International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) e-ISSN: 2395-5325
Volume 5, Issue 2 (April '2019)

# DNA MICROARRAY DATA REDUCTION METHOD FOR DIMENSIONALITY PROBLEMS

**R.Nandhakumar,**
Assistant Professor,
Department of Computer Science,
Nallamuthu Gounder Mahalingam College,
Pollachi, Tamilnadu, India-642001.

**Dr. Antony Selvadoss Thanamani,**
Associate Professor& Head,
Department of Computer Science,
Nallamuthu Gounder Mahalingam College,
Pollachi, Tamilnadu, India-642001.

***Abstract:*** High Dimensional Non-Linear data reduction has become inevitable for pre-processing of high dimensional data. "Gene expression microarray data" is an instance of such high dimensional data. Gene expression microarray data displays the maximum number of genes (features) simultaneously at a molecular level with a very small number of samples. The copious numbers of genes are usually provided to a learning algorithm for producing a complete characterization of the classification task. However, most of the times the majority of the genes are irrelevant or redundant to the learning task. It will deteriorate the learning accuracy and training speed as well as lead to the problem of over fitting. Thus, High Dimensional Non-Linear data reduction of microarray data is a crucial preprocessing step for prediction and classification of disease. Various feature selection and feature extraction techniques have been proposed in the literature to identify the genes that have direct impact on the various machine learning algorithms for classification and eliminate the remaining ones. This paper describes the taxonomy of High Dimensional Non-Linear data reduction methods with their characteristics, evaluation criteria, advantages and disadvantages. It also presents a review of numerous dimension reduction approaches for microarray data, mainly those methods that have been proposed over the past few years.

***Keywords: DNA microarrays, High Dimensional Non-Linear data reduction, classification, prediction.***

## I. INTRODUCTION

The theory of microarrays methodology was first introduced and demonstrated by Chang TW in 1983, for antibody microarrays in a scientific publication and registered a series of patents [1]. In 1990s, Microarrays were developed as a consequence of the efforts to speed up the process of drug discovery [2]. Traditional drug discovery was shaped for developing a number of candidate drugs and trying them one by one against diseases of interest. The long and limited method of trial and error based centered drug discovery could not be very effective for some particular diseases. A group of researchers established a photolithography system to achieve this in a fashion similar to the synthesis of VLSI (Very Large Scale Integration) chips in the semiconductor industry [3]. The first version developed by the sister company Affymetrix came to be known as the Gene chip [4]. The "gene chip" industry started to grow significantly after the 1995 with the publication of Science Paper by the Ron Davis and Pat Brown labs at Stanford University. Simultaneously researchers at the Pat Brown's lab of Stanford University developed a different type of microarray [7].

In the past few years, multivariate statistics for microarray data analysis has been the subject of thousands of research publications in Statistics, Bioinformatics, Machine learning, and Computational biology. Most of the traditional issues of multivariate statistics have been studied in the context of high-dimensional microarray data. The main types of data analysis needed for biomedical applications include [9,10]:

- *Gene Selection*: the procedure of feature selection, that finds the genes, strongly associated with a particular class.
- *Classification*: classifying diseases or predicting outcomes based on gene expression patterns, and perhaps even identifying the best treatment for given genetic signature [10].
- *Clustering*: finding new biological classes or refining existing ones [11].

Clustering can be used to find groups of similarly expressed genes in the aspiration of finding that both have a similar function [12]. On the other hand, another topic of interest is the classification of the microarray data for prediction of disease such as cancer using gene expression levels. Classification of gene expression data samples involves dimension reduction and classifier design. Thus, in order to analyze gene expression profiles correctly, dimension reduction is an important process for the classification.

The goal of microarray data classification of cancer is to build an efficient and effective model that can differentiate the gene expressions of samples, i.e. classify tissue samples into different classes of the tumor. Nearest neighbor classification, Artificial Neural Network,

Bayesian, Decision tree, Random forest methods and Support Vector Machine (SVM), are the most well-known approaches for classification. An overview of the methods mentioned above can be found in Lee et.al. [10].

