TECHNICAL UNIVERSITY OF CRETE
DEPARTMENT OF ELECTRONIC AND COMPUTER ENGINEERING
DIGITAL SIGNAL & IMAGE PROCESSING LAB

Comparison of Statistical Methods for Genomic Signature Extraction

Diploma Thesis

Nikolaos-Kosmas Chlis

Thesis Committee
Professor Michael Zervakis, Thesis Supervisor
Associate Professor Costas Balas
Professor Apostolos Dollas

Chania, October 2013

Abstract

In recent years microarray technologies have gained a lot of popularity for their ability to quickly measure the expression of thousands of genes and provide valuable information for linking complex diseases such as cancer to their genetic underpinnings. Feature selection methods are used in order to extract small and informative sets of genes that can maximize the performance of classification methods used to map unknown samples into classes of interest, leading to new and efficient methods for prognosis of several diseases, which are personalized to the genome of each specific patient. Moreover, the biological interpretation of these sets of genes, often referred to as "genomic signatures", can help biologists and physicians better understand the biological processes related to complex diseases, such as cancer and may potentially lead to the discovery of new methods of treatment.

Nevertheless, the large number of parameters to be estimated in relation to the small number of available samples gives rise to an "ill posed" problem where the performance assessment of feature selection and classification methods is not stable under slight changes the dataset. In this thesis, a generic evaluation framework named "Stable Bootstrap Validation" (SBV) is presented, that utilizes resampling of the original dataset and an explicit criterion that determines the stability of the observed classification accuracy, as well as the genomic signature. The proposed methodology works in an iterative manner and converges to a stable solution that combines good accuracy with biologically meaningful feature selection. The methodology is orthogonal to the specific feature selection and classification algorithms used. Moreover, methodologies for assessing the statistical significance and consistency of the observed results are also introduced. Some of the most widely used classifiers are compared, based on their average discrimination power and the size of the derived gene signature. According to our proposed model, a unified '77 common-gene signature' was selected, which is closely associated with several aspects of breast tumorigenesis and progression, as well as patient-specific molecular and clinical characteristics.

Aknowledgements

I would like to thank:

My thesis supervisor, Professor Michalis Zervakis, for his guidance and for giving me the chance to expand my knowledge in the exciting field of bioinformatics.

Dr. Katerina Bei for her support and biological insight.

M.Sc. Stelios Sfakianakis for providing the dataset and sharing his knowledge.

Associate Professor Costas Balas and Professor Apostolos Dollas, for their contribution as members of the thesis committee.

My high school biology teacher, Katerina Papadaki, for inspiring me to work hard and explore the field of biology.

Last but not least; my friends, including my family.

Table of Contents

List of Figures

List of Tables

1 - Introduction

1.1 Introduction to Genomic Analysis

While the mapping of the human genome has has been a subject of study for decades, it was until the more recent advent of DNA microarray technology that scientists have been given a valuable tool in measuring the expression levels of different genes in a biological system. The genomic analysis using DNA microarrays, serves a dual purpose. First, Scientists can observe patterns in the data that can lead to different expression profiles among distinct classes of interest. In that manner, the need arises for identification of sets of genes that strongly differentiate their expression levels among classes of interest. These sets of genes are also called “genomic signatures”. Second, using these sets of genes along with the patters that have been observed, scientists can design classification methodologies that assign class labels to new unknown samples. For example, when sets of genes that differentiate their expression levels between cancerous and non-cancerous tissue samples, they can be used to identify whether an unknown sample belonging to a patient corresponds to cancerous tissue or not. Moreover, these specific genomic signatures can be used to provide insight into biological processes, such as cancer and possibly lead to new methods of treatment.

However, the analysis of genomic datasets is prone to the problem known as “curse of dimensionality” since typically the number of available samples is considerably smaller than the number of features (genes) used for classification. To be precise, the number of samples is usually in the order of a few hundred in a best case scenario, while there are thousands of genes, approximately 20,000 in the human genome. The effect of the “curse of dimensionality” implies significant decrease in classification performance, instability of the derived signature, as well as difficulties in generalizing the results. The above problems call for methods that perform dimensionality reduction by eliminating “irrelevant” sets of features, which are called feature selection methods. There are several categorizations of feature selection methods e.g.: filter methods, following a univariate approach that examines one feature at a time; wrapper and embedded methods, which are multivariate approaches for simultaneously examining different sets of features. Univariate methods select features that strongly differentiate their behaviour between classes of interest and as such, they focus on features aimed at improving class separability. Multivariate methods, aim at selecting a set of features that maximizes the performance of a classification method and aim at selecting sets of features that improve class prediction of unknown samples. In this manner, feature selection as a methodology is often intertwined with the classification process of new samples.

While classification methodologies are often mixed with feature selection to produce sets of informative features, the problem of classification of new samples is also an important aspect of microarray analysis by itself, since it can lead to new and efficient prognosis methodologies. Given that the effect of the “curse of dimensionality” can been counterfeited by feature selection, with an informative and relatively small set of features being extracted, classification methods are used in order to classify new data into known classes of interest. Several different categories of classifiers have been used for this purpose. In this thesis, the classification methods examined are based on regression processes, and include RLS classifiers, PLS classifiers, Support Vector Machines (SVMs) and Nearest Neighbor approaches. A small set of specific features (genes) that achieves a high classification accuracy when used in conjunction with a classification method, is called a “genomic signature”.

Another aspect of DNA microarray analysis is the stability of the observed results. Most evaluation methods for determining the performance of feature selection as well as classification methodologies lead to observations that vary considerably when small variations take place in training and testing data, as well as algorithmic parameters. The need for stability of results has lead to the development of methodologies aimed at extracting more stable, robust and generalizable performance estimates. These methodologies often rely on random sampling or splitting of the original dataset multiple times in order to generate a large number of training, as well as test sets, which are used to infer the performance estimates of a given feature selection and classification scheme. In accordance to this goal, Davis et al. in [1] perform random splitting of the original dataset a large number of times in order to extract stable feature selection and classification performance assessments over all datasets generated. Suzuki et al. in [3] generate multiple dataset using random sampling with replacement and take into account the results of leave one out cross validation over all datasets in order to extract performance estimates. Barrier et al. in [8] utilize Monte Carlo cross validation, splitting the dataset a large number of times in training and test sets of various sizes. Armañanzas et al. [6] propose bootstrap resampling as a means to extract a stable bayesian model of dependent genes. However, while these methodologies lead to stable results, they lack a formal definition of stability, as well as an objective criterion that defines when a sufficient level of stability is reached for the resulting genomic signature and the corresponding classification accuracy. The lack of such a criterion is

bypassed using an arbitrary large number of bootstrap iterations in order to achieve stability, which range from 400 to thousands in the studies mentioned. Considering that feature selection and classification methods tend to be computationally intensive, performing such a large number of iterations can be impractical. Moreover, many of the studies mentioned utilize resampling methods to extract a stable genomic signature but assess classification performance based on typical cross validation techniques [6], [7]. Even if the genes in the signature are stable, the size of the signature itself (i.e. the number of selected genes) may differ considerable during the iterations [6] [7] [8]. This thesis aims at introducing a framework that utilizes an explicit definition of stability and objective criterion for determining when a sufficient level of stability is achieved for the extracted genomic signature and the classification accuracy, while performing a minimum number of bootstrap iterations.

Another need associated with biological problems is to determine whether the results extracted from feature selection and classification, even if they are stable, are observed as a result of the underlying biological system or are merely observed by chance. In this direction, statistical tests determining the randomness of results have been developed. Such tests often utilize permutation in order to measure the statistical significance of the observed results and assess their reliability [2] [19]. Results that are stable and reflect the biological model should also be consistent across different executions of the feature selection and classification methodologies. This aspect is also addressed in our methodological framework, which consists of several generic modules for specific tasks, as outlined in Figure 1.1.

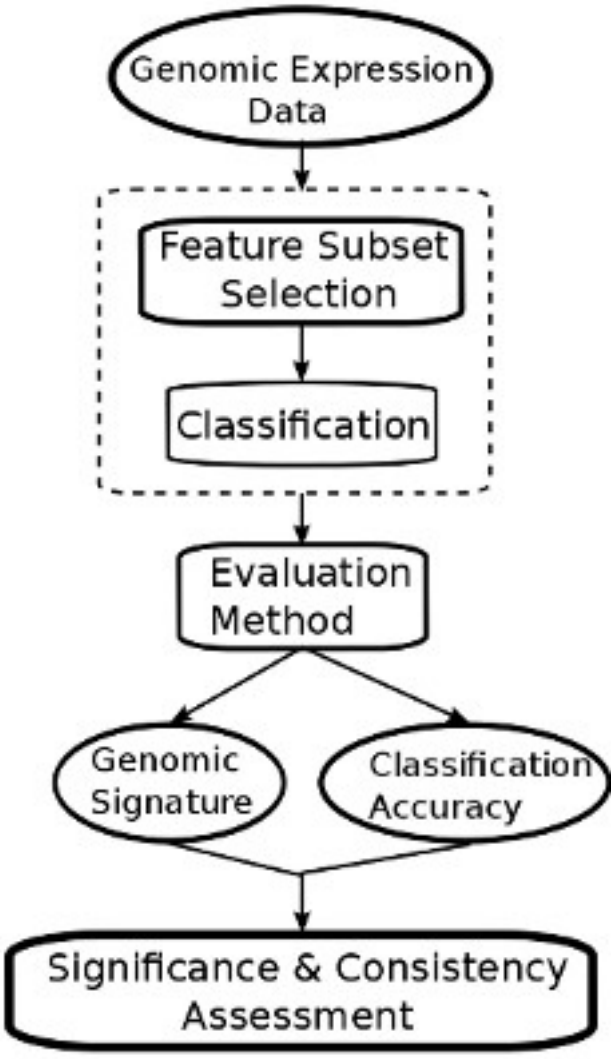

Figure 1.1 Abstract block diagram of the different algorithm categories used in DNA microarray analysis. The genomic dataset is first processed by a combination of feature subset selection and classification methods. The performance estimates of feature selection and classification are extracted using an evaluation method, resulting in a set of selected genes: the genomic signature and a classification accuracy. The significance and consistency of the observed results are then assessed using appropriate methodologies.

1.2 Related Work

The evaluation of stability and reliability of results concerning genomic analysis has been the focus of several studies in the field of Bioinformatics. Many studies focus on random sampling of the original dataset in order to infer stable performance estimates. Bootstrap resampling, that is random sampling with replacement, as a method to estimate the sampling distribution of a random variable based on the observed data was first introduced by B. Efron in 1979 [9]. In the same study bootstrapping was compared to the Jackknife and standard leave one out cross validation, outperforming both methods. Davis et al. in [1] study the stability of genomic signatures and it's impact in the stability of classification accuracy. They also propose a methodology that utilizing random splitting for determining efficient combinations of feature selection and classification models depending on the stability of signatures as well as efficient classification performance. Soek Ying Neo et al. in [2] utilize Monte Carlo simulations in order to determine whether the clustering performance is statistically significant or not. Maglietta et al. In [19] rank each gene not only depending on performance of a ridge regression classifier when only that specific gene is used as a feature, but also examine the statistical significance of that gene's observed classification accuracy. Suzuki et al. in [3] propose a model for the performance assessment of feature selection and classification methods, that takes advantage of the low bias of leave one out cross validation, while it aims to counter it's large estimation variance by utilizing bootstrap resampling. Haury et al. In [4] assess the influence in terms of stability, performance and interpretability, of different feature selection methods when used in conjunction with a set of classifiers. They also compare the performance of the genomic signatures to sets of randomly selected genes, a notion introduced by Ein-Dor et al. in [5]. Armañanzas et al. in [6] propose bootstrap resampling since it leads to reliability, robustness and few false positives in the observed results. They propose a scheme which utilizes bootstrap resampling in order to generate a large number of 1000 datasets and then univariate feature selection method called “correlation feature selection” is performed on each dataset in order to reduce the dimensionality. A k-Dependence Bayesian classifier is then trained using each bootstrap dataset resulting in a directed acyclic graph where each arc represents statistical dependence between the connected nodes (genes). To achieve stability of the model, only arcs whose appearance frequency over all bootstrap datasets is over a fixed threshold, are included in the final model. To assess the classification performance, 5 fold cross-validation is performed. The same approach is followed by García-Bilbao et al. in [7] in order to construct a k-Dependence Bayesian classifier utilizing bootstrap resampling. However, instead of 5 fold cross validation on the constructed model, a set of 10 features selected by the model is used in conjunction with a set of different classification methods and their performance is evaluated using leave one out cross validation. The concept of using bootstrap resampling for the estimation of confidence in selecting a feature in a bayesian network was first introduced by Friedman et al. in [11] it was reported to lead to low rate of false positive rate for selected features and also achieve reliable conclusions about the selected features, even if the dataset used was relatively small. Barrier et al. in [8] propose Monte Carlo cross validation, which generates multiple random splits of the dataset using random sizes for the training and tests sets. That is, for each of the 16 different values for training test size, 100 datasets are performed by random splitting of the dataset leading into 1600 total datasets generated. Then, a filter feature selection method and a diagonal linear discriminant analysis classifier is trained on the training set, while classification performance is assessed using the corresponding test set. In that study it is also reported that many different signatures lead to similar classification performance, a result shared by [4] and [5][8]. Kerr et al. in [10] perform bootstrap resampling from the original dataset in order to assess the stability of cluster analysis results. At the first level of bootstrapping, 10,000 bootstrap simulations are run in order to eliminate irrelevant features using a filter feature selection method. Then, at the second level of bootstrapping 499 additional datasets are generated from the filtered original dataset and each gene is clustered to one of 7 possible temporal patterns of yeast sporulation. Finally, the gene clusterings considered stable are only those being “95% stable” , that is they appear in at least 95% of the generated datasets, as well as in the clusters of the original dataset.

1.3 Thesis Outline and Innovation

The necessary theoretical background concerning the human genome and methodologies concerning the analysis of DNA microarray data in the field of bioinformatics is covered in chapter 2. That includes the biological concepts regarding DNA microarrays, feature selection and feature weighting methodologies, as well as classifiers. Different evaluation methods are also presented, followed by an introduction to the statistics theorem known as the “law of large numbers”. The proposed methodology for extraction of stable signatures and performance estimates, while assessing the statistical significance and consistency of results is covered in chapter 3. The innovative concept involves utilizing bootstrap resampling in order to generate a large number of datasets for training and testing a pair of feature selection and classification methods. Under the assumption that the observed classification accuracy and signature size are independent identically distributed random variables, according to the Law of Large Numbers the evaluation methodology is guaranteed to lead to stable assessment of both metrics, given that the number of bootstrap datasets used is large enough. Moreover, unlike similar methods the proposed methodology employs an explicit criterion that determines when stability has been achieved for the mean classification accuracy, as well as the genomic signature size. The results of the proposed methodology, including the performance estimates of several feature selection and classification techniques are presented in chapter 4, followed by consistency assessment of the accuracy reached by the proposed genomic signatures. Moreover a comparison between the proposed evaluation technique and standard 10-Fold cross validation is implemented. Finally, the biological interpretation of the extracted genomic signatures is presented.

2 - Theoretical Background

In this chapter the necessary background concerning the human genome and the bioinformatics aspects of DNA microarray analysis are covered. The human genome and the technology of DNA microarrays is introduced in section 2.1, followed by an introduction to the scientific field of machine learning and pattern recognition in section 2.2. Then, the subject of feature subset selection is examined in section 2.3 including the differences of filter, wrapper and embedded methods, while the recursive feature elimination algorithm is introduced as well. Different classification methods are covered in section 2.4. including regularized least squares, partial least squares, support vector machines and nearest neighbor classifiers. In section 2.5 different cases of evaluation methods are examined, such as holdout validation, K-fold cross validation, leave one out cross validation, repeated random sub-sampling validation and bootstrap resampling. Finally, a theorem of statistics called the “weak law of large numbers” is introduced in section 2.6.

2.1 The Human Genome - DNA Microarrays

In this section the necessary background is covered concerning the structure of the human genome as well as measuring expression values of different genes using the technology of DNA microarrays.

The Human Genome

The human genome refers to the complete set of human genetic information, the study, analysis and mapping of which, has been the subject of the “Human Genome Project”[12]. The majority (~98%) of the human genome located in genetic material in the nucleus of human cells (with the exception of red blood cells), while the rest (~2%) is located in organelles called mitochondria which are responsible for converting the energy from food into a form usable by human cells. The genome located in the nucleus is organized into 23 pairs of chromosomes. These 46 chromosomes consist of 44 autosomes and 2 sex chromosomes, XX or XY for females and males respectively. Every chromosome has a constriction along it's length, called the centromere that divides the chromosome into a long and a short “arm”. Each chromosome can be thought as a string of thousands of genes, which are in turn made of DNA. The human genome is made of approximately 20,000 genes, most of them located in the nucleus, while only 37 refer to mitochondrial genes. Moreover, the genes located in the nucleus are not organized in chromosomes. The DNA that makes up the genes is called “coding DNA”, while the DNA “string” between each gene is called “non-coding DNA”. Only a fraction of the genome refers to coding DNA, which is transcribed into RNA and then transcribed into proteins. Most of the genome consists of non-coding DNA that is associate with other known, or yet unknown, biological procedures.

DNA

As mentioned above, each gene is made of DNA [13]. Deoxyribonucleic acid (DNA) consists of two long complementary strands of nucleotides that take the form of a double stranded helix. DNA consists of four primary types of nucleotide molecules. Each nucleotide consists of a phosphate, a sugar (deoxyribose) and one of four possible nitrogen bases, each represented by a letter: adenine (A), guanine (G), cytosine (C) and Thymine (T). These distinct nitrogen bases are also used to distinguish the four types of nucleotides from one another. Each nucleotide of a strand is connected by a hydrogen bond to it's complementary nucleotide in the opposing DNA strand in order for the helix to maintain it's structure independent of the nucleotide sequence. These complementary nucleotide pairs are called the base pairs and correspond to G-C and A-T. The genetic information of each strand is read in the form of non-overlapping triplets of nucleotides. Given that there are 4 nucleotides, the possible number of different triplets is equal to $4^3=64$ combinations.

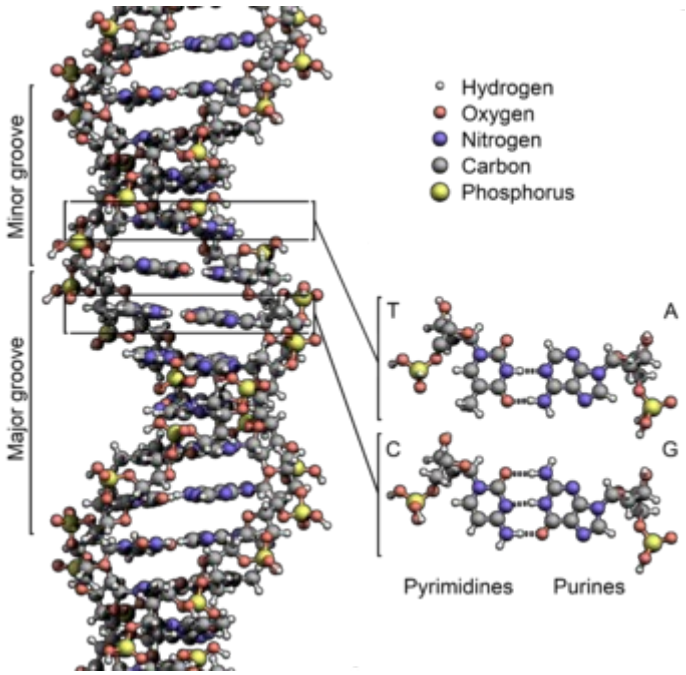

Figure 2.1 The DNA double helix. The nucleotide base pairs of A-T and G-C are also shown.

DNA Microarrays – Gene Expression – DNA transcription

DNA Microarrays [13] [14] are tools that allow the measurement of the expression levels of different genes. A gene is considered to be expressed if it's DNA has been transcribed to RNA and gene expression refers to the level of transcription of the gene's DNA. During the process of transcription the DNA is used as a template for the enzyme RNA polymerase II to construct pre-mRNA utilizing complementary base pairing. However, since there is no Thymine in RNA, it is replaced by Uracile (U). Finally, the enzyme recognizes signals in the DNA chain that lead to the termination of the transcription process and the pre-mRNA chain is released into the nucleus where it is processed into mRNA. DNA microarrays measure the levels of mRNA. DNA microarrays measure gene expression assessing the levels of mRNA present in the samples of interest indirectly. The assessment is indirect since DNA microarrays in reality measure the levels of cDNA, which is produced by mRNA using a process called Reverse Transcription (RT). The cDNA sequences used to bind target cDNA sequences of interest on the microarray are called “probes”. Probes bind target cDNA sequences by forming hydrogen bonds between complementary nucleotide base pairs, while multiple probes may be used to measure the same gene in order to reduce the noise present in the signal. The sequences bound by the probes are then detected using fluorescent dyes. If the genes of interest are found to be expressed, their expression levels are compared to those of known control samples in which the same genes are not expressed. Different technologies of DNA microarrays have been introduced. The “spotted cDNA microarray” developed at Stanford University utilizes robotic spotting of aliquots of purified cDNA clones, while category of microarrays developed by Affymetrix, Inc. Utilizes photo-lithography for embedding cDNA probes on silicon chips.

