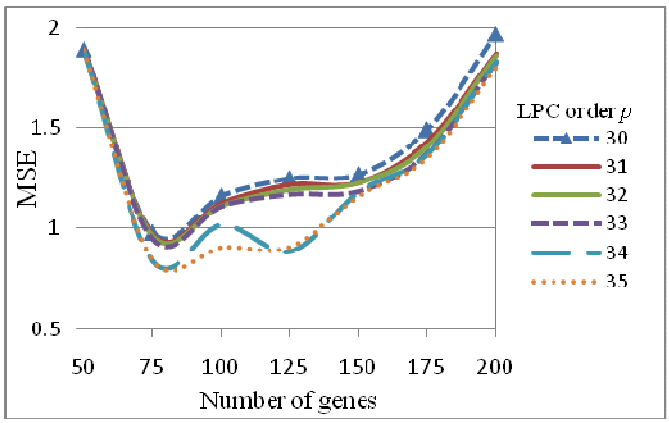
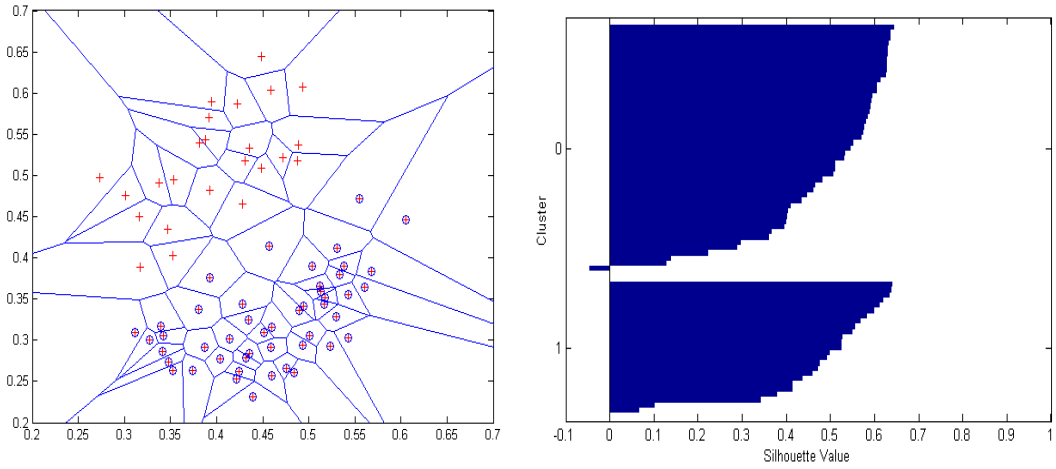


Figure 8 miLPC analysis for different order (*p*) and selected genes of the Leukaemia dataset

Fig. (9a) shows Voronoi clustering of the Leukaemia sample set using the miLPC method with *p*=34 and *g*=75: samples are plotted according to their distortion distances to the two classes. Fig.(9b) presents their associated silhouettes as defined in the previous section. Since the global silhouette index, ASW, is equal to 0.49, the formed clusters are likely to partition accurately the samples in the dataset. On a sample basis, silhouette width values are generally positive which suggests accurate clustering. However, Fig. (9b) shows one exception (sample 21) which displays a negative value and leads to conclude that its grouping is unreliable. Consequently, this sample

should not be associated with any of the clusters. Actually, class labels provided with the dataset reveals that all samples were clustered accurately by miLPC, even the one which was judged as unreliable.



a b

Figure 9 Voronoi clustering associated with silhouettes of Leukaemia dataset obtained by miLPC

Experimental results concerning miDWD method for the Leukaemia dataset are of better quality since a very good predictive MSE of 0.95 can be obtained using g={100,125,150} genes when processed with a DWD filter of level 2. This leads to an ASW of 0.63 and the absence of any unreliable sample. Even better results are obtained using the miFD method. Since fractal property allows localised description of expression data, an excellent ASW of 0.91 is achieved using g=100 genes. The performance of these GSP methods on the five datasets is summarized in Table 3.