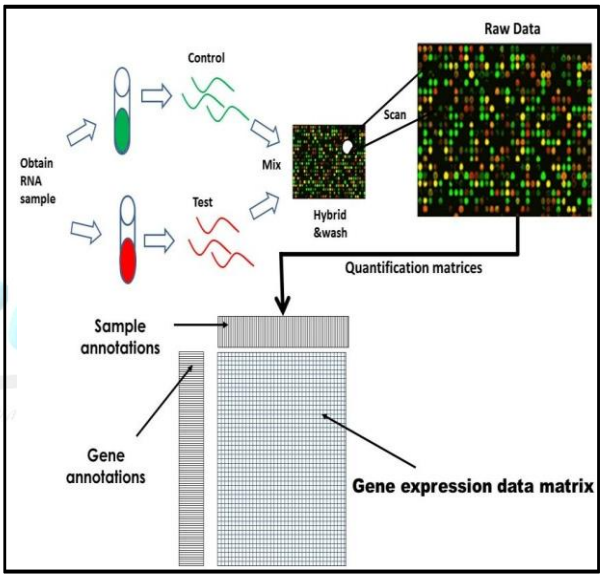
Recently, many gene expression data classification and High Dimensional Non-Linear data reduction techniques have been introduced. You W et al. applied feature selection and feature extraction for High Dimensional Non-Linear data reduction of microarray by using Partial Least Squares (PLS) based information [16]. Xi M et al. used a binary quantum-behaved Particle Swarm Optimization and Support Vector Machine for feature selection and classification. Shen et al. proposed a modified particle swarm optimization that allows for the simultaneous selection of genes and samples [14]. Xie et al. developed a diagnosis model based on IFSFFS (Improved F-score and Sequential Forward Floating Search) with support vector machines (SVM) for diagnosis of erythematic to squamous diseases [12]. Li et al. proposed an algorithm with a locally linear discriminate embedded in it, to map the microarray data to a High Dimensional Non-Linear data space, while Huang et al. recommended an upgraded decision forest method for the classification of microarray data that used a built-in feature selection method for fine-tuning.
In subsequent years, the use of gene expression profiles for cancer diagnoses has been the major focus in many microarray studies. Various gene selection methods and classification algorithms are proposed in the literature which are able to reduce the dimensionality by removing irrelevant, redundant and noisy genes for accurate classification of cancer.

## II.MICROARRAY GENE EXPRESSION DATA ANALYSIS CHALLENGES

Microarray technology provided biologists with the ability to measure the expression levels of thousands of genes in a single experiment. The data from microarray consists of a small sample size and high dimensional data. A characteristic of gene expression microarray data is that the number of variables (genes) *m* far exceeds the number of samples *n*, commonly known as "Curse of dimensionality" problem. Processing of microarray gene expression data is shown in Figure 1. To avoid the problem of the "curse of dimensionality", High Dimensional Non-Linear data reduction plays a crucial role in DNA microarray analysis. Microarray experiments provide huge amount of data to the scientific community, without appropriate methodologies and tools, significant information and knowledge hidden in these data may not be discovered.

The vast amount of raw gene expression data leads to statistical and analytical challenges. The challenge experienced by statisticians is the nature of the microarray data. The best statistical model will largely depend on the total number of possible gene combinations. Therefore, the impact of microarray technology on biology will depend heavily on data mining and statistical analysis. Conventional statistical methods give improper result due to high dimension of microarray data with limited number of patterns. Therefore, there is a need for methods capable of handling and exploring large data sets. The field of data mining and machine learning provides a wealth of methodologies and tools for analyzing large data sets. A sophisticated data-mining and analytical tool is required to correlate all of the data obtained from the arrays and which can help to group them in a meaningful way.

Data Mining with machine learning is a process to discover meaningful, non-trivial, and potentially useful knowledge (patterns, relationships, trends, rules, anomalies, dependencies) from the large amount of data by using different types of automatic or semi-automatic techniques. Compared with classical statistical approaches, data mining is feasibly best seen as a process that incorporates a wider range of integrated methodologies and tools, including databases, machine




learning, knowledge-based techniques, network technology, modeling, algorithms, and uncertainty handling.

**Figure :1 Formation of Microarray gene expression data.**

Gene expression data of DNA microarray which represent the state of a cell at a molecular level, have a great prospective as a medical diagnosis tool. Typical microarray data mining analysis include discriminate analysis, regression, clustering, and association and deviation detection. Several machine learning techniques such as, Support Vector Machines (SVM) , k-Nearest Neighbors (kNN) , Artificial Neural Networks (ANN),Naïve Bayes (NB), Genetic Algorithms, Bayesian Network, Decision Trees, Rough Sets, Emerging Patterns, Self-Organizing Maps, have been used by different research for different analysis of microarray gene expression data.