2.2 Machine Learning and Pattern Recognition

In machine learning [15], pattern recognition is the act of selecting an appropriate action based on the patterns observed in raw data.

Supervised learning [15] aims to generate a function given a set of labeled samples. That function can then be used to assign labels to new unknown data. In regression, the label of each sample, is a continuous variable, often called the response variable. In the case of classification, the label can only take one among a set of discrete values.

Unsupervised learning [15] aims to find groups of data that share similar properties. It differentiates from supervised and reinforcement learning, since the samples are unlabeled and there is no explicit feedback.

Reinforcement learning [15] is usually employed by A.I. agents and aims to maximize a cumulative reward function, given a set of variables determining the environment and the actions available at a given time. Instead of labels, reinforcement learning utilizes a positive or negative reward signal sent to the agent after an action is completed.

DNA microarray analysis is a case of supervised learning. The raw data consists of a set of N samples, each represented by a vector $\mathbf{x}_i \in R^P$ i=1, ...,N. Where P is the number of features/genes, also called predictors. To each of the samples, a class label y is assigned. In the case of cancer/control binary classification, $y \in \{-1,+1\}$. The data can also be expressed in array form as $\mathbf{X} \in R^{N,P}$ where each row represents a sample containing the expression values of P genes, while the class labels of all samples are expressed as a vector $\mathbf{y} \in R^N$.

2.3 Feature Subset Selection (FSS)

Feature subset selection [17] [18] is an important aspect of microarray analysis, since it aims to counter the "curse of dimensionality" that is encountered in DNA microarray datasets. That is, classifier performance is deteriorated when the number of features is larger than the number of available training samples. The goal of FSS methods is to reduce the number of features by keeping only the most "important" set, while discarding all others. The set of kept features is then used for classification. In DNA microarray analysis, the set of kept features (genes) is usually referred to as "genomic signature". There are three different approaches to feature subset selection: filter, wrapper and embedded methods.

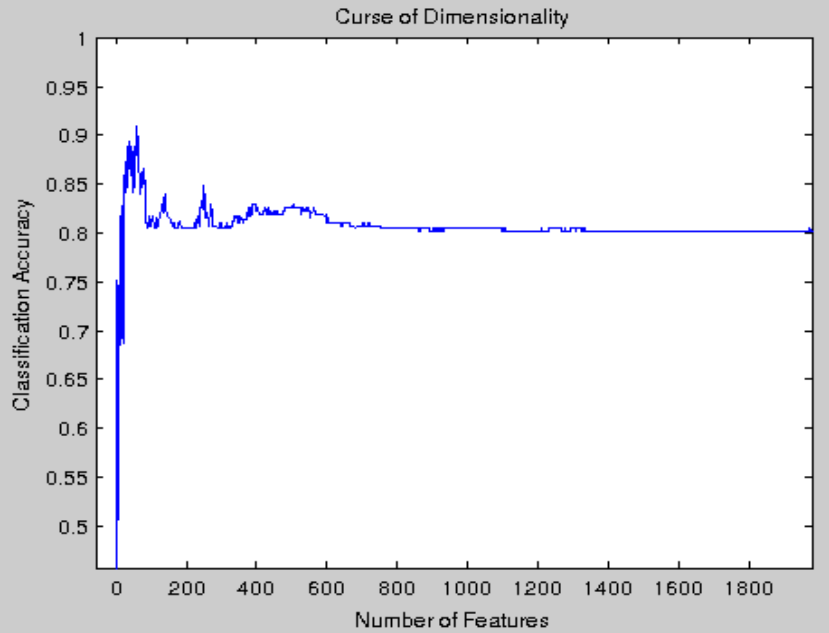

Figure 2.2 Demonstration of the curse of dimensionality on DNA microarray data, using a linear SVM classifier. Classification performance deteriorates when the number of features is comparable to, or greater than the number of available samples.

Filter Methods (Univariate)

Filter methods [17] [18] form univariate approaches, which act as a preprocessing step, independent of the classifier used. They rank each feature independent of others, based on its ability to discriminate between different classes of interest. They generally are simple to implement, computationally efficient and provide insight into class differences. However, filter methods produce a feature set that is not tuned to the performance of a specific classifier.

Wrapper Methods (Multivariate)

Wrapper methods [17] [18] fall within a multivariate approach. They evaluate a feature subset based on the prediction accuracy of the classifier when that specific subset is used. In that manner, given a classifier, they aim to find the set of features which maximizes the prediction performance. The classifier is perceived as a black box, independent of the feature selection method. Since they need to evaluate different combinations of features, they can be computationally expensive. In that manner, greedy algorithms have been proposed in order to reduce the computational complexity, such as forward selection and backward elimination.

Embedded Methods (Multivariate)

Embedded [17] [18] methods also evaluate a feature subset based on the prediction accuracy of the classifier. They differentiate from wrapper methods however, since the search for the feature subset is embedded in the training of the classifier, while in wrapper methods the feature selection step is independent of the classifier used. Compared to wrapper methods, embedded methods are more computationally efficient. However, due to the embedding of feature selection in the training process, they can prove to be harder to implement.

2.3.1 Recursive Feature Elimination (RFE)

Recursive Feature Elimination [16] is a popular embedded feature selection method that aims at preserving the minimal set of features maximizing the classification accuracy of a given classification method. RFE proceeds iteratively, eliminating a fixed number of least significant features during each iteration and then reassessing the classification performance. The elimination procedure stops when a predetermined small number of features are left. Then, the set of features across all iterations maximizing the classification accuracy is chosen as the optimal feature set, tuned for the specific classifier used. In order for the least significant feature to be determined, a feature weighting scheme is required. Such a weighting scheme can be the weight given to each feature by a linear classifier or by non-linear feature weighting methods such as RELIEF.

2.3.2 Feature Weighting Methods and I-RELIEF

As mentioned above, the weighting of features is a prerequisite for the implementation of a recursive feature elimination scheme. That is, feature weights are necessary in order to determine the least significant set of features during each iteration of RFE. In this study, two categories of feature weighting methods are examined: linear and non-linear feature weighting algorithms.

Linear Feature Weighting Methods – Linear Classifiers

The most common method to assign weights to features are linear classifiers, including RLS methods like RR and the LASSO, PLS methods like PLS-VIP and PLS-BETA and linear SVM. All these classifiers are presented in detail in section 2.4. The common characteristic of these methods is that they assign a label $\hat{y}$ to an unknown sample $\hat{\boldsymbol{x}}$ based on the formula $\hat{y} = f\left(\hat{\boldsymbol{x}} \cdot \boldsymbol{w}\right)$,while the weight vector w is inferred during the training process of the classifier.

Non-Linear Feature Weighting Methods

Non-linear feature weighting methods assign weights to features according to a non-linear criterion. In this study, the original RELIEF algorithm, as well as I-RELIEF are examined. Both are used in conjunction with the K-NN classifier which is introduced in section 2.4. They rank features according to their ability to discriminate the neighboring samples of different classes.

RELIEF

RELIEF [26] is a feature weighting algorithm that practically is an online solution to a convex optimization problem that maximizes a margin based objective function, which is defined by a 1-Nearest Neighbor classifier. Unlike methods such as Support Vector Machines that assign feature weights and perform classification of samples, RELIEF by itself is a methodology dedicated to feature weighting and does not provide a means to classify samples. In this manner, RELIEF is used in conjunction with a Nearest Neighbor classification method. RELIEF tends to perform better than filter methods since it utilizes a non-linear classifier and is faster than wrapper methods because it is expressed in closed form as a solution to an optimization problem, so only minimal computations are required. In the case of binary classification and a given dataset , RELIEF ranks features $D=\{(\boldsymbol{x_n}, y_n): \boldsymbol{x_n} \in R^P \; y_n \in \{-1,+1\}\}, n=1,\dots,N$ based on their ability to discriminate the classes of neighboring samples. Let NH(**x**n) be the "Nearest Hit", the nearest sample to **x**n belonging to the same class and NM(**x**n) the "Nearest Miss", the nearest sample to **x**n belonging to the different class. The margin for **x**n can be described as $\rho_n = d(\boldsymbol{x_n} - NM(\boldsymbol{x_n})) - d(\boldsymbol{x_n} - NH(\boldsymbol{x_n}))$, where $d(\cdot)$ a distance function defined as $d(\boldsymbol{x}) = \sum_{p=1}^{P} |x_p|$. In that case $\rho_n > 0$ only if the sample has been correctly classified. The idea is to use feature weights in order to scale each feature so that the average margin in the weighted feature space is maximized:

maximize in **w**

$$\sum_{n=1}^{N} \left(\sum_{p=1}^{P} w_p \left| \boldsymbol{x_n}^{(p)} - NM^{(p)}(\boldsymbol{x_n}) \right| - \sum_{p=1}^{P} w_p \left| \boldsymbol{x_n}^{(p)} - NH^{(p)}(\boldsymbol{x_n}) \right| \right)$$

subject to $\|\boldsymbol{w}\|_2^2 = 1, \boldsymbol{w} \geq 0$

Where $\|\boldsymbol{w}\|_2^2 = 1$ is used so that the maximization does not increase without bound and $\boldsymbol{w} \geq 0$ ensures that the weight vector **w** is a distance metric. In order to simplify the above optimization problem we define the vector **z** as $\sum_{n=1}^{N} |\boldsymbol{x_n} - NM(\boldsymbol{x_n})| - |\boldsymbol{x_n} - NH(\boldsymbol{x_n})|$ the above problem can be rewritten as

maximize in **w**

$$\boldsymbol{w}^T \cdot \boldsymbol{z}$$

subject to $\|\boldsymbol{w}\|_2^2 = 1, \boldsymbol{w} \geq 0$

By utilizing the Lagrangian technique and the Karush-Kuhn-Tucker condition, the solution is expressed in closed form as $\boldsymbol{w} = \frac{(\boldsymbol{z})^+}{\|\boldsymbol{z}\|^+}$, where $(z_i)^+ = max(z_i, 0)$. In case the Euclidean distance is used instead of the above distance metric, the resulting algorithm is called Simba [28]. After feature weighting has taken place, feature subset selection can be performed based on the absolute value of the weight of each sample.

I-RELIEF

I-RELIEF [27] is a feature weighting scheme that aims to improve on the drawbacks of the original RELIEF algorithm. The first drawback of RELIEF is that it makes the implicit assumption that the nearest neighbors of a sample in the original feature space are the same in the weighted feature space, which is generally not the case. The second drawback of RELIEF is it's sensitivity to outliers. I-RELIEF follows the

principle of the EM algorithm and threats the nearest neighbor as well as the identity (outlier or not) of a sample as hidden variables and iteratively estimates them until convergence is achieved. Given the dataset $D=\{(\boldsymbol{x_n}, y_n): \boldsymbol{x_n}\in R^P\ y_n\in\{-1,+1\}\}, n=1,\dots,N$, two sets are determined for each sample $\boldsymbol{x_n}$: $M_n=\{i: 1\le i\le N, y_i\neq y_n\}$ the set of all samples with a label different than $\boldsymbol{x_n}$ and $H_n=\{i: 1\le i\le N, y_i=y_n, i\neq n\}$ the set of all samples sharing the same label as $\boldsymbol{x_n}$. Let $S_n=\{s_{n1}, s_{n2}\}$ where $s_{n1}\in M_n$ the nearest miss of $\boldsymbol{x_n}$ and $s_{n2}\in H_n$ the nearest hit of $\boldsymbol{x_n}$. Moreover, the vector of binary parameters $\boldsymbol{o}=[o_{1,\dots}, o_N]^T$ such as $o_n=0$ if $\boldsymbol{x_n}$ is an outlier, otherwise $o_n=1$. The objective function that needs to be optimized is $C(\boldsymbol{w})=\sum_{\{n=1,\, o_n=1\}}^{N}\|\boldsymbol{x_n}-\boldsymbol{x_{s1}}\|\cdot w-\|\boldsymbol{x_n}-\boldsymbol{x_{s2}}\|\cdot w$. However, the set $S=\{S_n\}_{n=1}^{N}$ and the outlier vector **o** are still unknown at this point. Under the assumption that the elements of both $S=\{S_n\}_{n=1}^{N}$ and **o** are random variables, the method proceeds to iteratively derive their probability distributions. First, the weight vector w is set to an initial value that does not violate the constraints $\|\boldsymbol{w}\|_2^2=1, \boldsymbol{w}\ge 0$. Given that all pairwise distances among training samples have already been computed when searching for nearest hits and misses, the probability of the i-th sample being the nearest miss of $\boldsymbol{x_n}$ is defined as $P_m(i|\boldsymbol{x_n}, \boldsymbol{w})=\frac{f(\|\boldsymbol{x_n}-\boldsymbol{x_i}\|\cdot w)}{\sum_{j\in M_n} f(\|\boldsymbol{x_n}-\boldsymbol{x_i}\|\cdot w)}$ and the probability of the i-th sample being the nearest hit of $\boldsymbol{x_n}$ is defined as $P_h(i|\boldsymbol{x_n}, \boldsymbol{w})=\frac{f(\|\boldsymbol{x_n}-\boldsymbol{x_i}\|\cdot w)}{\sum_{j\in H_n} f(\|\boldsymbol{x_n}-\boldsymbol{x_i}\|\cdot w)}$. Where $f(\cdot)$ a kernel function, with a commonly used example being $f(d)=e^{-d/\sigma}$, the kernel width σ being a user defined parameter. In a similar manner, the probability of $\boldsymbol{x_n}$ being an outlier can be defined as $P_o(o_n=0|D, \boldsymbol{w})=\frac{\sum_{i\in M_n} f(\|\boldsymbol{x_n}-\boldsymbol{x_i}\|\cdot w)}{\sum_{\boldsymbol{x_i}\in D-\boldsymbol{x_n}} f(\|\boldsymbol{x_n}-\boldsymbol{x_i}\|\cdot w)}$. In order to estimate these probabilities an iterative algorithm similar to EM will be implemented. However it needs to be noted that it is not an EM algorithm since the objective function is not a likelihood. For brevity of notation, the following formulas are defined: $\alpha_{i,n}=P_m(i|\boldsymbol{x_n}, \boldsymbol{w}^{(t)})$, $\beta_{i,n}=P_h(i|\boldsymbol{x_n}, \boldsymbol{w}^{(t)})$, $\gamma_n=1-P_o(o_n=0|D, \boldsymbol{w}^{(t)})$, where t refers to the t-th iteration of the algorithm. $W=\{\boldsymbol{w}: \|\boldsymbol{w}\|_2^2=1, \boldsymbol{w}\ge 0\}$, $\boldsymbol{m_{n,i}}=|\boldsymbol{x_n}-\boldsymbol{x_i}|\ \text{if i}\in M_n$ and $\boldsymbol{h_{n,i}}=|\boldsymbol{x_n}-\boldsymbol{x_i}|\ \text{if i}\in H_n$. The following iterative algorithm of I-RELIEF consists of the following two steps:

step 1: After the t-th iteration, calculate the Q function as:

$$Q(\boldsymbol{w}|\boldsymbol{w}^{(t)})=\boldsymbol{E}_{\{S,o\}}[C(\boldsymbol{w})]=\sum_{n=1}^{N}\gamma_n\Big(\sum_{i\in M_n}\alpha_{i,n}\|\boldsymbol{x_n}-\boldsymbol{x_i}\|\cdot\boldsymbol{w}-\sum_{i\in H_n}\beta_{i,n}\|\boldsymbol{x_n}-\boldsymbol{x_i}\|\cdot\boldsymbol{w}\Big)$$

$$=\sum_{n=1}^{N}\gamma_n\Big(\sum_j w_j\sum_{i\in M_n}\alpha_{i,n}m_{n,i}^j-\sum_j w_j\sum_{i\in H_n}\beta_{i,n}h_{n,i}^j\Big)=\sum_{n=1}^{N}\gamma_n\Big(\sum_j w_j\bar{m}_n^j-\sum_j w_j\bar{h}_n^j\Big)$$

$$=\boldsymbol{w}^T\sum_{n=1}^{N}\gamma_n(\bar{\boldsymbol{m}}_{\boldsymbol{n}}-\bar{\boldsymbol{h}}_{\boldsymbol{n}})=\boldsymbol{w}^T\cdot\boldsymbol{v}$$

step 2: re-estimate w in the (t+1) iteration using the formula:

$$\boldsymbol{w}^{(t+1)}=arg\max_{\boldsymbol{w}\in W}Q(\boldsymbol{w}|\boldsymbol{w}^{(t)})=(\boldsymbol{v})^+/\|(\boldsymbol{v})\|$$

The above two steps iterate alternatively until convergence is achieved, that is until $\|\boldsymbol{w}^{(t)}-\boldsymbol{w}^{(t+1)}\|<\theta$.

It can be proved that I-RELIEF always converges, provided that the kernel width σ is large enough [27]. Moreover, since the initial version of relief is a batch algorithm for binary classification of samples, extensions that cover online learning and multiclass classification are introduced [27].

2.4 Classification Methods

2.4.1 Regularized Least Squares (RLS) Classifiers

Linear Regression

Regression [19] [20] is a statistical model for estimating the relationship among the observed and response variables of a system. The regression model is linear, when the response variable is modeled as a linear combination of the observed variables. The above problem is expressed in matrix form as $y = X \cdot w + \varepsilon$ In DNA microarray analysis the response variable **y** $\in R^N$ is the vector of class labels (cancer/control) while the observed variables are the gene expression measurements per sample expressed as **X** $\in R^{N,P}$ in matrix form. Finally, the weight vector **w** $\in R^P$ is the vector of regression coefficients that need to be estimated and **ε** is the error term that corresponds to random noise.

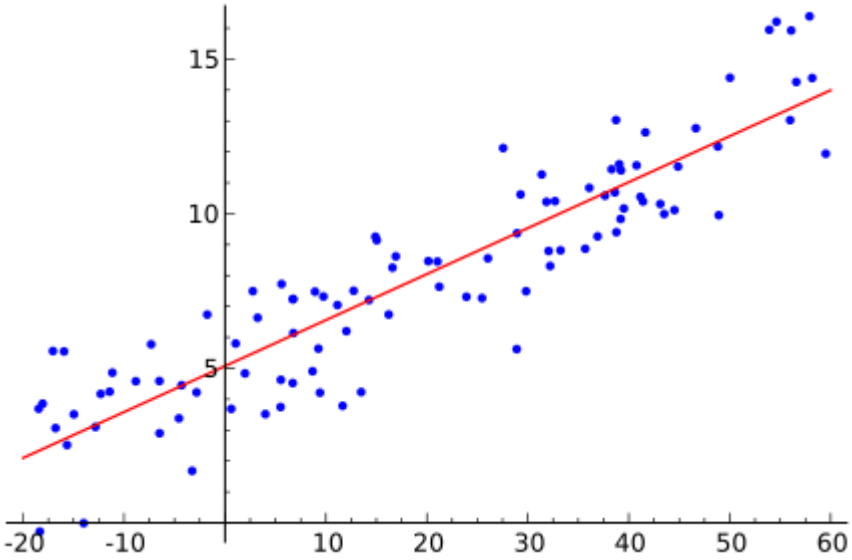

Figure 2.3 Linear regression example, of one independent variable on the x-axis.

Ordinary Least Squares (OLS) [20]

Given a set of N samples $x_n \in R^P$ n=1, ...,N of P features expressed as in matrix form as **X** $\in R^{N,P}$ and a response variable $y \in \{-1,+1\}$ for each sample, the ordinary least squares method aims to infer a function that estimates the labels $\hat{y}$ of a new set of test samples $\hat{X}$. The function suggested by the OLS model is the linear approach $\hat{y} = \hat{X} \cdot w$. In order to solve the OLS problem, the weight vector **w** needs to be estimated. Following the OLS approach, the optimal vector w is the one that minimizes the function:

$$w = argmin\ f(w) \quad , \quad f(w) = \sum_{n=1}^{N} y_n - y'_n = \sum_{n=1}^{N} (y_n - x_n \cdot w)^2$$

Regularized Least Squares (RLS)

While standard OLS approach provides a solution to the classification of new samples, it achieves low classification accuracy and generally does not provide insight into the importance of different features. In that manner, RLS methods intent to improve the performance of the standard OLS approach by further restraining the weight vector **w**. The importance of each feature for OLS and RLS methods is related to the ability of correctly predicting the response variable, when given a set of input samples.