Table (3) comparative Performance of the GSP methods

	miLPC approach					miDWD approach					miFD approach			
Datasets	Samples non Clustered	Clustering accuracy	Min no. of genes	LPC order	Predictive error	Samples non Clustered	Clustering accuracy	Min no. of genes	DWD level	Predictive error	Samples non Clustered	Clustering accuracy	Min no. of genes	FD
Leukemia [7]	0	100%	75	34	0.838	0	100%	100	2	0.956	0	100%	100	0.87
Colon [8]	3	95%	100	32	2.97	2	97%	25	3	3.7	1	98%	75	0.55
Hepato-cellular [12]	14	76%	125	29	0.528	7	90%	50	8	4.2	5	92%	100	0.2
Prostate [16]	14	90%	125	28	0.92	8	94%	175	6	3.42	9	93%	100	0.92
Gliomas [17]	5	90%	75	26	1.47	4	92%	50	9	1.3	3	94%	100	0.56

These comparative results indicate that miFD consistently achieves significantly better results than the other GSP methods and miLPC is the least accurate method. Although miFD provide better performance than miDWD, miDWD requires fewer genes to produce accurate clustering.

Fig. (10) presents the validation indices of the GSP methods. The figure shows that DWD and FD have generally average silhouette widths which are either close or greater than 0.5 which indicates they produce reasonable partitions of the data samples. Moreover, FD's values are systematically higher than DWD's. On the other hands, LPC clustering generates low ASWs and even in one instance, where the width is smaller than 0.2, it is not able to produce structured clusters. This figure also provides DB-indices which are in lines with ASWs. This analysis of validation indices confirms the earlier conclusion based on performance: clustering based on the miFD method is consistently the best GSP approach.

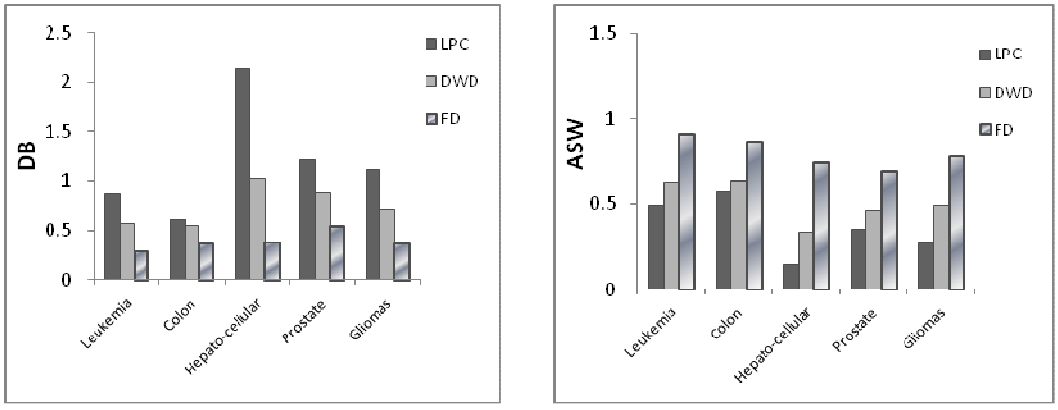


Figure 10 Validation of the GSP methods with DB and ASW

Table 4 shows the complete analysis of the GSP methods compared with earlier clustering methods described in section 2.1. These results demonstrate the superior clustering performance of the miFD method over all other methods. Since only partial results are available regarding the GA/SVM method, the fact it outperforms miFD for the Colon dataset is not fully conclusive. In any case, compared to our approach, GA/SVM has limitations. First, since it is based on GA optimisation, GA/SVM is very computationally expensive. Secondly, unlike the miFD method, it requires a training dataset, which may not be available in some applications.

Table (4): Complete Comparative analysis and comparative of GSP and traditional clustering methods using different microarray dataset.