In classification, available training data sets are generally

of a fairly small sample size compared to a large number of genes involved. Theoretically, increasing the size of the genes is expected to provide more discriminating power but in practice, large genes significantly slow down the learning process. As well as cause the classifier to over fit the training data and compromise model simplification. Dimension reduction can be used to successfully extract those genes that directly influence the classification. In this paper, we focus our discussion on popular machine learning techniques for High Dimensional Non-Linear data reduction and identification of potentially relevant genes for molecular classification of cancer.

## III.DIFFERENT HIGH DIMENSIONAL NON-LINEAR DATA REDUCTION TECHNIQUES

For microarray data classification, the main difficulty with most of the machine learning technique is to get trained with a large number of genes. A lot of candidate features (genes) are usually provided to a learning algorithm, for constructing a complete characterization of the classification task. In the past ten years, due to the applications of machine learning or pattern recognition, the domain of features have expanded from tens to hundreds of variables or features used in those applications. Several machine learning techniques are developed to address the problem of reducing irrelevant and redundant features which are a burden for different challenging tasks. The next section is about feature selection methods (filters, wrappers, and embedded techniques) applied on microarray cancer data. Then we will discuss feature extraction methods, special case of feature selection method for microarray cancer data and the final section is about combination of different feature selection method as a hybrid search approach to improve classification accuracy and algorithmic complexity.

- **Feature Selection**

In machine learning, feature selection also known as variable selection, attribute selection or variable subset selection. Feature selection is the process of selecting a subset of relevant and redundant features from a dataset in order to improve the performance of the classification algorithms in terms of accuracy and time to build the model. The process of feature selection is classified into three categories.

- **Filter**

Filter methods use variable ranking methods as the standard criteria for variable selection by ordering. Statistical ranking methods are used for their simplicity and good success is reported for practical applications. A different suitable ranking criterion of statistics is used to score the variables and select a threshold value in order to remove the variables below it. One definition that can be mentioned, which will be useful for a feature is that "A feature can be regarded as irrelevant if it is conditionally independent of the class labels". If a feature is to be relevant it can be independent of the input data but cannot be independent of the class labels i.e. the feature that has no influence on the class labels can be discarded. The filter methods grouped in ranking and space search methods according to the strategy utilized to select features. Filter ranking methods select features regardless of the classification model that are based on univariate and multivariate feature ranking techniques. The process of feature selection that follows the filter methods is depicted in Figure 2. This Figure shows that it selects features, which are similar to ones already picked. This provides a good balance between independence and discrimination. Since the data distribution is unknown, various statistical techniques can be used to evaluate different subsets of features with a chosen classifier. Some of the popular technique found in literature that can be used for feature ranking, with their advantage and disadvantage are listed in Table 1.

Different researchers used a different framework of filter methods in their works for the gene selection of microarray data. Lin and Chien, used statistical clustering, based on linear relationship and Coefficient correlation for Breast cancer cDNA micro-array data. Zhu et al. used model-based entropy for feature selection. Some of the researchers used signal-to-noise ratio approach in a leukemia dataset with k-fold and Holdout validation method .Wei et al. developed two recursive features elimination methods , i.e. Feature score based recursive feature elimination (FS-RFE) and Subset level score based re-cursive feature elimination (SL-RFE). Recently, a multiphase cooperative game theoretic feature selection approach has been proposed for microarray data classification by Mortazavi et al. in 2016. The average classification accuracy on eleven microarray data sets in this work shows that the proposed method improves both average accuracy and average stability.