2.4.1.1 Ridge Regression (RR)

The RR approach [19] [20] replaces f(**w**) of OLS with

$f(\boldsymbol{w})=\sum_{n=1}^{N}(y_n-\boldsymbol{x_n}\cdot\boldsymbol{w})^2, subject\ to \sum_{p=1}^{P} w_p^2<t$, t can be estimated using cross-validation.

The added constraint aims at preserving the most important features. By limiting the total sum of squared weights, the most important features are more likely to attract larger weight values while the less significant features will be given values close to 0. As such, the most important features in RR will have a greater impact at the classification process, compared to the case of OLS.

2.4.1.2 Least Absolute Shrinkage and Selection Operator (LASSO)

While RR shrinks a lot of features, it does not implement variable selection since it does not set any of them to exactly 0. That leads to a model which is not easily interpretable. In that manner, LASSO [21] aims to further shrink the available features while setting a considerable amount of them at exactly 0. The LASSO approach replaces f(**w**) with

$f(\boldsymbol{w})=\sum_{n=1}^{N}(y_n-\boldsymbol{x_n}\cdot\boldsymbol{w})^2, subject\ to \sum_{p=1}^{P} |w_p|<t$, t can be estimated using cross-validation.

The LASSO constraint is more limiting than the case of RR. Since feature weights are typically small numbers, smaller than 1. When the weight values are squared they lead to even smaller values and the RR constraint for the sum of squared weights is achieved, while the distinct weight of each feature is generally small but larger than 0. However, the LASSO constraint limits the absolute value of the sum of weights. As such, minimizing the distinct weight of each feature is more important in achieving the limitation of the constraint, compared to the case of RR. That leads to a large number of less important features being assigned weights that are exactly 0, leading to an embedded process of feature selection in the LASSO classification method.

Estimating the parameter t

Both RR and LASSO require the estimation of t beforehand. Given w_0 the estimates of the simple OLS model, then t can be expressed as $t=\alpha\cdot\sum_{p=1}^{P} w_{0_i}^2, \alpha\in[0,1]$. In that manner, α needs to be estimated instead of t. Cross validation for the estimation of α is easier since it only takes values between 0 and 1.

2.4.2 Partial Least Squares (PLS) Classifiers

Partial Least Squares Regression

PLS regression [22] [23] [24] aims to counter the effects of multicollinearity as well as the fact that the number of features is larger than the number of available samples, since both of these factors lead to poor performance of standard regression models. Instead of adopting a feature selection approach like the LASSO, the PLS approach simultaneously decomposes the input data matrix **X** as well as the response variable vector **y** utilizing a set of latent variables that aim to explain the covariance structure between **X** and **y**. PLS decomposition is focused on extracting latent variables that model the observations **X** and can also adequately predict the response variable **y**. The model is expressed as follows:

$\boldsymbol{X}=\boldsymbol{T}\,\boldsymbol{P}^{\boldsymbol{T}}+\boldsymbol{E}$

$\boldsymbol{y}=\boldsymbol{T}\,\boldsymbol{b}+\boldsymbol{f}$

Where **X** $\in R^{N,P}$ is the input data matrix, **y** $\in R^{N}$ the vector of response variables.

T (Scores) $\in R^{N,h}$ is the structural part of PLS that implies how the different rows of **X** (observations) relate to each other. The k-th column of **T** includes the scores for the k-th latent variable. The method assumes that the scores **T** are good predictors of **y** [23]. In this manner, **T** is used to model **X** and predict **y** as well.

P (Loadings) $\in R^{P,h}$ is also a structural part of PLS. The loadings show the influence of different observations of **X** on the scores of **T**.

E (Residuals) is the error that is not predicted by the $\boldsymbol{T}\,\boldsymbol{P}^{\boldsymbol{T}}$ part of the model. In that manner, E should not be large since it would lead to poor performance of the model.

b $\in R^h$, the k-th element of b explains the relation between the response variable vector **y** and and the k-th column vector of the scores **T**.

f $\in R^N$ refers to random errors in the representation of **y** that is not predicted by the model.

Finally, h is the number of latent variables. The value of h is selected beforehand and is theoretically bound by $1 \leq h \leq min\{N, P\}$. However, in practice only the first few components are utilized otherwise a considerable amount of noise is embedded in the structural part of PLS.

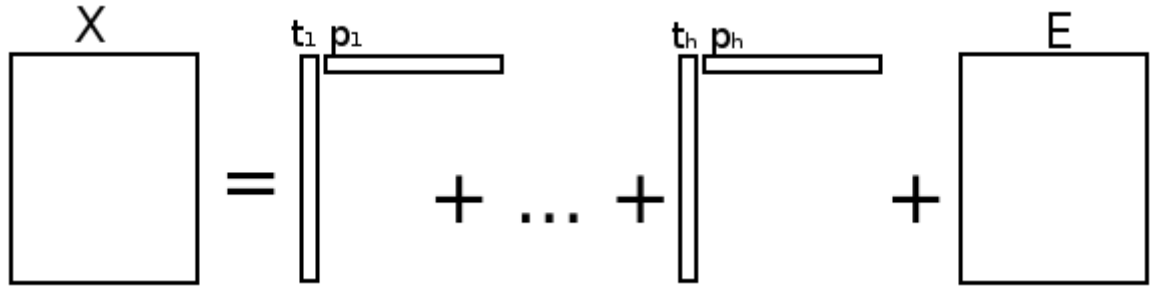

Figure 2.4 PLS decomposition of the input data matrix **X**.

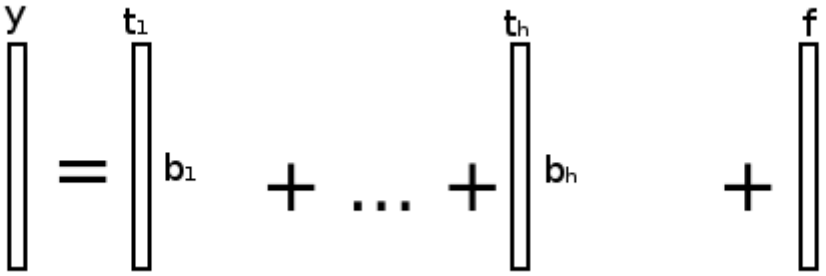

Figure 2.5 PLS decomposition of the response variable vector **y**.

Non-Linear Iterative Partial Least Squares (NIPALS) Algorithm

Given **X**, **y** and a fixed number of h latent variables, the NIPALS algorithm [25] is used in order to calculate **T**, **P**, **W** and **b**. NIPALS is an iterative approach that calculates a single latent variable during each iteration, but leads to the same results as methods that examine the covariance of **X** and **y** [24]. Since the number of latent variables used in genomic datasets is typically small, NIPALS tends to be a computationally efficient approach. However, the version of NIPALS implemented in this thesis and introduced in [22] assumes that the model parameters are adequately estimated in one iteration, in order to further reduce the computational load. As such, the PLS model is constructed using the NIPALS estimations of the model parameters. **W** $\in R^{P,h}$ is a weight matrix obtained in order to minimize the Euclidean norm of **f** in order to derive a useful relation between **X** and **y**. The NIPALS algorithm consists of the following steps, for k=1, …, h:

step 1: $\boldsymbol{y}_{(k)} \Leftarrow \boldsymbol{y}_{(k-1)} - \boldsymbol{b}_{k-1}\boldsymbol{t}_{k-1}; \boldsymbol{y}_{(1)} \Leftarrow \boldsymbol{y}$
$\boldsymbol{X}_{(k)} \Leftarrow \boldsymbol{X}_{(k-1)} - \boldsymbol{t}_{k-1}\boldsymbol{p}_{k-1}^T; \boldsymbol{X}_{(1)} \Leftarrow \boldsymbol{X}$

step 2: $\boldsymbol{w}_k^T = \boldsymbol{y}_{(k)}^T \boldsymbol{X}_{(k)} / \boldsymbol{y}_{(k)}^T \boldsymbol{y}_{(k)}$

step 3: $\boldsymbol{w}_k \Leftarrow \boldsymbol{w}_k / \|\boldsymbol{w}_k\|$

step 4: $\boldsymbol{t}_k = \boldsymbol{X}_{(k)} \boldsymbol{w}_k / \boldsymbol{w}_k^T \boldsymbol{w}_k$

step 5: $\boldsymbol{p}_k^T = \boldsymbol{t}_k^T \boldsymbol{X}_{(k)} / \boldsymbol{t}_k^T \boldsymbol{t}_k$

step 6: $\boldsymbol{t}_k \Leftarrow \boldsymbol{t}_k \cdot \|\boldsymbol{p}_k\|$

step 7: $\boldsymbol{w}_k \Leftarrow \boldsymbol{w}_k \cdot \|\boldsymbol{p}_k\|$

step 8: $\boldsymbol{p}_k \Leftarrow \boldsymbol{p}_k / \|\boldsymbol{p}_k\|$

step 9: $\boldsymbol{b}_{(k)} = \boldsymbol{y}_{(k)}^T \boldsymbol{t}_k / \boldsymbol{t}_k^T \boldsymbol{t}_k$

2.4.2.1 PLS-VIP Method

The PLS-VIP method [25] performs feature selection based on the Variable Importance in Projection (VIP) score of each feature. The VIP score implies each feature's significance in finding the h latent variables during the decomposition of the input data matrix. Since the average VIP score equals to 1, usually a 'greater than one' rule is used during VIP based feature selection.

Given that **T**, **P**, **W** and **b** have been calculated using NIPALS. The VIP score for the j-th feature is calculated using the following formula:

$$VIP_j = \sqrt{\frac{p\sum_{k=1}^{h} SS(b_k \boldsymbol{t_k})(\boldsymbol{w_{jk}}/\|\boldsymbol{w_k}\|)^2}{\sum_{k=1}^{h} SS(b_k \boldsymbol{t_k})}} \quad \text{, where} \quad SS(b_k \boldsymbol{t_k}) = b_k^2 \boldsymbol{t_k^T t_k}$$

After the feature selection procedure has been completed, new samples $\hat{\boldsymbol{X}}$ are classified using the formula $\hat{\boldsymbol{y}} = \hat{\boldsymbol{X}}\,\boldsymbol{b_{pls}}$, where $\boldsymbol{b_{pls}} = \boldsymbol{W}(\boldsymbol{P^T W})^{-1}(\boldsymbol{T^T T})^{-1}\boldsymbol{T^T y}$, **y** the known class labels of the training set.

2.4.2.2 PLS-BETA Method

PLS-BETA [25] is almost identical to PLS-VIP. They both utilize NIPALS in order to calculate **T**, **P**, **W** and **b**. Their difference lies in the criterion for feature selection. While PLS-VIP performs feature selection using the VIP score, PLS-BETA selects features based on the magnitude of their respective regression coefficients in $\boldsymbol{b_{pls}}$. Since variables are selected according to the “beta” vector $\boldsymbol{b_{pls}}$ of regression coefficients, the method is called PLS-BETA.
Similar to PLS-VIP, new samples are classified using the $\hat{\boldsymbol{y}} = \hat{\boldsymbol{X}}\,\boldsymbol{b_{pls}}$ formula, where $\boldsymbol{b_{pls}} = \boldsymbol{W}(\boldsymbol{P^T W})^{-1}(\boldsymbol{T^T T})^{-1}\boldsymbol{T^T y}$.

2.4.3 Support Vector Machine (SVM) Classifier

Support Vector Machines [25] are a machine learning algorithm than can be used for regression, and by extent for classification purposes. In the case of two-way classification, the SVM computes the hyperplane separating the classes of interest with the maximum margin across the closest samples of the two classes. The aim of the utilization of the maximum margin hyperplane is to minimize the generalization error of the classifier. The original SVM algorithm assumes that the data are linearly separable. If that is not the case, using a kernel function the data are mapped to a higher dimension space in which they are found to be linearly separable. Moreover, the SVM algorithm has been extended to what is called the “soft margin” SVM [25], that makes no assumption about the linear separability of the classes. Instead it normally functions as a typical SVM but in case the data are not linearly separable, it computes the hyperplane resulting in the lowest mis-classification rate, while it ensures the maximum margin between the closest correctly classified samples of the two classes. In order to understand the notion of the support vectors, the case of the simple SVM given linearly separable data is further explained.

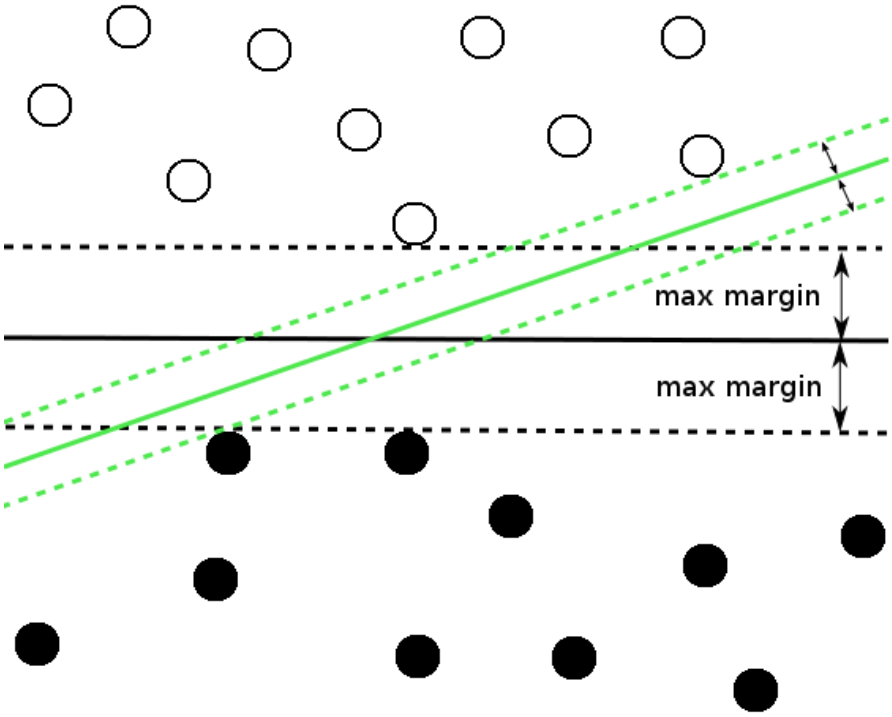

Figure 2.6 The black hyperplane separates the two classes, resulting in the maximum margin between their closest samples, and thus is selected as the SMV separating hyperplane.

Linear SVM

Given a dataset $D=\{(\boldsymbol{x_i}, y_i): \boldsymbol{x_i} \in R^P \; y_i \in \{-1,+1\}\}, i=1,...,N$ where $\boldsymbol{x_i}$ the samples and y_i the class labels, the goal of the SVM is to compute the hyperplane of dimension $R^{(P-1)}$ that separates all samples belonging to the class y=1 from those of y=-1, such as the margin of the closest samples of the two classes is maximized. If $\boldsymbol{x} \in R^P$ then any hyperplane can be expressed as $\boldsymbol{w} \cdot \boldsymbol{x} - b = 0$, where **w** the normal vector to the hyperplane and b a real constant. Then the parameter $\frac{b}{\|\boldsymbol{w}\|}$ expresses the offset of the hyperplane from the origin, along the normal vector **w**. Given that the data are linearly separable, there exist two hyperplanes $H_1: w \cdot x - b = 1$, $H_2: w \cdot x - b = -1$ that fully separate the two classes without any samples being misclassified. The region bounded by these two hyperplanes is called the "margin" between the two classes, which is equal to $\frac{2}{\|w\|}$. So in order to maximize the margin, $\|w\|$ needs to be minimized. While $\|w\|$ is minimized, samples of either class may appear inside the margin, for that to be avoided, further constraints need to be implemented: $\boldsymbol{w} \cdot \boldsymbol{x_i} - b \geq 1$ for samples of class $y_i = 1$ and $\boldsymbol{w} \cdot \boldsymbol{x_i} - b \leq -1$ for samples of class $y_i = -1$. Both constraints can be expressed in one equation as $y_i \cdot (\boldsymbol{w} \cdot \boldsymbol{x_i} - b) \geq 1$ for i=1, ..., N. The above can be expressed as an optimization problem:

Minimize in w,b

$\|\boldsymbol{w}\|$

subject to $y_i \cdot (\boldsymbol{w} \cdot \boldsymbol{x_i} - b) \geq 1$ **, for i=1,...,N**

or to avoid calculating the square root:

Minimize in w,b

$\frac{1}{2}\|\boldsymbol{w}\|^2$

subject to $y_i \cdot (\boldsymbol{w} \cdot \boldsymbol{x_i} - b) \geq 1$ **, for i=1,...,N**

By introducing the Lagrange multipliers α , the above can be expressed as a problem of quadratic programming:

$\min_{w,b} \max_{\alpha \geq 0} \{ \frac{1}{2} \|w\|^2 - \sum_{i=1}^{N} \alpha_i [y_i \cdot (w \cdot x_i - b) - 1] \}$ and then according to the stationary Karush-Kuhn-Tucker condition, the solution can be expressed as a linear combination of the training input vectors x_i : $w = \sum_{i=1}^{N} \alpha_i y_i x_i$.

Only a few of the Lagrange multipliers α_i are greater then zero. These multipliers correspond to the closest samples of the two classes, the support vectors, that lie on the margin and satisfy $y_i \cdot (w \cdot x_i - b) = 1$. Solving the previous equation for b we obtain $b = w \cdot x_i - y_i$ for a given support vector. In that manner, a more stable estimation of b is the mean value over all support vectors, given by the formula $\hat{b} = \frac{1}{N_{sv}} \sum_{i=1}^{N_{sv}} (w \cdot x_i - y_i)$.

Using the equations $\|w\| = w \cdot w$ and $w = \sum_{i=1}^{N} \alpha_i y_i x_i$ the optimization problem can be expressed in it's dual form as:

Maximize in α_i

$$L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

subject to $\alpha_i \geq 0$, $\sum_{i=1}^{N} \alpha_i \cdot y_i = 0$

where $K(x_i, x_j) = x_i \cdot x_j$ a kernel function.

After the Lagrange multipliers α_i have been computed, w can be determined using $w = \sum_{i=1}^{N} \alpha_i y_i x_i$.

The problem expressed in dual form is computationally efficient, since the classification task takes into consideration only the support vectors, which generally are a small subset of the original set of training samples.

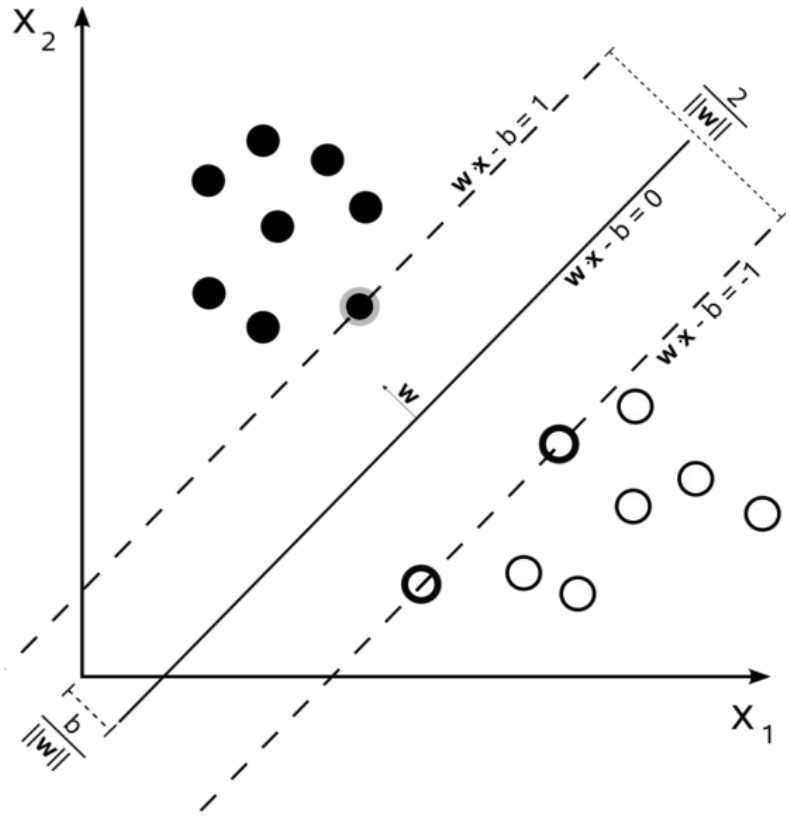

Figure 2.7 The separating hyperplane of a linear SVM.