Method	Author	Leukaemia	Colon	Hepatocellular	Prostate	Gliomas
T-test	Golub,1999[7]	85%				
T-test	Alone,1999[8]		87%			
FLC	Iizuka, 2003[12]			93%		
kNN	Singh,2002 [16]				90%	
kNN	Nutt,2003 [17]					86%
PAM	Tibhirani,2002[19]	95%	83%	59%		67%
MARS	Mukkamala,2005 [22]	85%	80%		92%	
CART	Mukkamala,2005 [22]	92%	95%		96%	
LGP	Mukkamala,2005 [22]	95%	85%		96%	
RF	Mukkamala,2005 [22]	100%	90%		88%	
PLSLD	Nguyen,2002[23]	97%	92%			
KPCA	Liu,2005[20]	97%	100%			
FJC	Jong,2003[13]	91%	54%			
Two-way	Chandra,2006[9]	96%	88%			
SVM	Furey,2000[11]	94%	90%			
MRMR	Ding,2004[25]	100%	94%			
GA/SVM	Huerta,2006[14]	100%	99%			
P-ICR	Huang, 2006[21]	95%	86%	62%		74%
miLPC		**100%**	**95%**	**76%**	**90%**	**90%**
miDWD		**100%**	**97%**	**90%**	**94%**	**92%**
miFD		**100%**	**98%**	**92%**	**93%**	**94%**

6. Conclusion

In this paper, we introduced a detailed comparative analysis of GSP methods for microarray clustering. The performance analysis of these methods on different well known test bench microarray datasets indicates that the miFD method outperform all the other GSP and traditional methods without the need for either training datasets or vector quantisation analysis. The quality of the results obtained from our miFD approach suggests that this method is able to partition the samples of the signal data space by extracting the relevant features. This can be explained by the fact that the miFD cluster method depends on intrinsic relationship in the sample cluster set, rather than geometric shape or distances. Furthermore, provides enhanced characterization property indicated by the interaction between the smallest partitions with the distribution of the samples to a degree that cannot be matched by traditional statistical measurements.

Also, our study indicates that the proposed methods can be applied in future GSP microarray

clustering studies for different diagnostic and personalised healthcare systems. Ongoing work is currently underway to integrate adaptive schemes into the presented methods to provide better processing capabilities for testing larger datasets, different diseases and genetic samples without the need of the relevant parametric selection procedures. Further work on non-stationary data samples using adaptive DSP methods is currently on going. The GSP methods present a suitable approach for real-time processing of different gene expression data sets that might be required in future studies in areas such as mobile healthcare or individualised medicine. Further work will focus on the application of the presented methods to gene selection instead of sample selection.

References

[1] R. Istepanian, "Microarray Image Processing; current status and future directions", IEEE Trans on Nanobioscience, Vol.2, No.2, 2003, pp. 173-175.

[2] T. Pham, C. Wells and D. Crane, "Analysis of microarray gene expression data", Current Bioinformatics, Vol. 1, No. 1, 2006, pp. 37-53.

[3] E. Dougherty, I. Shmulevich, J. Chen, and Z. Wang, "Genomic Signal Processing and Statistics", Hindawi Pub. Corp., 2005.

[4] L. Yeung, L. Szeto, A-C. Liew and H. Yan, "Dominant spectral component analysis for transcriptional regulations using microarray time-series data", Bioinformatics, Vol. 20, No. 5, 2004, pp. 742-749.

[5] W. Zhang and I. Shmulevich, "Computational and Statistical Approaches to Genomics", Kluwer Academic Publishers, Boston, 2002.

[6] H. Yan and T. Pham, "Spectral similarity for analysis of DNA microarray time-series data", Int. J. Data Mining and Bioinformatics, Vol. 1, No. 2, 2006, pp. 150-161.

[7] T. Golub, *et al*., "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring", Science, Vol. 286, No. 15, 1999, pp. 531–537.