The benefits of variable ranking are computationally easy and avoid over fitting and are proven to work well for certain datasets. Filter methods do not rely on learning algorithms which are biased and is equivalent to changing data to fit the learning algorithm. One of the drawbacks of ranking methods is that the selected subset might not be optimal because in that a redundant subset might be obtained. Finding a suitable learning algorithm can also become hard since there is no underlying learning algorithm for feature selection. Also, there is no ideal method for choosing the dimension of the feature space.

| Model search | Advantages | Disadvantages | Examples |
|---|---|---|---|
| | **Univariate** | | |
| | Fast, Scalable | Ignores feature dependencies | $\chi^2$ |
| | Independent of the classifier | Some features which as a group have | Euclidean distance |
| | | strong discriminatory power but are weak | t-test |
| | | as individual features will be ignored | Information gain |
| Filter | | Features are considered independently | Gain ratio |
| | **Multivariate** | | |
| | The models feature dependencies | Slower than univariate techniques | Correlation-based feature selection (CFS) |
| | Independent of the classifier | Less scalable than univariate techniques | Markov blanket filter (MBF) |
| | Better computational complexity | Ignores interaction with the classifier | Fast correlation-based |
| | than wrapper methods | redundant features may be included | feature selection (FCBF) |

**Table 1. Advantages and disadvantages of filters methods.**

- **Sequential Selection Algorithms**

The Sequential selection algorithm finds the minimum (or maximum) features by iterating the process. The Sequential

Feature Selection (SFS) algorithm starts with an empty set and adds one feature for the first step that increases the performance of the objective function. From the second step onwards the remaining features are added individually to the current subset and the performance of new subset is calculated. By this process the best feature subset can be found that gives the maximum accuracy of the classifier [15]. The process is repeated until the required numbers of features are added. The Sequential Floating Forward Selection (SFFS) algorithm is more flexible than the naive Sequential Floating Selection (SFS) because it introduces an additional backtracking step.

- **Evolutionary Selection Algorithms**

An evolutionary selection algorithm is a method that might not always find the best solution but definitely finds a good solution in reasonable time by sacrificing totality to increase efficiency. The objective of an evolutionary is to produce a solution in a reasonable time frame that is good enough for solving the problem at hand. The Evolutionary search algorithms evaluate different subsets to optimize the performance of the objective function. Different feature subsets are produced either by searching around in a search space or by generating solutions to the optimization problem. Evolutionary algorithms are based on the ideas of biological evolution, such as reproduction, mutation, and recombination, for searching the solution of an optimization problem. The main loop of evolutionary algorithms includes the following steps:

1. Initialize and estimate the initial population.
2. Implement competitive selection.
3. Apply different evolutionary operators to generate new solutions.
4. Estimate solutions in the population.
5. Repeat from the second steps, until some convergence criteria is fulfilled [13].

Some examples of Evolutionary algorithms are simulated annealing algorithm, Tabu search, Swarm intelligence. The most successful among evolutionary algorithms are Genetic Algorithms (GAs). They have been investigated by John Holland, and demonstrate essential effectiveness and filter feature selection procedure are depicted in Figure 2.

The performance of the filter versus wrapper gene selection technique was evaluated by Srivastava et al. in 2014 by supervised classifiers over three well known public domain datasets with Ovarian Cancer, Lymphomas & Leukemia. In the study, Relief F method was used as a filter based gene selection, and Random gene subset selection algorithm was used as a wrapper based gene selection. Recently, Kar et al. developed a computationally efficient but accurate gene identification technique "A particle swarm optimization based gene identification technique". In this technique at the onset, the t-test method has been utilized to reduce the High Dimensional Non-Linear data of the dataset and then the proposed particle swarm optimization based approach has been employed to find useful genes.