2.4.4 K Nearest Neighbor (K-NN) Classifier

Nearest Neighbor methods such as K-NN [27] utilize a non-linear approach that classifies new samples depending on a set of samples closest to them, which are called their "nearest neighbors". Given a set of known training samples, K-NN classifies a new test sample depending on the class label of the majority of K samples nearest to it, according to a distance metric. Supposing that in the case of binary classification the dataset is $D=\{(\boldsymbol{x_n}, y_n): \boldsymbol{x_n} \in R^P \; y_n \in \{-1,+1\}\}, n=1,\dots,N$, then a new sample $\hat{\boldsymbol{x}}$ is given a class label $\hat{y}$ according to the formula $\hat{y}=sign(\sum_{i=1}^{K} \tilde{y}_i)$. Where $\tilde{y}_i$ the class label corresponding to the i-th nearest neighbor of $\hat{\boldsymbol{x}}$. In the case Euclidean distance is used, the nearest neighbor of $\hat{\boldsymbol{x}}$ is expressed as $\tilde{\boldsymbol{x}}=argmin\|\hat{\boldsymbol{x}}-\boldsymbol{x_i}\|, i=1,\dots,N-1$.

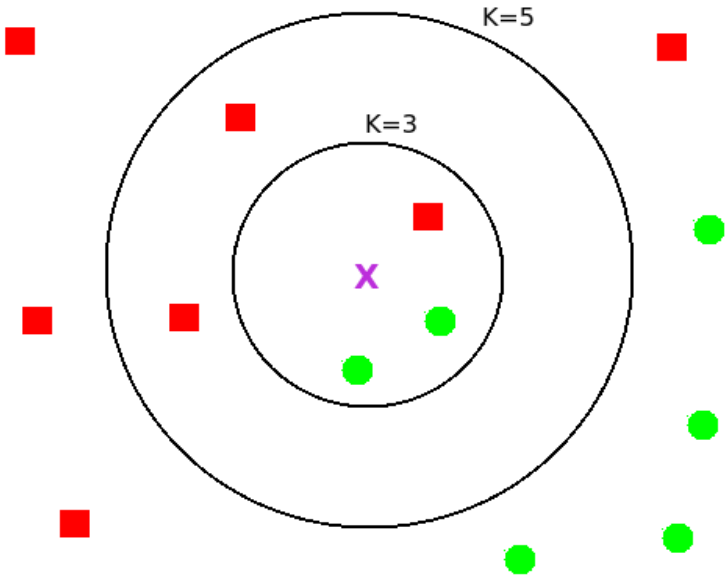

Figure 2.8 The test sample (purple X) will be classified in the first class of green circles in the case of K=3. However, in the case of K=5 it will be classified in the second class of red rectangles.

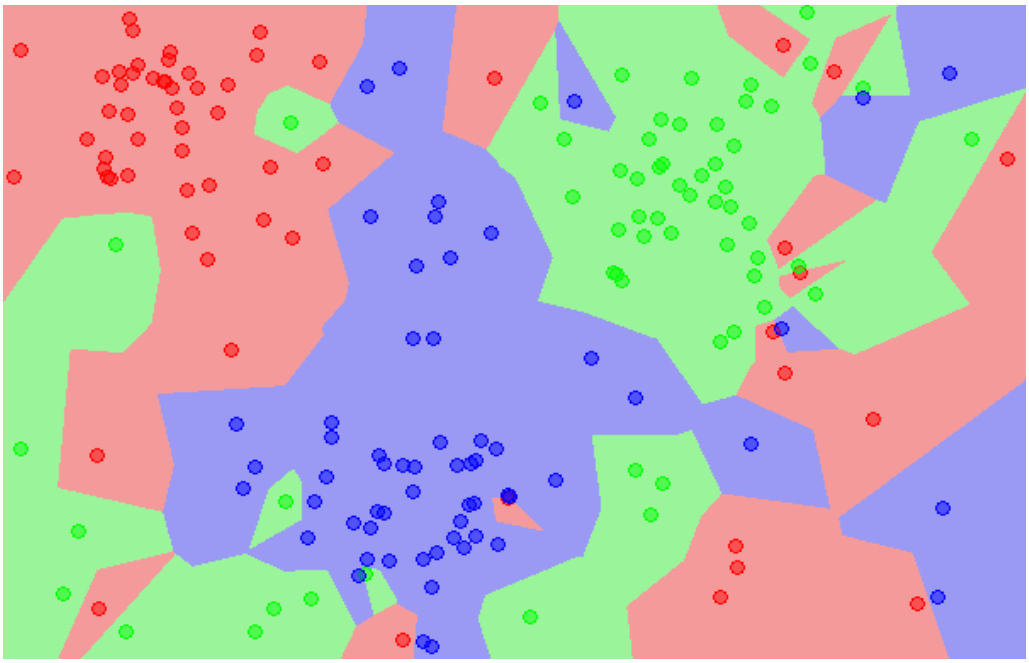

Figure 2.9 The class borders of an 1-NN classifier In the case of 3-way classification.

2.5 Evaluation Methods

Evaluation methods are used to estimate the ability of the model to generalize, that is to yield comparable results in unknown data as well data used during training. If all available data are used for training, there is no assessment of the FSS & classification performance on new data and as such, the generalization ability of the model remains unknown. In that manner, evaluation methods leave out a set of samples that are only used in order to assess the performance of the model on new data. That set of samples is called the test set, while the set of samples used while training the model is called the training set.

2.5.1 Holdout Validation

Holdout validation is probably the simplest validation method. It splits the available samples into two groups. The training set consists of the majority of available samples and is used for training the model while the test set corresponds to a smaller percentage of the available samples and is used in order to evaluate the model's generalization ability. However, excluding a portion of the dataset can be costly when the available samples are few. Moreover, the results obtained greatly depend on the random splitting of the dataset into training and test sets and the observed results can be misleading if both splits are do not reflect the structure of the original dataset. To counter these drawbacks of the simple holdout method at the expense of computational load, other validation techniques have been proposed.

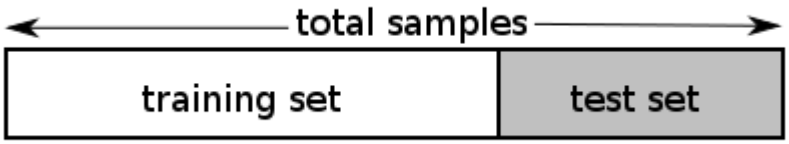

Figure 2.10 Holdout validation method.

2.5.2 K-Fold Cross Validation (K-Fold CV)

K-Fold Cross Validation splits the dataset into K different subsets of approximately the same size, called folds. It then proceeds to iteratively use k-1 folds for training and 1 fold for testing the FSS & Classification model, using a different fold for testing during each iteration. At the end of the procedure, k different test statistics have been observed. The average statistics over all folds are then calculated. If for example the only test statistic examined is the classification accuracy, it is calculated using the following formula: $\bar{a} = \frac{1}{K}\sum_{k=1}^{K} a_k$. Typical values used for k are K=3, 5 or 10. As the number of folds increases, the bias of the estimate decreases, so the estimation of performance is representative of the actual performance of the method. However, the variance of the estimation as well as the computational cost increase due to the large number of iterations. If the cross-validation method is “stratified”, then the class ratio is the same for all folds.

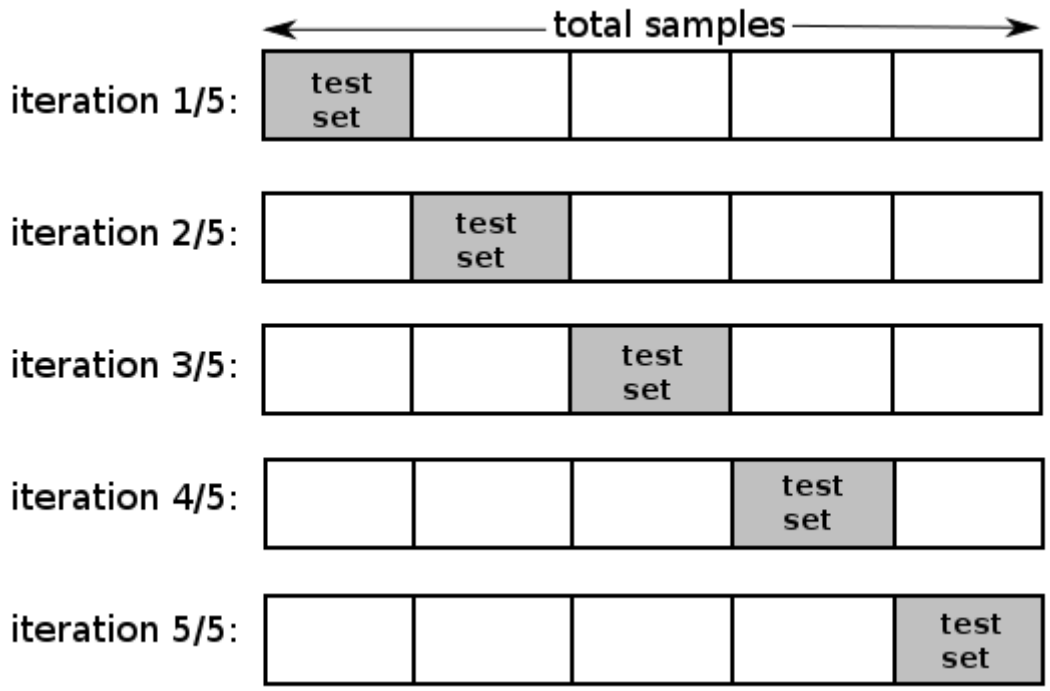

Figure 2.11 5-Fold Cross Validation.

2.5.3 Leave One Out Cross Validation (LOOCV)

Leave one out cross validation is a case of K-Fold CV where the number of folds K is equal to the number of samples in the dataset N. Since the number of samples is larger than the typical values of k used during simple K-Fold CV, LOOCV displays the characteristics of K-Fold CV when large K is utilized: small bias of the estimations accompanied by large variance of the test statistics as well as high computational cost.

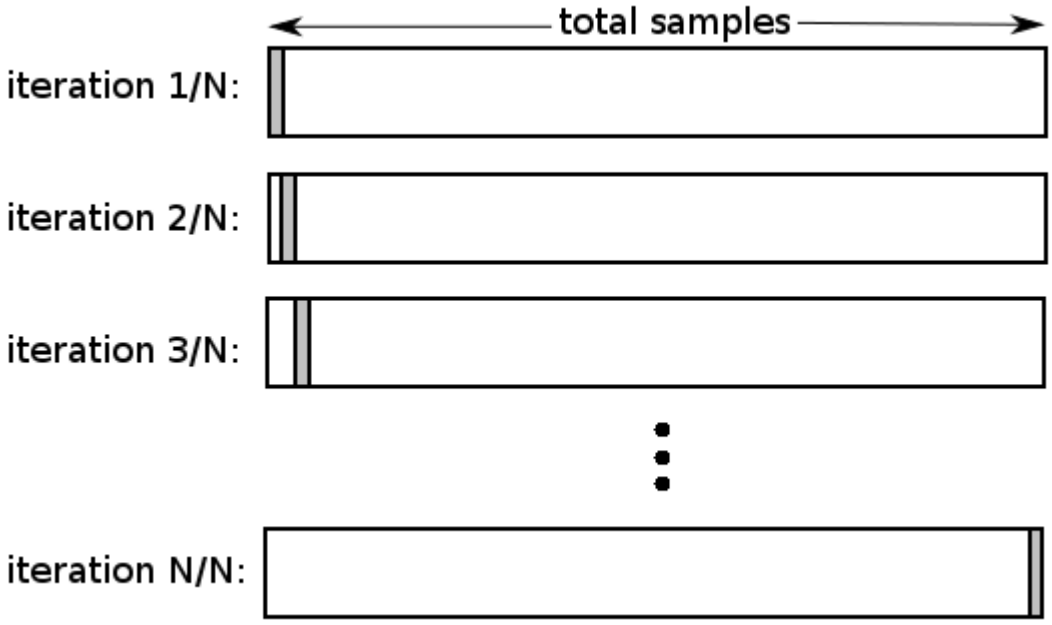

Figure 2.12 Leave One Out Cross Validation.

2.5.4 Repeated Random Sub-Sampling Validation

Repeated random sub-sampling validation is run for a fixed number of K iterations. During each iteration it utilized random sampling without replacement, in order to select a fixed number of S samples that make up the test set and are excluded from the training process of the model. The observed test statistics are then averaged over all iterations.

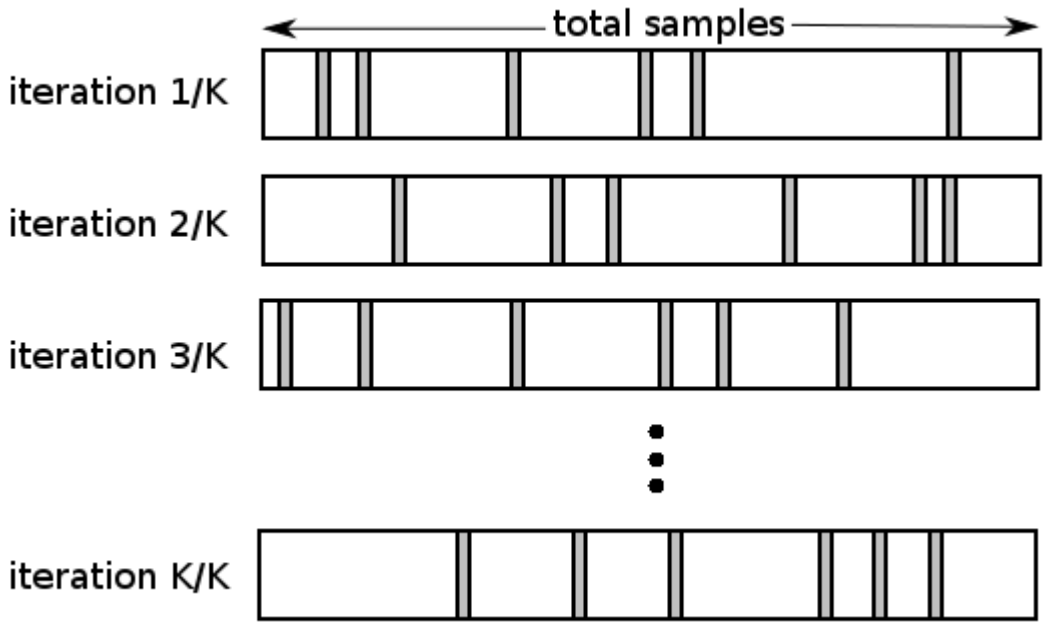

Figure 2.13 Repeated Random Sub-Sampling Validation.

2.5.5 Bootstrap Resampling Validation

Given an original dataset, bootstrap resampling, also called bootstrapping, utilizes random sampling with replacement in order to construct a number of B bootstrap datasets of fixed size, usually the same number of N samples as the original dataset. The class ratio in each dataset can either be random, or determined beforehand. Each bootstrap dataset can then be separated into training and test sets using the simple holdout method. The test statistics are then calculated for each bootstrap dataset and are averaged over all bootstrap datasets in order to get a stable estimation.

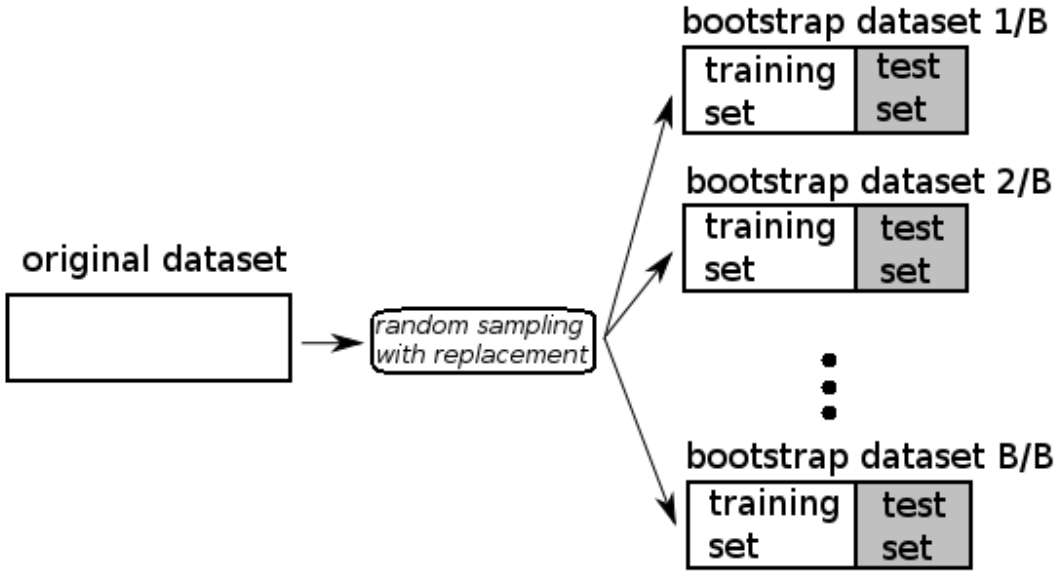

Figure 2.14 Bootstrap Resampling Validation.

2.6 Weak Law of Large Numbers

The weak law of large numbers (LLN) [29] [30] is a theorem of probability theory which states that given that a random experiment is executed a sufficiently "large" number of times, the mean value of the observed results will be close to the expected value, and will continue to converge as more experiments are performed. Stated formally, the theorem suggests that given a set of independent identically distributed (i.i.d) random variables $X_{1,}\dots,X_n$, each having a mean $\bar{X}_i=\mu$ and variance $var\left(X_i\right)=\sigma^2$.

A new random variable X can be defined, such as $X\equiv\frac{X_1+\dots+X_n}{n}$.

Then, as the number of trials $n\to\infty$: $\bar{X}=\frac{\overline{X_1+\dots+X_n}}{n}=\frac{\bar{X}_1+\dots+\bar{X}_n}{n}=\frac{n\cdot\mu}{n}=\mu$.

Moreover $var\left(X\right)=var\left(\frac{X_1+\dots+X_n}{n}\right)=var\left(\frac{X_1}{n}\right)+\dots+var\left(\frac{X_n}{n}\right)=\frac{\sigma^2}{n^2}+\dots+\frac{\sigma^2}{n^2}=n\cdot\left(\frac{\sigma^2}{n^2}\right)=\frac{\sigma^2}{n}$

and by the Chebyshev inequality, for all ε>0:

$P\left(\left|X-\mu\right|\geq\varepsilon\right)=var\frac{\left(X\right)}{\varepsilon^2}=\frac{\sigma^2}{n\cdot\varepsilon^2}$ and for $n\to\infty$: $\lim_{n\to\infty}P\left(\left|X-\mu\right|\geq\varepsilon\right)=0$

For example, let $X_{1,}\dots,X_n$ be the results of rolling a 6-sided die. Then each roll produces a result between that is one of the numbers 1, 2, 3, 4, 5 or 6 with equal probability. Then the expected value of the die roll is $\frac{1+2+3+4+5+6}{6}=3.5$. So, according to the weak law of large numbers, given a large enough number of repetitions, the average value of die rolls should converge towards 3.5. The results of such an experiment are shown in figures 2.6a and 2.6b.