[8] U. Alon, *et al*., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", PNAS., Vol. 96, No.12, 1999, pp. 6745-6750.

[9] B. Chandra, S. shanker and S. Mishra, "A new approach: Interrelated two way clustering of gene expression data", Jour of statistical methodology, Vol. 3, 2006, pp. 93-102.

[10] W. Yang, D. Dai, and H. Yan, "Biclustering of Microarray Data Based on Singular Value Decomposition", Emerging Technologies in Knowledge Discovery and Data Mining, Vol. 4819, 2009, pp. 194-205

[11] T. Furey, N. Cristianini, N. Du.y, D. Bednarski, M. Schummer and D. Haussler., "Support vector machine classification and validation of cancer tissue samples using microarray expression data". Bioinformatics, Vol. 16, No. 10, 2000, pp. 906–914.

[12] N. Iizuka, *et al*., "Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection", The Lancet, Vol. 361, No. 9361, 2003, pp. 923-929.

[13] K. Jong, E. Marchiori, A. Van Der, "Finding clusters using support vector classifiers", ESANN-1th European Symposium on Artificial Neural Networks Bruges, Belgium, April 23-25, 2003.

[14] E. Huerta, B. Duval and J-K Hao, "A Hybrid GA/SVM Approach for Gene Selection and Classification of Microarray Data", Evo Workshops, LNCS, Vol. 3907, 2006,.pp. 34-44

[15] L. Jourdan, E. Alba, J. Nieto and T. El-Ghazali, "Gene Selection in Cancer Classification using PSO/SVM and GA/SVM Hybrid Algorithms"**,** IEEE Congress on Evolutionary Computation CEC-05, Singapore, Sep, 2007.

[16] D. Singh, *et al* , "Gene expression correlates of clinical prostate cancer behavior". Cancer Cell, Vol.**1**, No.2, 2002, pp. 203-209.