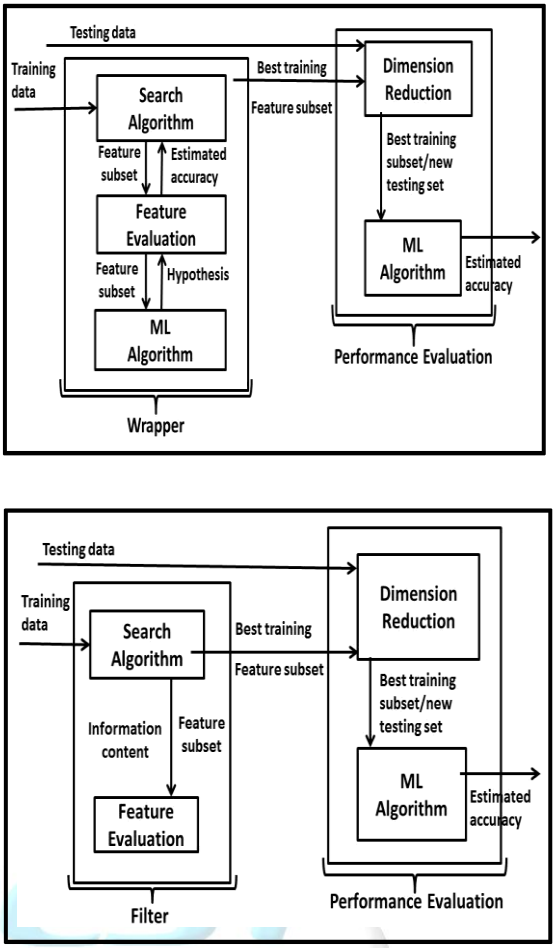



**Figure 2. Feature selection procedure of filter and wrapper approaches.**

Wrappers tend to perform better, in selecting features because they take the model hypothesis into account by training and testing in the feature space. The main disadvantage of Wrapper methods was the number of iterations required to obtain the best feature subset. For every subset evaluation, the predictor creates a new model, i.e. the predictor was trained for every subset and tested to obtain the classifier accuracy. If the number of samples were large, most of the algorithm execution time was spent in training the predictor. Another drawback of using the classifier performance as the objective function was that the classifiers were prone to over fitting. Overfitting occurs if the classifier model, well learned the data and provides poor generalization capability. The classifier can introduce bias and increases the classification error. Using classification accuracy in feature subset selection, can result in a bad feature subset with high accuracy, but poor generalization power .

- **Linear Feature Extraction**

Linear feature extraction assumes that the data are linearly separable in a lowest dimensional subspace. It transforms them on this subspace by using matrix factorization method. Given a dataset X:N.D, there exists a projection matrix U:D.K and a projection Z:N.K, where Z = X × U. Using

UUT = I (orthogonal property of eigenvectors), we get X = Z × UT . The most famous linear feature extraction method is principal component analysis (PCA). PCA uses the covariance matrix and its reign values and eigenvectors, to find the "principal components" in the data that are uncorrelated eigenvectors, each demonstrating some proportion of variance in the data. PCA and its several versions have been applied to reduce the dimensionality of the cancer microarray data. These methods were highly effective in identifying important features of the data. PCA cannot easily capture nonlinear relationship that frequently exists in high dimensional data, especially in complex biological systems; this is the main drawback of PCA. Classical multidimensional scaling (classical MDS) or Principal Coordinates Analysis that estimates the matrix of dissimilarities for any given matrix input are the similar linear approach for data extraction. It was used for high dimensional gene expression datasets because it is effective in combination with Vector Quantization or K-Means that assigns each observation to a class, from the total of K classes.

- **Non-Linear Feature Extraction**

Non-Linear feature extraction works in different ways for High Dimensional Non-Linear data reduction. In general kernel functions can be considered to create the same effect without using any type of lifting function. Kernel PCA is an important nonlinear method of feature extraction for classification. It has been widely used for biological data. Since, High Dimensional Non-Linear dataality reduction helps in the understanding of the results . Nonlinear feature extraction using Manifolds is another similar approach for dimensional reduction. It has been built on the hypothesis that the data (genes of interest) lie on an embedded nonlinear manifold that has lower dimension than the raw data space and lies within it. Many methods exist working in the manifold space and applied to reduce the High Dimensional Non-Linear data of microarrays, such as Locally Linear Embedding (LLE) and Laplacian Eigenmaps. Kernel PCA and extraction using manifold methods are widely used feature extraction method for dimension reduction of the microarray. Self-organizing maps (SOM) can also be used for reducing the High Dimensional Non-Linear data of gene expression data but it was never generally accepted for analysis. As, it needs just the accurate amount of data to implement well . SOM can often be better separated using manifold LLE but kernel PCA is far faster than the other two. Kernel PCA has an important limitation in terms of space complexity since it stores all the dot products of the training set and therefore, the size of the matrix increases quadratically with the number of data points. Independent component analysis (ICA) is also widely used in microarrays. Independent component analysis (ICA) is a Feature extraction technique, which was proposed by Hyvarinen to solve the typical problem of the non-Gaussian processes and has been applied successfully in different fields. The extraction process of ICA is very similar to the algorithm of PCA. PCA maps the data into another space with the help of principal component. In place of Principal component, the ICA algorithm finds the linear representation of non-Gaussian data so that the extracted components are statistically independent.