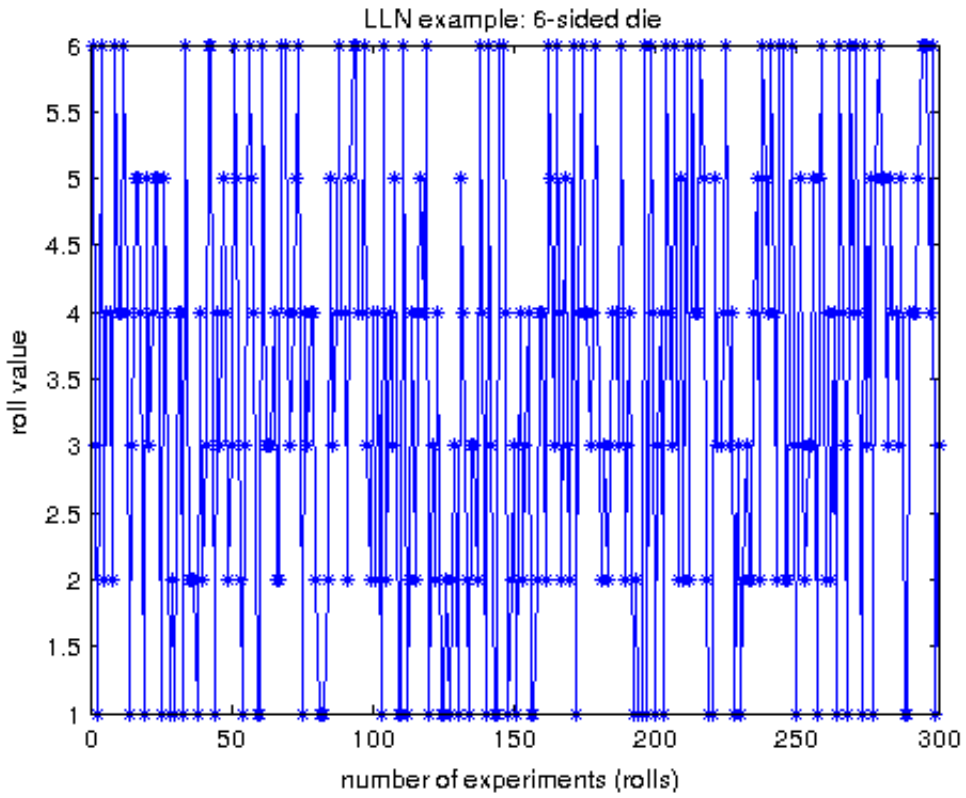

Figure 2.15 Instantaneous values of the 300 rolls of a 6-sided die.

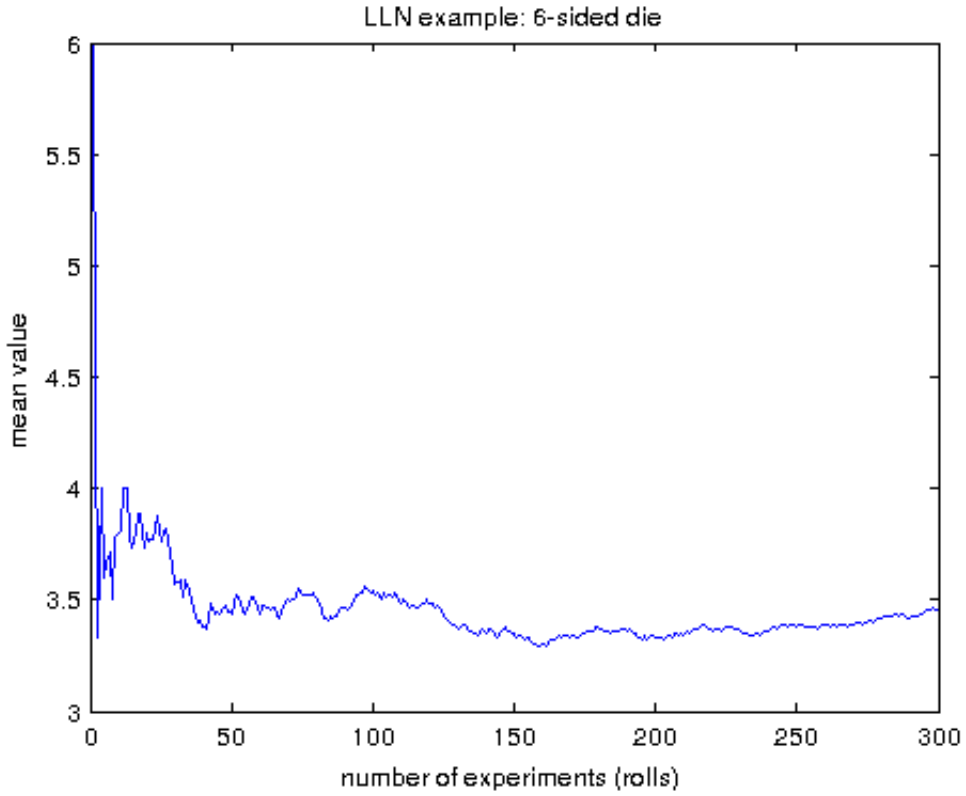

Figure 2.16 Demonstration of the law of large numbers: the mean value over all rolls converges towards 3.5, the expected value of the experiment, as more repetitions of the experiment take place.

The law of large numbers can be utilized in order to assess the stability of results in genomic datasets. First, bootstrap resampling can be used to generate a large number of datasets to be used for the evaluation of feature selection and classification methods. Then, under the assumption that the observed results are independent identically distributed random variables, the law of large numbers can guarantee the stability of the average estimates given that the sample size is sufficiently large. Thus, the average estimates can be used as a measure of stability. In order to determine when the sample size is large enough and no more bootstrap datasets are required, an explicit criterion determining the stability of results can be used. The use of bootstrap resampling, in conjunction with the law of large numbers and a stability criterion constitute the concept behind the stable evaluation methodology proposed in this thesis.

3 - Methodology

3.1 Methodology Overview

This section is aimed at proposing a methodology for performing stable and robust feature selection and classification, while evaluating the statistical significance and consistency of the observed results. Stability of performance assessments is an important aspect of microarray analysis, since slight variations in the training or testing data can lead to significant variations in the set of features selected, as well as the observed classification accuracy. Due to to that sensitivity to the training and test data variations, known and widely used methods such as K-Fold CV result in performance assessments of FSS & classification methods that vary between different executions of the evaluation method. To address that issue, methodologies that utilize repeated resampling or splitting of the original dataset have been proposed, in order to extract stable performance estimates. Davis et al. in [1] notice that after a sufficiently large number of datasets have been generated by random splitting of the original dataset and are used to extract performance estimates, the average value of the classification accuracy tends to stabilize. However, “sufficiently large” is a subjective term that can vary between different FSS & classification methods. One approach, as adopted in StabPerf [1] is to run the bootstrap validation scheme for a fixed large number of repetitions, for example 400, plot the observed results and manually assess their stability for each test. Likewise, the performance assessment model proposed by Suzuki et al. in [3] performs LOOCV on an arbitrary large number of 100 bootstrap datasets in order to get a reliable LOOCV accuracy estimates. The framework introduced by Armañanzas et al. [6] requires and arbitrary number of 1000 bootstrap iterations followed by univariate filtering and training a k Dependence Bayesian classifier, in order to result in a stable set of genes selected in the model. In a similar manner, the Monte Carlo CV methodology of Barrier et al. [8] results in 1600 datasets produced by splitting the original dataset into training and test set of various sizes and are used for univariate filtering and training a diagonal linear discriminant analysis classifier. However, FSS & classification methods are generally computationally intensive, so running an unnecessarily large number of evaluation iterations, just to be sure that statistics will be stable, can prove to be impractical.

Moreover, while the methodology of Armañanzas et al. [6] results in a stable set of genes included in the model, the accuracy estimate is extracted by standard cross validation techniques that do not guarantee the stability of the observed classification accuracy: 5-fold CV in [6], or LOOCV in [7]. Another issue is that while many methods select reliable genes in the genomic signature, the size of the signature itself is not a stable quantity, but depends on the number of genes that surpass and arbitrary threshold [6] [7] [8].

To address the above issues, in section 3.2 a methodology called “Stable Bootstrap Validation” is proposed, that utilizes a formal criterion for stability that determines when a sufficient level of stability is reached for the resulting genomic signature as well as the observed classification accuracy and no further iterations are required. The stable estimates can be reproduced resulting in minimal variations during independent executions of the evaluation method. Thus, stable estimates lead to a greater degree of generalization of results.

Meanwhile, section 3.3 introduces a methodology for assessing the statistical significance of the observed results. It is used to determine whether the results are observed merely by chance or reflect the underlying biological model. The statistical significance of the classification accuracy is assessed by calculating it's corresponding p-value using permutation tests. If the observed accuracy is significant and it is observed due to the correlation between the gene expression levels and the class of the sample, then the corresponding p-value should be lower than the 0.05 threshold. As for the genomic signature, it's relevance to the biological model is assessed by comparing it's performance to signatures of the same size, whose genes have been selected at random. If the genomic signature reflects the biological model, then it should considerably outperform random signatures.

Section 3.4 introduces a methodology concerning the assessment of consistency regarding the observed classification performance of a genomic signature. If a classification method is consistent, it is should lead to considerable repeatability of results. That is, it should yield similar results on the same test set across different iterations. To assess the consistency of a classifier, a single bootstrap test dataset is generated, while a number of different bootstrap datasets are used in order to train the classifier.

The overview of the proposed framework is presented as a block diagram in figure 3.1. This forms an iterative approach, but it is preceded by a filter methodology as the preprocessing step applied to the dataset in order to eliminate the less descriptive features. In this form, our framework forms a two-step selection approach, exploiting the benefits of both univariate and multivariate techniques.

Preprocessing of the Dataset: Two-step FSS

The dataset provided for this study has been preprocessed. That is, it has undergone feature subset selection using a filter (univariate) method. The filter method used during the preprocessing of the dataset is called “Significance Analysis of Microarrays” (SAM) [31] and employs a modified t-statistic and repeated permutations of the data to determine if the expression of genes is strongly related to the response. On the other hand, during stable bootstrap validation the multivariate feature subset selection method of recursive feature elimination is implemented. By utilizing both univariate and multivariate methods in what can be called a “2-step FSS” the resulting signatures are expected to harness the advantages of both schemes. That is, genomic signatures extracted should be small in size and provide insight into class differences (univariate FSS) while also be tuned to optimize the classification performance of the specific classifier used (multivariate FSS).

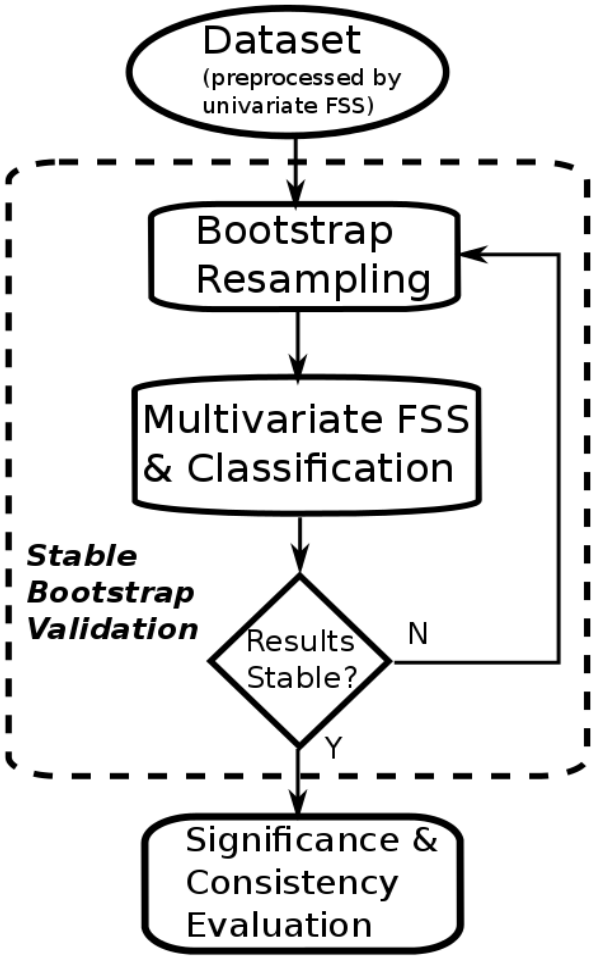

Figure 3.1 Overview of the proposed methodology. The dataset, preprocessed using univariate FSS is used as an input to stable bootstrap validation for the extraction of stable FSS and classification performance estimates. The significance evaluation of the results of stable bootstrap validation is then assessed.

3.2 Stable Bootstrap Validation (SBV)

Given a pair of FSS and classification methods, SBV aims at using a large number of datasets generated from bootstrap resampling of the original dataset, in order to extract stable estimates for the classification accuracy, as well as the size of the genomic signature. If the FSS and classification methods are evaluated on a large enough number of bootstrap datasets, then according to LLN the average estimates for the classification accuracy and the size of the genomic signature will be stable. To ensure that no more bootstrap datasets than necessary are generated, SBV utilizes an explicit dual criterion that determines

whether stability has been reached for the average classification accuracy, as well as the signature size. The criterion assesses the stability of results over consecutive batches of bootstrap datasets and determines whether a desired level of stability has been reached, or generating another batch of datasets is required. Unlike similar methodologies which lack a stability criterion and are executed for an arbitrary number of iterations, SBV is only executed until the necessary level of stability is reached. As such, SBV is a more computationally efficient methodology. Moreover, similar methods tend to extract stable estimates only for the classification accuracy while selecting an arbitrary number of genes or only a number of genes surpassing a selection frequency threshold. On the other hand, SBV also leads to a stable estimate concerning the number of genes in the signature, also called the genomic signature size and then proceeds to select the genes with the highest selection frequency over all iterations of the method.

The SBV procedure proceeds as follows. First, the batch datasets called the "bootstrap window" B is defined as a fixed number of bootstrap datasets. Then, a number of 3B bootstrap datasets are generated from the original dataset by random sampling with replacement. The size of the bootstrap datasets is arbitrary, however in most cases it is selected to be the same as the size of the original dataset. The class ratio is also arbitrary and typical values include the same class ratio as in the original dataset, or equal class ratio for all classes. The FSS & classification method is then executed 3B times, resulting in values $A_{1,}\dots,A_{3B}$ for the classification accuracy and $G_{1,}\dots,G_{3B}$ for the number of features selected, also called the genomic signature size. Assuming that A_i and G_i are sets of independent identically distributed (i.i.d) random variables, then according to the weak law of large numbers the average values over all samples $\bar{A}$ and $\bar{G}$ should converge towards the expected value of the classification accuracy and the genomic signature size, respectively. Next, the stability of the observed results is assessed in batches of subsequent B trials. Let $A_{wi}, i=1, 2, 3$ be the mean accuracies at the end of the fist, second and third bootstrap window, respectively. Then, the maximum difference of mean accuracy between windows 1, 2 and 1, 3 is defined as $\Delta A=max\left(\left|A_{w1}-A_{w2}\right|,\left|A_{w1}-A_{w3}\right|\right)$. The method uses three windows instead of two, in order to overcome the risk of local stability. Given that the mean accuracy differences between the three last windows have been calculated, stabilization is assessed. The classification accuracy estimate $\bar{A}$ is considered stable if $\Delta A<acc_{thresh}$, where acc_{thresh} a fixed threshold. In a similar manner, let $G_{wi}, i=1, 2, 3$ be the genomic signature sizes at the end of the fist, second and third bootstrap window, while $\Delta G=max\left(\left|G_{w1}-G_{w2}\right|,\left|G_{w1}-G_{w3}\right|\right)$ the maximum difference of mean signature size between windows 1, 2 and 1, 3. However, there is a significant difference when assessing the stability of signature size. While classification accuracy of all methods varies in [0,1], different FSS methods can lead to genomic signatures whose size differs in orders of magnitude. For this reason the corresponding threshold for the signature size is normalized by the largest signature size and is defined as $gen_{thresh}=\frac{\left|G_{wi}-G_{wj}\right|}{max\left(G_{wi},G_{wj}\right)}$, where i, j the windows being compared.

If both $\bar{A}$ and $\bar{G}$ are found to be stable, the SBV procedure ends. Otherwise, another set of B datasets is generated and the stability assessment is performed again for the 3 windows, which now extend to cover the additional datasets. The above steps are repeated until stability for the classification accuracy as well as the signature size is reached. During each iteration, the following formula applies for the mean accuracy of a given window:

$$A_{w_j}^{(n)}=\frac{1}{(n+j-1)B}\sum_{b=1}^{(n+j-1)}acc_b$$

While the same formula applies for the mean signature size:

$$G_{w_j}^{(n)}=\frac{1}{(n+j-1)B}\sum_{b=1}^{(n+j-1)}gen_b$$

Where n is the iteration number, j is the window being checked (1, 2 or 3), b runs all the bootstrap datasets, acc is the accuracy achieved by the classification method in each dataset and gen the number of genes selected (signature size).

After the SBV procedure has been completed, $\bar{A}$ is considered to be the stable assessment of classification performance, while $\bar{G}$ is the stable assessment of the genomic signature extracted by the FSS method. Finally, the $\bar{G}$ genes with the highest selection frequency across all bootstrap datasets are selected as the genomic signature of the specific combination of FSS & classification methods. A flowchart of SBV if shown in figure 3.2a, while an instantiation of the bootstrap process is illustrated in figure 3.2b.

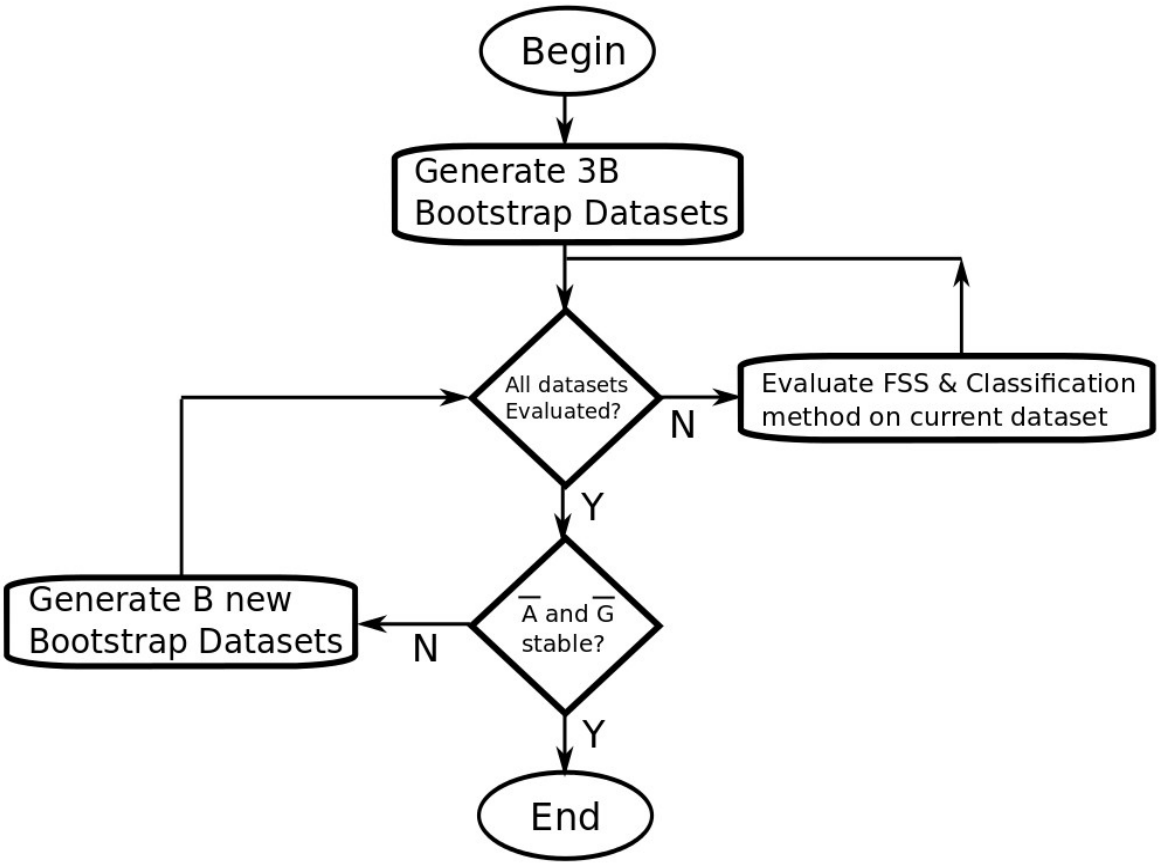

Figure 3.2 Flowchart of the SBV method.