[17] C. Nutt, et al , “Gene expression-based classification of malignant gliomas correlates better with survival than histological classification”, Cancer Research, Vol. 63, April 2003, pp. 1602-1607.
[18] M. De Souto, *et al.*, “Clustering cancer gene expression data: a comparative study”, BMC Bioinformatics Vol. 9, No. 497, 2008,
[19] R. Tibshirani, T. Hastie, B. Narasimhan and G. Chu, “Diagnosis of multiple cancer types by shrunken centroids of gene expression”, PNAS, Vol. 99, No.10, 2002, pp. 6567–6572.
[20] Z. Liu, D. Chen and H. Bensmail, “Gene Expression Data Classification with Kernel Principal Component Analysis”, J. Biomed. Biotechnol., Vol. 2, 2005, pp. 155–159.
[21] D. Huang, and C. Zheng, “Independent component analysis-based penalized discriminant method from tumor classification using gene expression data”, Bioinformatics, Vol.22, No.15, 2006, pp. 1855-1862.
[22] S. Mukkamala, Q. Liu, R. Veeraghattam and A. Sung, “Computational Intelligent Techniques for tumor classification using microarray gene expression data”, Int. J. of Lateral Computing, Vol. 2, No. 1, 2005, pp. 38-45.
[23] D. Nguyen and D. Rocke, “Tumor classification by partial least squares using microarray gene expression data”, Bioinformatics, Vol. 18, No. 1, 2002, pp. 39-50.
[24] G. Fort and S. Lacroix, “Classification using partial least squares with penalized logistic regression”, Bioinformatics, Vol. 21, No. 7, 2005, pp. 1104-1111;
[25] C Ding, and H Peng, “Minimum redundancy feature selection from microarray gene expression data”, IEEE Computer Society Bioinformatics conf.- CSB, Aug 2003, Stanford, CA, USA., pp. 523-529
[26] A. El Akadi, et al , “A New gene selection approach based on Minimum Redundancy-Maximum Relevance (MRMR) and Genetic Algorithm (GA)”, AICCSA2009, IEEE/ACS International Conference on Computer Systems and Applications, Vol. 10-13, May 2009, pp.69 – 75
[27] Lio P., (2003), “Wavelets in bioinformatics and computational biology: state of art and perspectives”, Bioinformatics, Vol. 19, pp. 2–9.
[28] Liu Y., (2008), “Detect Key Gene Information in Classification of Microarray Data”, EURASIP Journal on Advances in Signal Processing, Vol. 2008.
[29] Otazu X. and Pujol O., (2006), “Wavelet based approach to cluster analysis. Application on low dimensional datasets”, Pattern Recognition Letters, Vol. 27, pp. 1590–1605
[30] Moesa H. A. et al., (2005), “Efficient Determination of Cluster Boundaries for Analysis of Gene Expression Profile Data Using Hierarchical Clustering and Wavelet Transform”, Genome Informatics, Vol. 16, No.1, pp. 132-141.
[31] Sheikholeslami G. et al., (1998), “WaveCluster: A multiresolution clustering approach for very large spatial dataset”, Proc. of the 24th Int. Conf. on Very Large Data Bases, Morgan Kaufmann, New York, USA, pp. 428–439.
[32] Liu Y., (2009), “Wavelet feature extraction for high-dimensional microarray data”, Neurocomputing, Vol.72, No. 4, pp. 985– 990.
[33] Jelinek F. et al., (1998), “Is there meaning in fractal analyses”, Complex systems, Conference 98 UNSW Sydney, pp.144-149.
[34] Seymour G., (2004), “Fractal properties of the human genome”, Journal of Theoretical Biology, Vol. 230, pp. 251–260.
[35] Zhi-Yuan S. et al., (2007), “Local scaling and multifractal spectrum analyses of DNA sequences – GenBank data analysis”, Chaos, Solitons and Fractals, Vol.40, No. 4, pp. 1750-65.
[36] Carlin M., (2000), “Measuring the complexity of non-fractal shapes by a fractal method”, Pattern Recognition Letters, Vol. 21, No. 11, pp. 1013–1017.
[37] Wang W. *et al.*, (1997), “STING: A statistical information grid approach to spatial data mining”, Proc. of the 23rd Int. Conf. on Very Large Data Bases, Morgan Kaufmann, Athens, Greece, pp. 186–195.
[38] Peter G. and Borisov A., (2002), "Using Grid-Clustering Methods in Data Classification”, Int Conf. on Parallel Computing in Electrical Engineering (PARELEC'02), pp. 425.
[39] H. Kriegel, P. Kröger, A. Zimek, (2009), “Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering”, ACM Transactions on Knowledge Discovery from Data (TKDD), Vol. 3, No. 1.
[40] T. Quatieri, (2002), “Discrete time speech signal processing”, Prentice Hall.

[41] R. Istepanian, A. Sungoor, Jc. Nebel, (2007),“ Linear predictive coding for enhanced microarray data clustering”, Fifth IEEE Int. Workshop on Genomic signal processing and statistics, GENSIPS’07, Tuusula, Finland,
[42] K. K. Paliwal and W. Kleijn, (1995), “Quantization of LPC parameters”, in speech coding and synthesis, Elsevier Science, pp. 433-466.
[43] J. Chen, H. Li, K. Sun, B. Kim, (2003),”How will Bioinformatics impact signal Processing research,” IEEE Signal Processing Magazine, Vol. 20, No. 6, pp. 16-26.
[44] J. Li and H. Zha, (2002), “Simultaneous classification and feature clustering using discriminate vector quantization with application to microarray data analysis”, IEEE Proc. computer society Bioinformatics Conf., Stanford, USA.
[45] F. Camastra., (2003), “Data dimensionality estimation methods: a survey”, Pattern Recognition, Vol. 36, No. 12, pp. 2945–2954.
[46] N. Bolshakova, and F. Azuaje, (2003), “Cluster validation techniques for genome expression data”, Signal Processing, Vol. 83, pp. 825–833.