- **Hybrid Methods For High Dimensional Non-Linear Data Reduction**

Recently, a hybrid search technique has been used for dimension reduction that was proposed by Huang et al., that has the advantages of the both filter/extraction and the wrapper method. A hybrid High Dimensional Non-Linear data reduction technique consists two stages, in the first step, a filter/extraction method are used to identify best relevant features of the data sets. In the second step, which constitutes a wrapper method, verifies the previously identified relevant feature subsets are verified by a method that gives higher classification accuracy rates. It uses different evaluation criteria in different search stages, to improve the efficiency and classification accuracy with better computational performance . In the hybrid search algorithm, the first subset of features is selected or extracted based on the filter/extraction method and after that the wrapper method is used to select the final feature set. Therefore the computational cost of the wrapper method becomes acceptable due to the use of reduced size features. Information gain and a Memetic algorithm ,Fishers core with a GA and PSO mRAR with ABC algorithm have recently used the hybrid method to solve the problem of dimensionality reduction of the microarray.

## IV.CONCLUSION

This paper has presented two different ways of reducing the dimensionality of high dimensional non-linear microarray data. The first is to select the best features from the original feature set this is called feature selection. On the other hand, feature extraction methods transform the original features into a lower dimensional space by using linear or a nonlinear combination of the original features. To analyze microarray data, dimensionality reduction methods is essential in order to get meaningful results. In this whole paper different aspects feature selection and extraction methods were described and compared. The advantage and disadvantage of these methods are streamlined to get the clear idea about, when to use which method, in order to save computational time and resources. In addition, we have also described a hybrid method that incorporates increasing the classifier accuracy and reducing the computational complexity of an existing method.

## V.REFERENCE

[1]. Chang TW (1983) Binding of cells to matrixes of distinct antibodies coated on solid surface. *J Immunol Methods* 65: 217–223.

[2]. Lenoir T, Giannella E (2006) The emergence and diffusion of DNA microarray technology. *J Biomed Discov Collab* 1: 11–49.

[3]. Pirrung MC, Read LJ, Fodor SPA, et al. (1992) Large scale photolithographic solid phase synthesis of polypeptides and receptor binding screening thereof: US, US5143854[P].

[4]. Peng S, Xu Q, Ling XB, et al. (2003) Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. *Febs Lett* 555: 358–362.

[5]. Eisen MB, Brown PO (1999) DNA arrays for analysis of gene expression. *Method Enzymol* 303: 179–205.
[6]. Leng C (2008) Sparse optimal scoring for multiclass cancer diagnosis and biomarker detection using microarray data. *Comput Biol Chem* 32: 417–425.
[7]. Quackenbush J (2001) Computational analysis of microarray data. *Nat Rev Genet* 2: 418–427.
[8]. Piatetsky-Shapiro G, Tamayo P (2003) Microarray data mining: facing the challenges. *ACM Sigkdd Explor Newslett* 5: 1–5.
[9]. Golub TR, Slonim DK, Tamayo P, et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286: 531–537.
[10]. Lee JW, Lee JB, Park M, et al. (2005) An extensive comparison of recent classification tools applied to microarray data. *Comput Stat Data An* 48: 869–885.
[11]. You W, Yang Z, Yuan M, et al. (2014) Totalpls: local High Dimensional Non-Linear data reduction for multicategory microarray data. *IEEE T Hum Mach Syst* 44: 125–138.
[12]. Xi M, Sun J, Liu L, et al. (2016) Cancer feature selection and classification using a binary quantum-behaved particle swarm optimization and support vector machine. *Comput Math Method Med* 2016: 1–9.
[13]. Wang L, Feng Z, Wang X, et al. (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* 26: 136–138.
[14]. Shen Q, Mei Z, Ye BX (2009) Simultaneous genes and training samples selection by modified particle swarm optimization for gene expression data classification. *Comput Biol Med* 39: 646– 649.
[15]. Pinkel D, Segraves R, Sudar D, et al. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* 20: 207–211.
[16]. Cheadle C, Vawter MP, Freed WJ, et al. (2003) Analysis of microarray data using Z score transformation. *J Mol Diagn* 5: 73–81.
[17]. Saeys Y, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics* 23: 2507–2517.