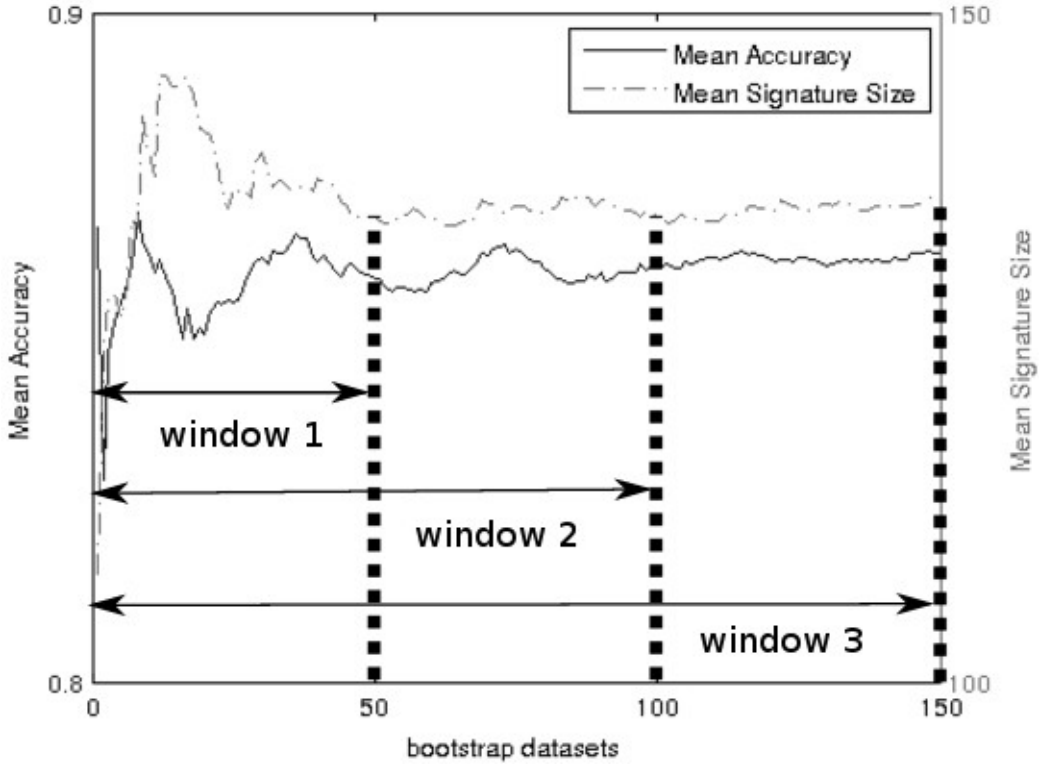

Figure 3.3 Graphical representation of the SBV bootstrap windows and stabilization of mean accuracy as well as mean signature size values. The left horizontal axis refers to the classification accuracy, while the right horizontal axis refers to genomic signature size. The bootstrap window size B has been set to 50. Performance assessment was stabilized within the first 3B datasets, so no additional extensions were required.

3.3 Statistical Significance Evaluation

After the completion of SBV the stable values that have been extracted are the classification accuracy $\bar{A}$ and signature size $\bar{G}$. Another aspect of microarray analysis is the statistical significance of the observed results, which include the classification accuracy, as well as the genomic signature. In other words, it is necessary to determine whether the performance assessment reflects the underlying biological model, or it is observed by chance due to random noise. In that manner, permutation tests are performed to determine the statistical significance of the classification accuracy. On the other hand, the significance of the observed genomic signature is determined by comparing it's performance to that of random signatures of the same size.

First statistical significance of the observed classification accuracy $\bar{A}$ needs to be assessed. The aim is to infer to what extent the classification accuracy is achieved due to the correlation between the gene expression levels of the samples and the class labels. In that manner, a classical method of hypothesis testing is performed. Let the null hypothesis H_0 be that the random variables $\boldsymbol{x_n}$ (samples) and y_n (class labels) are independent. To evaluate the p-value corresponding to $\bar{A}$ the probability density function of the classification accuracy is required. Since it is unknown, non parametric permutation tests are performed in order to estimate the empirical probability density function and calculate the corresponding p-value [32]. More specifically, a fixed number of 1000 bootstrap datasets are generated and the labels of the two classes are permuted. Then, the FSS & Classification method is performed on the permuted datasets. Given and observed classification accuracy $\bar{A}$ it's corresponding p-value is defined as the number of times that accuracy “greater or equal to” $\bar{A}$ was observed when a permuted dataset was used for training the model, divided by the number of permuted bootstrap datasets. If the p-value calculated is less than 0.05, then the observed $\bar{A}$ is considered to be statistically significant.

In order to test the significance of the extracted genomic signature, a bootstrap resampling approach is followed as well. However, unlike the case of classification accuracy, permutation tests are not performed. Instead, the performance of the extracted signature is compared to that of random signatures of the same size. The significance testing proceeds as follows. A fixed number of 1000 bootstrap datasets are generated. For each dataset, the FSS method is omitted and the classifier is trained on a bootstrap dataset given a random signature of size $\bar{G}$. That is, only $\bar{G}$ random genes remain in the dataset, while all others have been eliminated. Testing the classifier results into two performance metrics. First, the mean accuracy across all random gene datasets. Second, Prs: the percentage of times that equal or greater accuracy than $\bar{A}$ was observed when the model was trained using a random signature. Prs is not a p-value since no permutations of labels were performed.

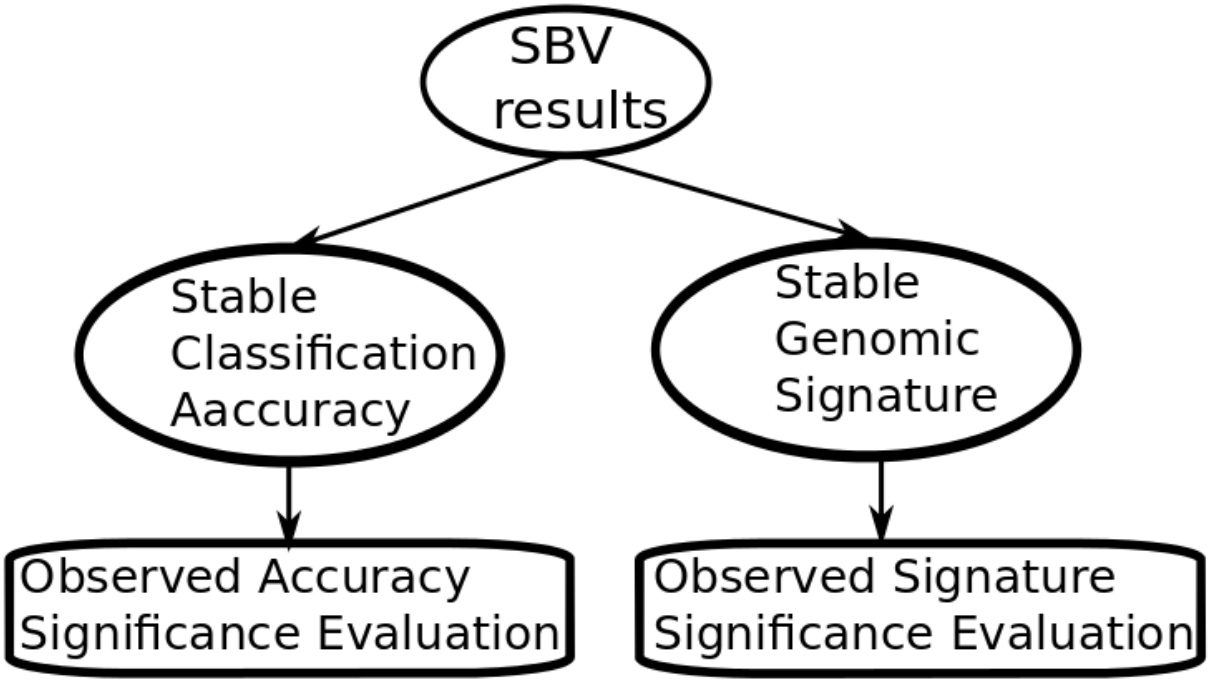

Figure 3.4 Flowchart of the significance evaluation methodology

3.4 Consistency Evaluation of Signature Classification Accuracy

The consistency of a classification method is the ability to yield similar performance when applied on the same test set multiple times, while using different training sets [59]. The consistency of a classification method can either be assessed given a fixed set of selected features, or the FSS method can be executed during each iteration to extract a feature set. In this thesis, the fixed signature scenario was implemented. To assess the consistency of the classification accuracy of the extracted signature, a bootstrap resampling scheme was implemented in order to create a large number of datasets for training a classifier, while only a single bootstrap dataset is used for testing. If the method is consistent, then there should be small standard deviation of the classification accuracy achieved on the test set, using all different bootstrap training datasets. A fixed number of 30 bootstrap training datasets were generated while only one bootstrap test dataset was generated for each signature. As such, each classifier produced 30 different classification accuracies on the same test set, depending on the training set used. The average observed accuracy as well as the corresponding variance and standard deviation are then extracted. The above procedure is repeated a total of 100 iterations and the results are averaged to produce a more stable estimation of the consistency of observed accuracy.

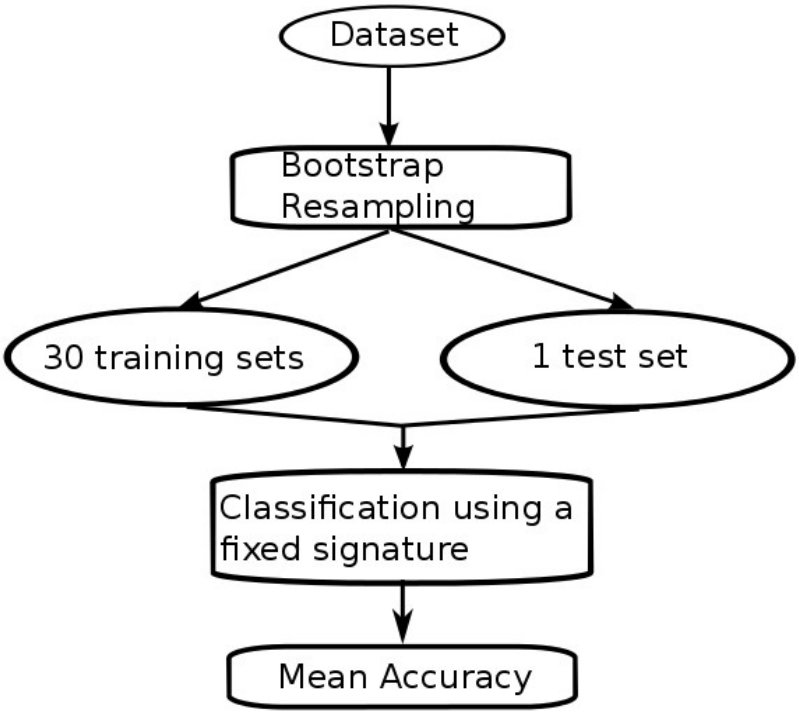

Figure 3.5 Flowchart corresponding to one iteration of the consistency evaluation methodology. The process is repeated 100 times and the results over all iterations are averaged.

4 - Results

The performance metrics of different FSS & classification methods extracted by SBV are presented and compared in section 4.1. Results are stabilized and the plots of observed stability prove that the original assumption that the classification accuracy as the signature size are i.i.d random variables. The embedded FSS method is RFE while the methods used for classification are the RLS classifiers RR and LASSO, the PLS Classifiers PLS-BETA and PLS-VIP, a linear SVM, I-RELIEF feature weighting in conjunction with the K-NN classifier, as well as PLS feature weighting in conjunction with the K-NN classifier. To be precise, while I-RELIEF was proved to be too computationally expensive for the high dimensionality dataset used, the K-NN classifier performed well. That lead to the idea of using the K-NN classifier but adopting a different feature weighting scheme which proved to be fast and efficient: PLS feature weighting. The statistical significance of the above SBV results is then assessed in section 4.2. The performance assessment of SBV are compared to those extracted using 10-Fold CV in section 4.3. Finally, the biological evaluation of the signatures extracted using SBV is included in section 4.4

Original Dataset

As mentioned in section 3, the dataset provided for this study has been preprocessed by a univariate FSS method called “Significance Analysis of Microarrays” (SAM) [31]. The implementation of both univariate (as preprocessing) and multivariate (during SBV) FSS aims at harnessing the advantages of both FSS methodologies, resulting in small sets of features that discriminate between classes of interest and lead to good classification performance. The original dataset consists of 529 samples related to breast cancer, 104 of which correspond to non-cancerous control and 425 to cancer samples and is produced by the integration of 5 publicly available datasets (GEO access numbers: 22820, 19783, 31364, 9574, 18672). For each sample, there are measurements of all remaining 4174 genes after the first step of univariate FSS using the SAM algorithm.

4.1 SBV Results

The SBV methodology proceeds to evaluate the classification accuracy and genomic signature extracted from a pair of FSS and classification methods, on batches of bootstrap datasets. Each batch of bootstrap datasets has the same size, called the “bootstrap window” B. After the method is run for a sufficient number of bootstrap windows and stability has been reached according to the criterion for classification accuracy and signature size, the SBV procedure terminates and returns the stable performance estimates.

SBV Parameters - Bootstrap Dataset Structure

The bootstrap window B of SBV was set to 50 bootstrap datasets, the accuracy threshold acc_{thresh} was set to 0.01 and the signature size threshold gen_{thresh} was set to 0.05. Each bootstrap dataset has the same size as the original dataset and is split into a training (90%) and a test set (10%). Moreover, each of the training and test sets has the same cancer/control ratio as the original dataset (4 to 1). The SBV method was set to stop if no convergence had taken place at 1000 bootstrap datasets, a scenario that never took place as all methods converged at most 200 bootstrap datasets, half of what was required in StabPerf [1].

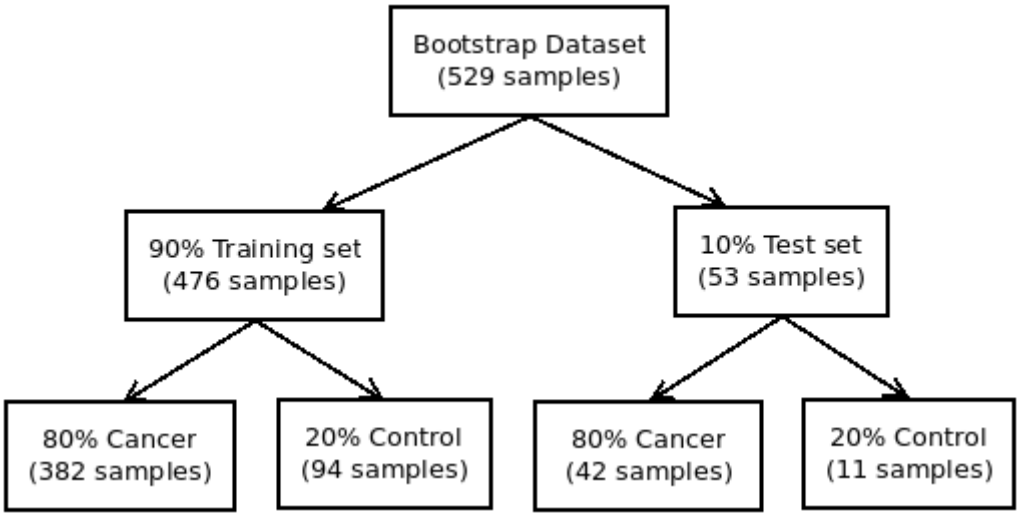

Figure 4.1 Structure of the bootstrap datasets used.

4.1.1 RLS Classifiers

Estimation of the t threshold

As mentioned in the theoretical background, the regularization threshold t was expressed as $t=\alpha\cdot\sum_{p=1}^{P} w_{0_i}^2, \alpha\in[0,1]$ and estimated using 3 different executions 10-Fold CV on the original dataset. The value of α=0.3 proved to be best for classification performance of both the RR and LASSO methods.

RFE & Ridge Regression

Ridge Regression achieves good classification accuracy, however it tends to keep a very large number of features, resulting in a genomic signature of large size, which is difficult to assess biologically.
It requires a moderate, yet reasonable, amount of running time.

Classification Accuracy	Genomic Signature Size	Time per bootstrap dataset (seconds)
88%	1372	142.95

Table 4.1 SBV results of RR

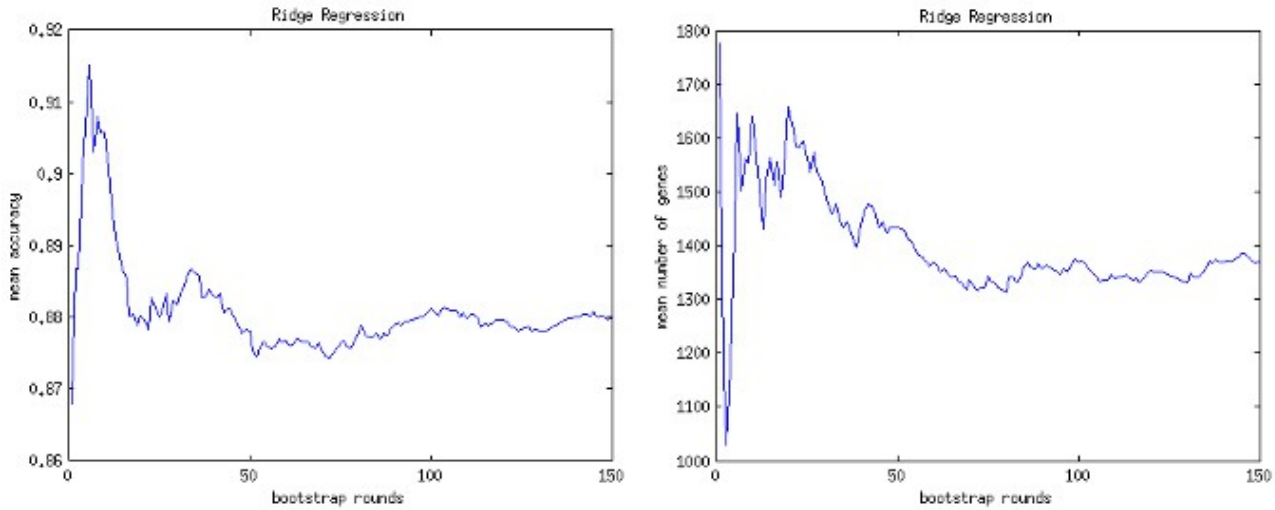

Figure 4.2 Left: Stabilization of RR mean accuracy over all bootstrap datasets
Right: Stabilization of RR mean signature size over all bootstrap datasets

RFE & LASSO

LASSO regression achieves similar classification accuracy as well as running time to RR, while the resulting genomic signature is considerably (an order of magnitude) smaller in size, leading to a more easily interpretable model. RFE was implemented in conjunction with the embedded feature selection of the LASSO.

Classification Accuracy	Genomic Signature Size	Time per bootstrap dataset (seconds)
86.4%	136	118.52

Table 4.2 SBV results of LASSO

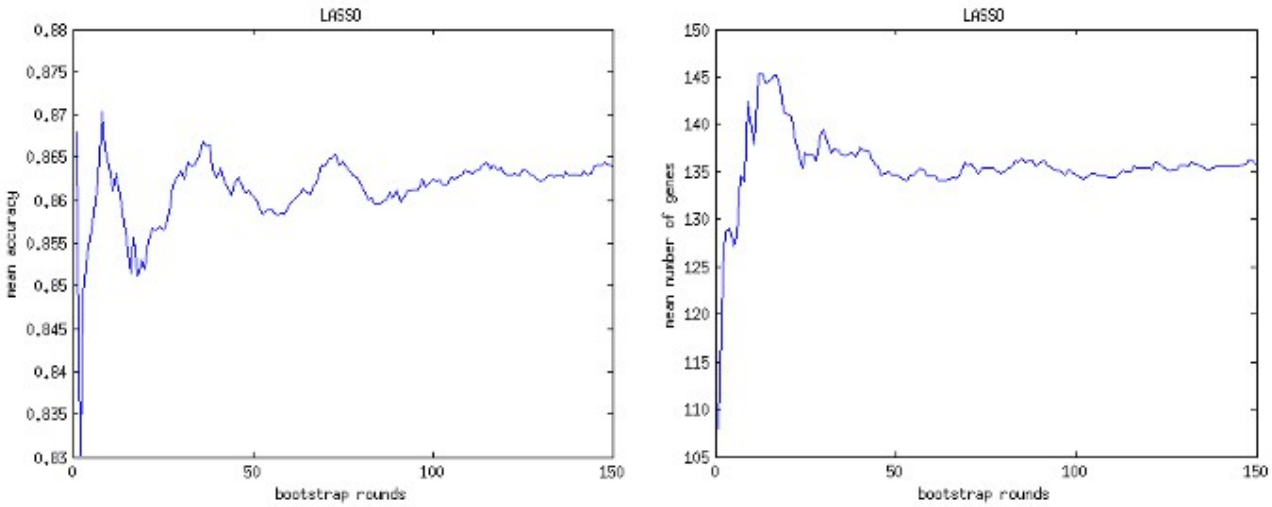

Figure 4.3 Left: Stabilization of LASSO mean accuracy over all bootstrap datasets
Right: Stabilization of LASSO mean signature size over all bootstrap datasets

4.1.2 PLS Classifiers

Estimation of the number of latent variables h

The number of latent variables, also called principal components, was estimated using 10-fold CV on the original dataset. The value h=2 proved to maximize the classification of both PLS-VIP and PLS-BETA methods. This is to be expected, since a small number of principal components are used in practice during PCA, otherwise a considerable amount of noise is embedded in the structural part of PCA, the matrices of scores T and loadings P.

Tuning the VIP score

The performance of the PLS-VIP method is strongly correlated to the choice of the cut-off value for the VIP score used for the selecting variables. The cut-off value used commonly in literature is the average VIP>1. However, in this study is observed that increasing the cut-off value results in considerably smaller sizes of genomic signatures, while the loss of classification accuracy is relatively small. The VIP score was also used in PLS-BETA to determine the initial number of features kept. That is, the number of selected features was selected according to the count of features surpassing a VIP threshold. However, the specific features were selected according to their corresponding value of $\boldsymbol{b}_{pls}$, as determined by PLS-BETA.

RFE & PLS-VIP

The PLS-VIP classification method is very fast, requiring up to just 7 seconds in the worst case scenario of VIP>1. The classification accuracy is good, and slightly decreases as the VIP threshold increases. However, increasing the VIP threshold results in signatures that are smaller in orders of

magnitude. While, the signature selected for VIP>1 is large, the signatures corresponding to the VIP>1.5 or VIP>2 criteria are considerably smaller. The noticeable change in signature size is reflected by the selection frequency of genes, which is shown in figure 4.1.2a.

VIP score	Classification Accuracy	Genomic Signature Size	Time per bootstrap dataset (seconds)
> 1	87.6%	825	6.9
> 1.5	83.6%	88	1.4
> 2	82.2%	18	0.87

Table 4.3 SBV results of PLS-VIP

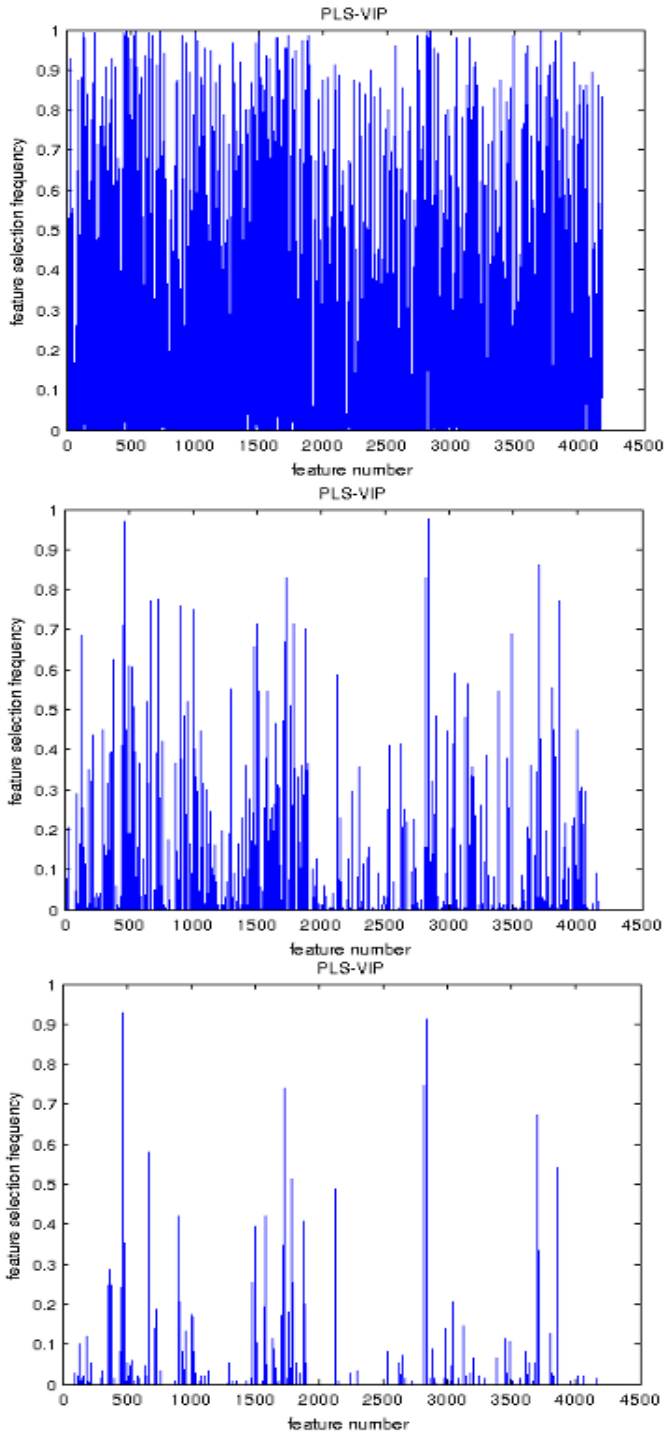

Figure 4.4 PLS-VIP gene selection frequency histograms, in the case of VIP>1, VIP>1.5 and VIP>2, respectively. Only the most significant genes are frequently selected as the VIP cut-off value increases.

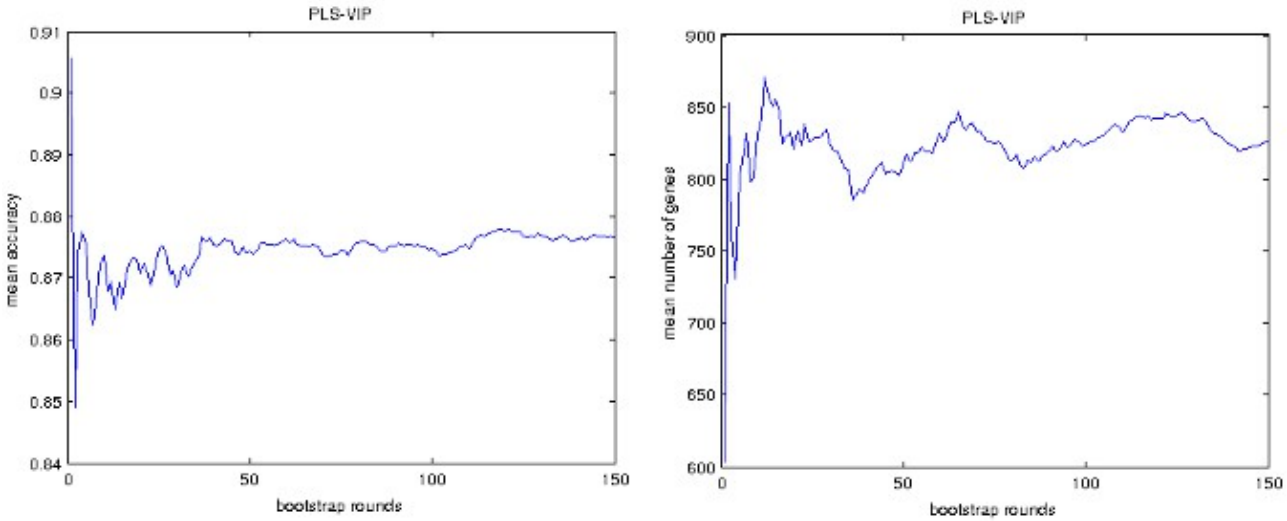

Figure 4.5 Left: Stabilization of PLS-VIP mean accuracy over all bootstrap datasets, when VIP>1. Right: Stabilization of PLS-VIP mean signature size over all bootstrap datasets, when VIP>1.

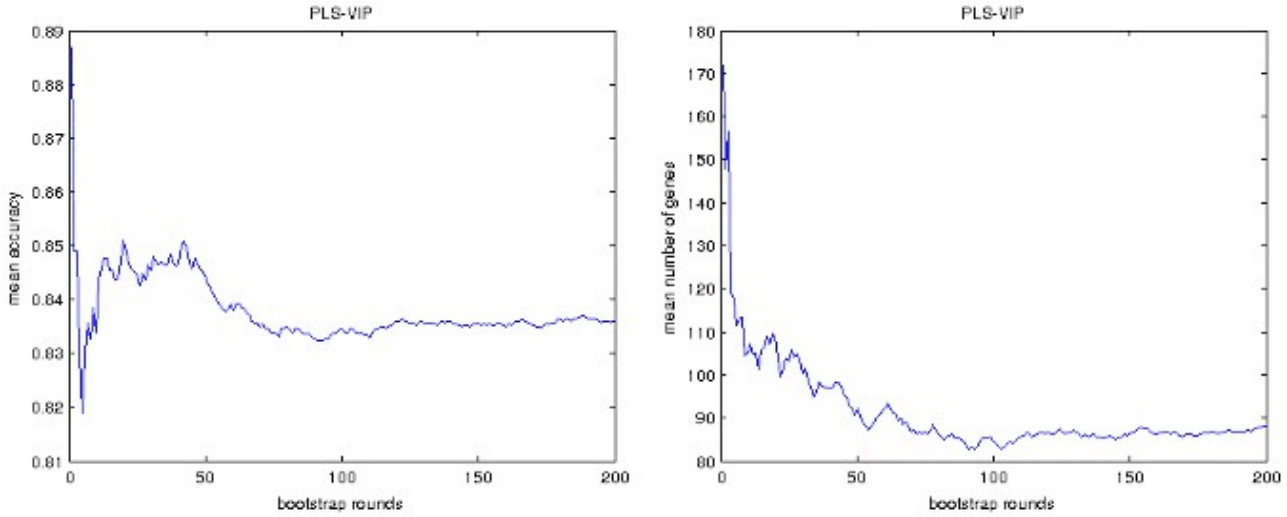

Figure 4.6 Left: Stabilization of PLS-VIP mean accuracy over all bootstrap datasets, when VIP>1.5. Right: Stabilization of PLS-VIP mean signature size over all bootstrap datasets, when VIP>1.5.

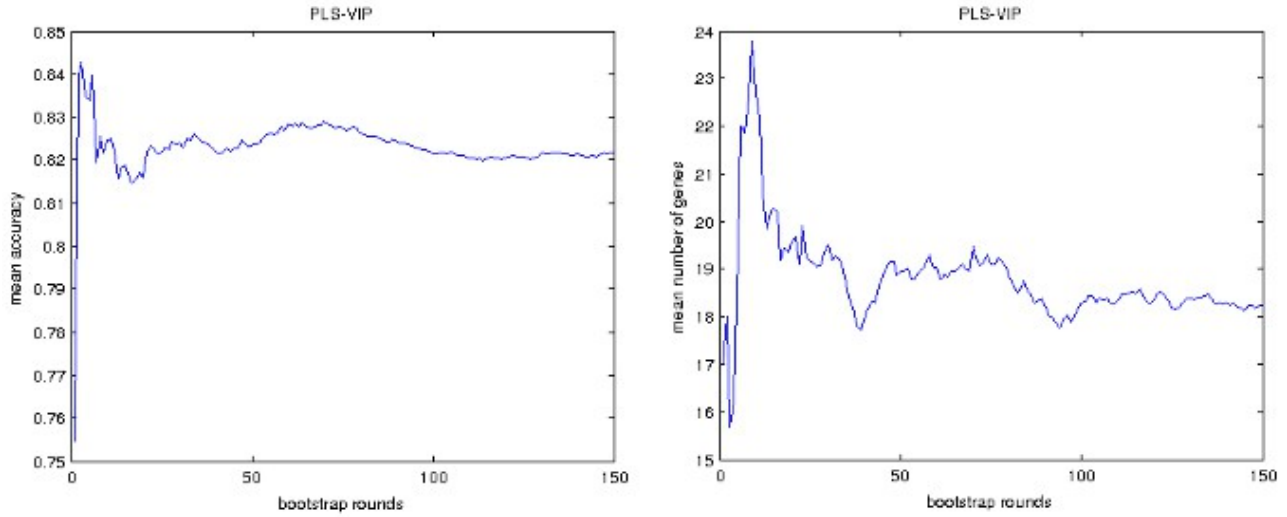

Figure 4.7 Left: Stabilization of PLS-VIP mean accuracy over all bootstrap datasets, when VIP>2. Right: Stabilization of PLS-VIP mean signature size over all bootstrap datasets, when VIP>2.

RFE & PLS-BETA

The PLS-BETA classification method performs similarly to PLS-VIP. In that manner, it is very computationally efficient, has good classification accuracy and tends to select small sets of features when the VIP threshold is set to 1.5 or more. Like the case of PLS-VIP change in signature size is reflected by the selection frequency of genes, which is shown in figure 4.1.2h.

VIP score	Classification Accuracy	Genomic Signature Size	Time per bootstrap dataset (seconds)
> 1	88.8%	1159	7.04
> 1.5	82.1%	92	1.76
> 2	81.2%	16	1.2

Table 4.4 SBV results of PLS-BETA

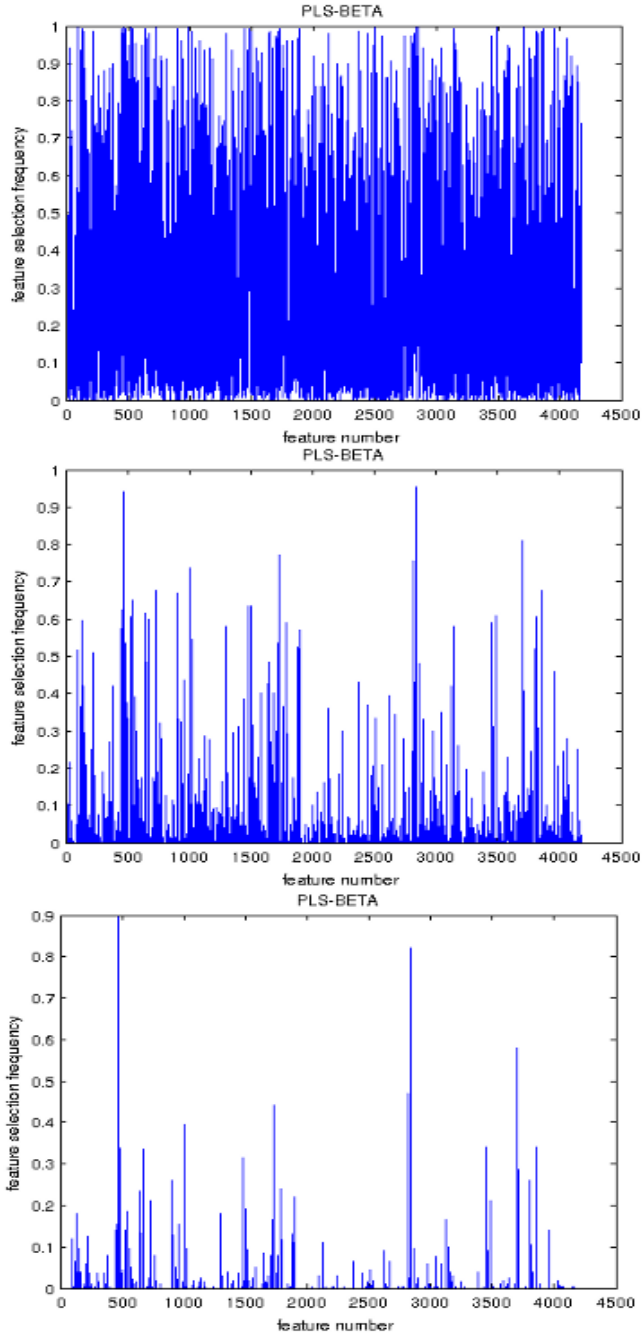

Figure 4.8 PLS-BETA gene selection frequency histograms, in the case of VIP>1, VIP>1.5 and VIP>2, respectively. Only the most significant genes are frequently selected as the VIP cut-off value increases.

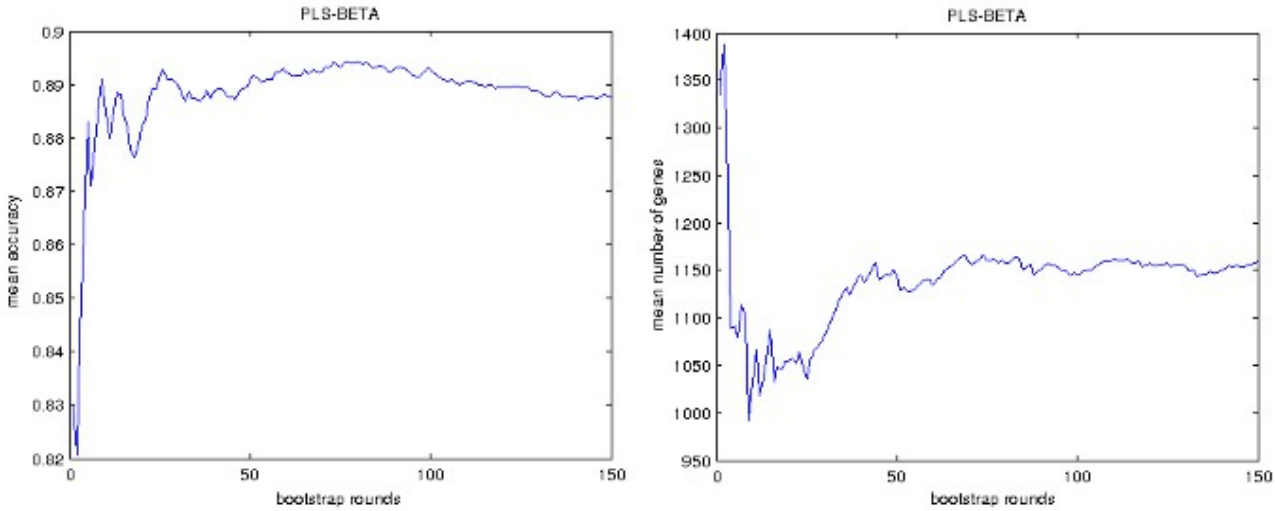

Figure 4.9 Left: Stabilization of PLS-BETA mean accuracy over all bootstrap datasets, when VIP>1. Right: Stabilization of PLS-BETA mean signature size over all bootstrap datasets, when VIP>1.

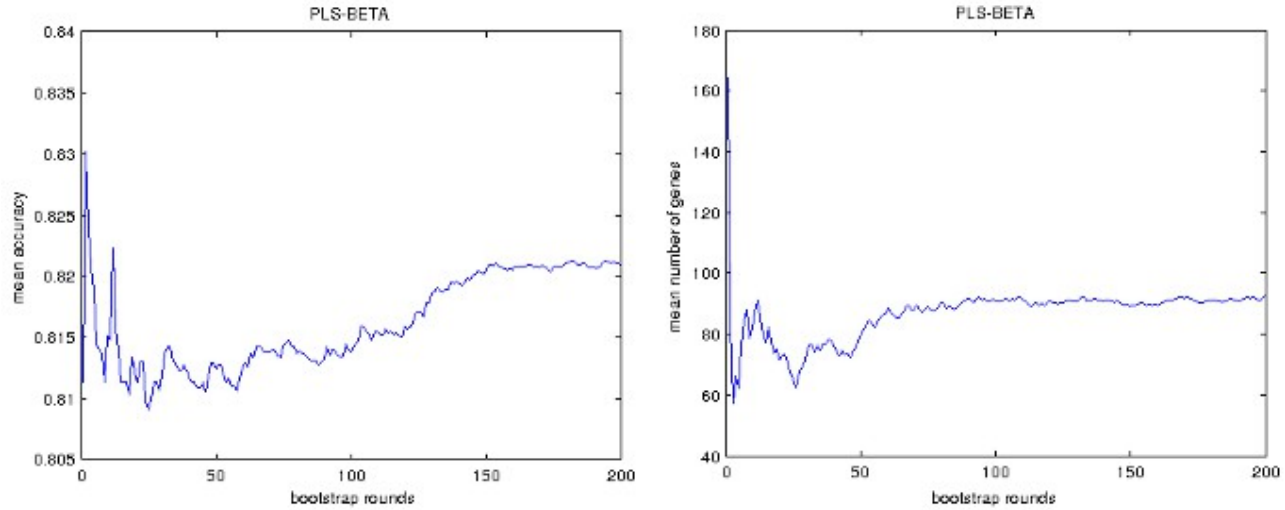

Figure 4.10 Left: Stabilization of PLS-BETA mean accuracy over all bootstrap datasets, when VIP>1.5. Right: Stabilization of PLS-BETA mean signature size over all bootstrap datasets, when VIP>1.5.

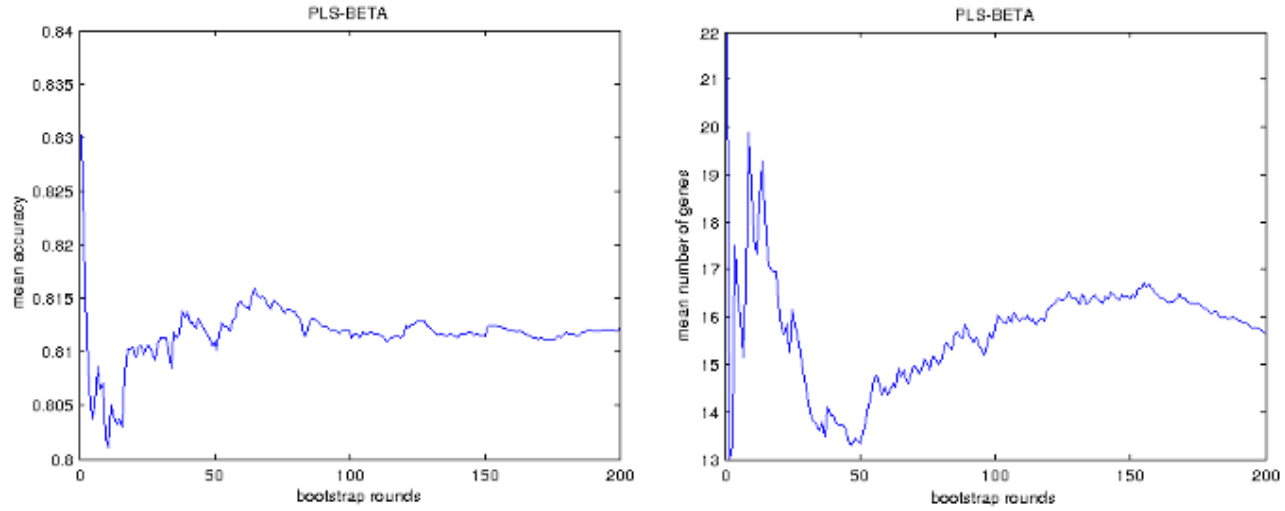

Figure 4.11 Left: Stabilization of PLS-BETA mean accuracy over all bootstrap datasets, when VIP>2. Right: Stabilization of PLS-BETA mean signature size over all bootstrap datasets, when VIP>2.

4.1.3 SVM Classifier

RFE & SVM

The classification accuracy of the SVM method comparable to that of RR and PLS methods in the VIP>1 case while it leads to a genomic signature of smaller size. However, the genomic signature of the SVM method can still be considered large, while the LASSO and PLS methods for VIP>1.5 or more, sacrifice a small amount of classification accuracy, while resulting in considerably smaller signatures. Finally, the runtime of SVM is moderate, comparable to that of the RLS methods RR and LASSO.

Classification Accuracy	Genomic Signature Size	Time per bootstrap dataset (seconds)
89.9%	640	106.72

Table 4.5 SBV results of SVM

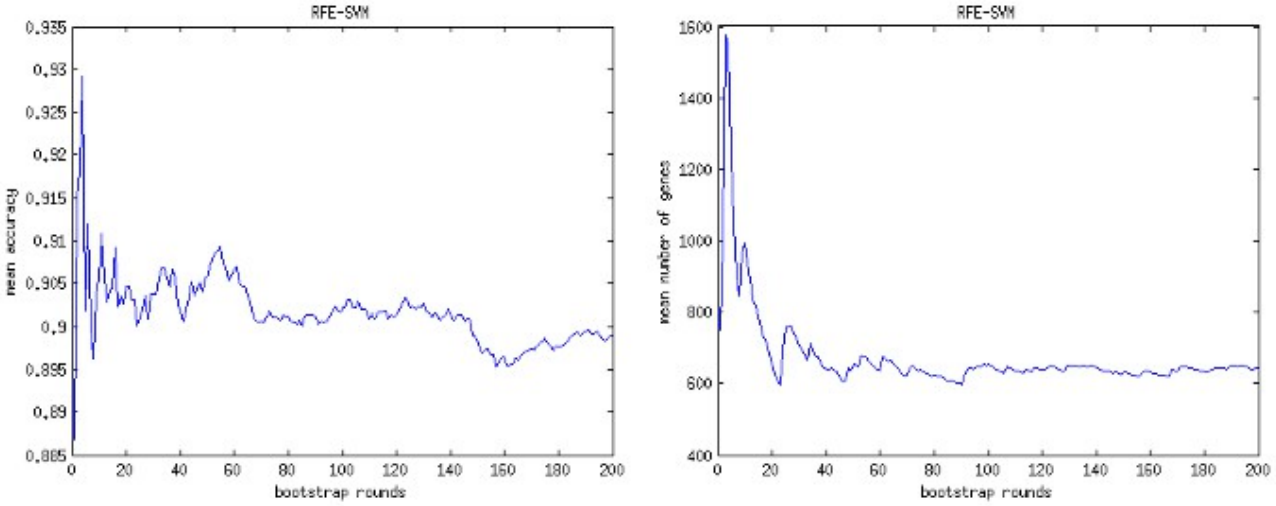

Figure 4.12 Left: Stabilization of SVM mean accuracy over all bootstrap datasets.
Right: Stabilization of SVM mean signature size over all bootstrap datasets.

4.1.4 K-NN Classifier

4.1.4.1 K-NN with I-RELIEF Feature Weighting

Reduction of the training set size – Setting the method-specific parameters

The feature weighting scheme of I-RELIEF proved to computationally expensive in a manner that made it practically impossible to run on an average personal computer for a dataset of the same size as other methods. The computational complexity of I-RELIEF is $O(N^2 \cdot P)$, where N the number of samples and P the number of features. While the computational complexity itself does not seem to be that extreme in theory, the running times in practice proved than a reduced training set was necessary. In that manner, the training as well as the test set size was set to 20 samples and the class ratio was the same as the original dataset (4 cancer to 1 control). Moreover, the bootstrap window B of SBV was set to 10 datasets, in order to lead to faster convergence of results. The kernel width σ of I-RELIEF was set to σ=20 in order to achieve fast convergence of the weights. Finally, two cases for the number of nearest neighbors of K-NN are tested (K=3, K=5).

RFE & K-NN with I-RELIEF Feature Weighting (Non-Linear)

K-NN with I-RELEF feature weighting leads to acceptable classification accuracy, despite using only a very small set of training samples, while it leads to extremely compact genomic signatures, which are in turn easily interpretable. However, even for a reduced training set of 20 samples, I-RELIEF leads to training time that is 20 times larger than RLS or SVM methods, while it is over 1000 times slower than the VIP>1.5

case of PLS methods.

K	Classification Accuracy	Genomic Signature Size	Time per **reduced** bootstrap dataset (seconds)
3	79.1%	9	2185.94
5	79.6%	15	2168.87

Table 4.6 SBV results of I-RELIEF K-NN

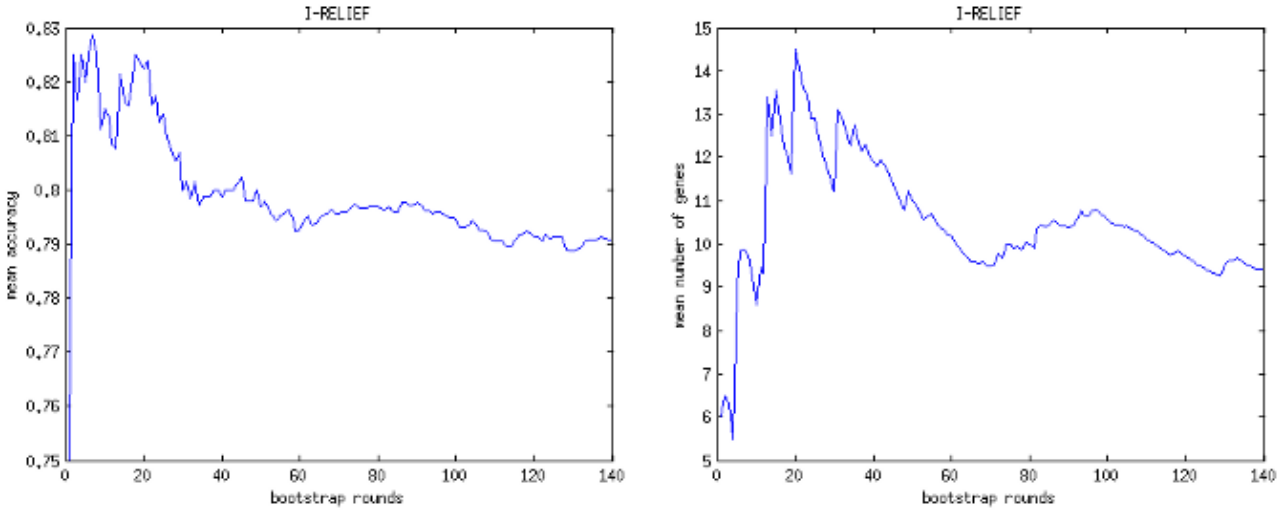

Figure 4.13 Left: Stabilization of I-RELIEF-KNN mean accuracy over all bootstrap datasets. Right: Stabilization of I-RELIEF-KNN mean signature size over all bootstrap datasets.

4.1.4.1 K-NN with PLS Feature Weighting

The good classification performance of the K-NN classifier and the small signatures and fast execution time of PLS methods lead to the idea of integrating both methodologies. The outcome of that integration is feature weighting for RFE using PLS methods, and classification using the K-NN classifier. The resulting methods are PLS-VIP K-NN and PLS-BETA K-NN. Different cases of the VIP score threshold are tested (VIP>1, VIP>1.5, VIP> 2), as well as different cases of nearest neighbors (K=3, K=5).

RFE & PLS-VIP K-NN

PLS-VIP K-NN lead to the best observed classification performance, while keeping the small genomic signatures associated with PLS methods in the cases where VIP>1.5 or 2. Running time was the same as that of PLS methods, ranging from one second to less than 11 seconds, depending on the VIP score. The execution time for the K=5 case is slightly larger, as is to be expected since more comparisons are undertaken in order to select the 2 extra nearest neighbors, compared to the case of K=3.

K=3 Nearest Neighbors

VIP score	Classification Accuracy	Genomic Signature Size	Time per bootstrap dataset (seconds)
>1	94.7%	793	7.33
>1.5	91.7%	86	1.5
>2	90.5%	19	0.95

Table 4.7 SBV results of PLS-VIP K-NN for K=3

K=5 Nearest Neighbors

VIP score	Classification Accuracy	Genomic Signature Size	Time per bootstrap dataset (seconds)
>1	94.3%	776	10.56
>1.5	92.3%	94	2.07
>2	90.4%	18	1.31

Table 4.8 SBV results of PLS-VIP K-NN for K=5

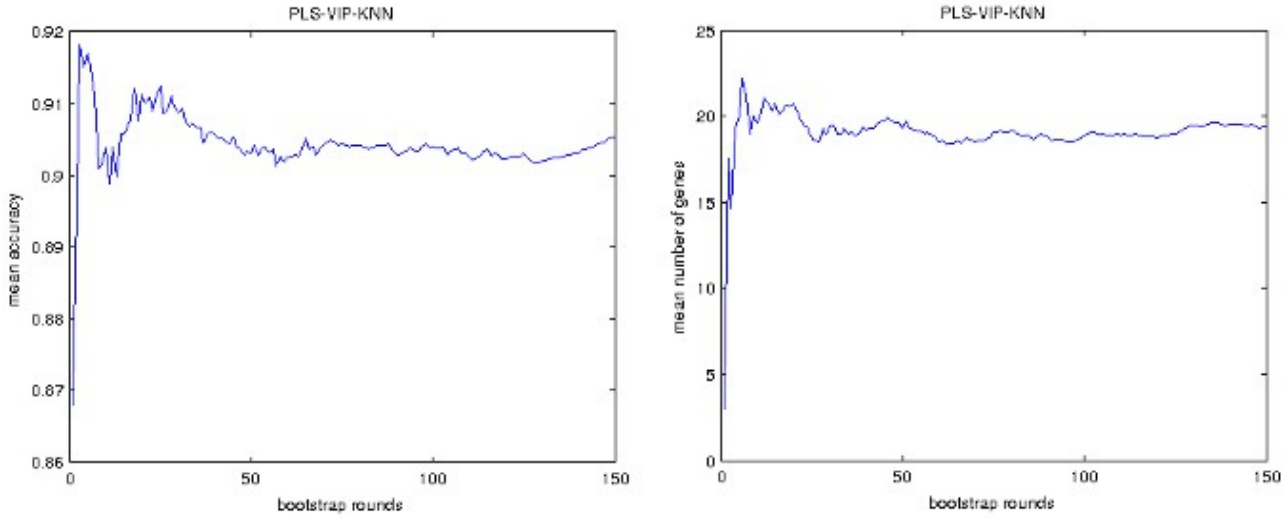

Figure 4.14 Left: Stabilization of PLS-VIP-KNN mean accuracy over all bootstrap datasets, for K=3, VIP>2. Right: Stabilization of PLS-VIP-KNN mean signature size over all bootstrap datasets, for K=3, VIP>2.

RFE & PLS-BETA K-NN

The results of PLS-BETA K-NN where identical to those of PLS-VIP K-NN. That is, exceptional classification performance, small genomic signatures for VIP>1.5 or 2 and very fast execution time of 1 up to approximately 11 seconds. Once again, the execution time of the K=5 case was slightly larger compared to that of K=3.

K=3 Nearest Neighbors

VIP score	Classification Accuracy	Genomic Signature Size	Time per bootstrap dataset (seconds)
>1	94.3%	1194	7.95
>1.5	91.8%	88	1.83
>2	90.7%	15	1.24

Table 4.9 SBV results of PLS-BETA K-NN for K=3

K=5 Nearest Neighbors

VIP score	Classification Accuracy	Genomic Signature Size	Time per bootstrap dataset (seconds)
>1	94.3%	1142	11.41
>1.5	91.6%	85	2.52
>2	89.7%	16	1.94

Table 4.10 SBV results of PLS-BETA K-NN for K=5

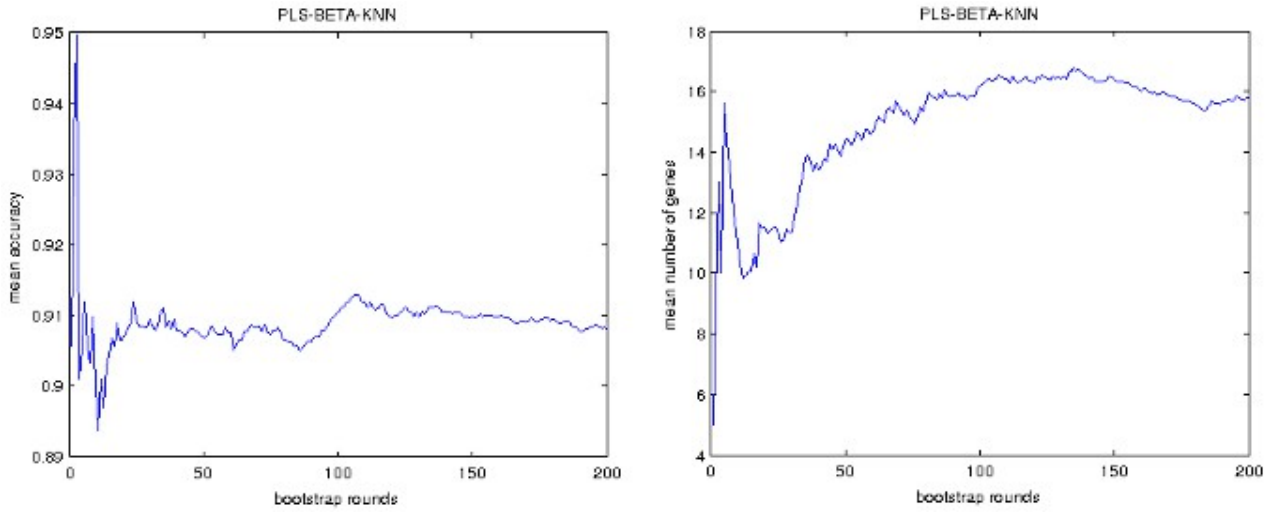

Figure 4.15 Left: Stabilization of PLS-BETA-KNN mean accuracy over all bootstrap datasets, for K=3, VIP>2. Right:Stabilization of PLS-BETA-KNN mean signature size over all bootstrap datasets, for K=3, VIP>2.

4.1.5 Synopsis of SBV Results

In the case of classification accuracy, the PLS K-NN methods outperformed all other classifiers, reaching accuracies of 90%, while extracting the smallest genomic signatures for VIP>1.5 or more. The SVM classifier was second in terms of classification accuracy but it kept a relatively large number of features. In the case of RLS classifiers, RR achieved higher classification accuracy than LASSO, however it lead to an approximately ten times larger genomic signature. LASSO lead to the second smallest signature of 136 genes, larger only than those extracted by PLS methods for VIP>1.5 or more. PLS methods lead to similar classification accuracy to RLS classifiers for VIP=1 but the extracted signature was too large. When the VIP criterion became more strict in the cases of VIP>1.5 and VIP>2 the resulting signatures are orders of magnitude smaller, but a small loss of classification accuracy is observed, as well. Finally, I-RELIEF K-NN leads to the smallest genomic signatures among all methods and has good generalization ability when it comes to the classification of unknown samples. However, the execution time of I-RELIEF, which even for a reduced dataset is 1000 times larger than that of PLS methods (VIP>1.5 or VIP>2) and 20 times larger than that of LASSO, makes the I-RELIEF K-NN method practically unusable.

FSS & Classification Method	Classification Accuracy	Genomic Signature Size	Time per bootstrap dataset (seconds)
RFE & LASSO	86.4%	136	118.52
RFE & Ridge Regression	88%	1372	142.95
RFE & PLS VIP (VIP>2)	82.2%	18	0.87
RFE & PLS VIP (VIP>1.5)	83.6%	88	1.4
RFE & PLS VIP (VIP>1)	87.6%	825	6.9
RFE & PLS BETA (VIP>2)	81.2%	16	1.2
RFE & PLS BETA (VIP>1.5)	82.1%	92	1.76
RFE & PLS BETA (VIP>1)	88.8%	1159	7.04
RFE & SVM	89.9%	640	106.72
RFE & I-RELIEF (K=3)	79.1%	9	2185.94
RFE & I-RELIEF (K=5)	79.6%	15	2168.87
RFE & PLS-BETA K-NN (VIP>1, K=3)	94.3%	1194	7.95
RFE & PLS-BETA K-NN (VIP>1.5, K=3)	91.8%	88	1.83

RFE & PLS-BETA K-NN (VIP>2, K=3)	90.7%	15	1.24
RFE & PLS-BETA K-NN (VIP>1, K=5)	94.3%	1142	11.41
RFE & PLS-BETA K-NN (VIP>1.5, K=5)	91.6%	85	2.52
RFE & PLS-BETA K-NN (VIP>2, K=5)	89.7%	16	1.94
RFE & PLS-VIP K-NN (VIP>1, K=3)	94.7%	793	7.33
RFE & PLS-VIP K-NN (VIP>1.5, K=3)	91.7%	86	1.5
RFE & PLS-VIP K-NN (VIP>2, K=3)	90.5%	19	0.95
RFE & PLS-VIP K-NN (VIP>1, K=5)	94.3%	776	10.56
RFE & PLS-VIP K-NN (VIP>1.5, K=5)	92.3%	94	2.07
RFE & PLS-VIP K-NN (VIP>2, K=5)	90.4%	18	1.31

Table 4.11 Synopsis of SBV results.

4.2 Significance Evaluation Results

4.2.1 Classification Accuracy Significance

The p-value of each observed classification accuracy is estimated utilizing permutation tests which are introduced in section 3.3. According the the estimated p-values, the observed accuracies of all classification methods, according to SBV, are statistically significant and reflect the underlying biological system. The usual threshold declaring when an observation is statistically significant is to have a corresponding p-value < 0.5. The maximum p-value observed across all classification methods was 0.007, one order of magnitude smaller than the typical threshold. Another observation is that increasing the VIP score threshold for PLS methods, slightly increases the corresponding p-value but the statistical significance of the results is not hindered. That is, randomness in the results is slightly increased but they still strongly reflect the underlying biological model.

FSS & Classification Method	Classification Accuracy	p-value
RFE & Ridge Regression	88%	<0.001
RFE & LASSO	86.4%	0.001
RFE & PLS VIP (VIP>2)	82.2%	0.005
RFE & PLS VIP (VIP>1.5)	83.6%	0.002
RFE & PLS VIP (VIP>1)	87.6%	<0.001
RFE & PLS BETA (VIP>2)	81.2%	0.007
RFE & PLS BETA (VIP>1.5)	82.1%	0.011
RFE & PLS BETA (VIP>1)	88.8%	<0.001
RFE & SVM	89.9%	<0.001
RFE & PLS-BETA K-NN (VIP>1, K=3)	94.3%	<0.001
RFE & PLS-BETA K-NN (VIP>1.5, K=3)	91.8%	<0.001
RFE & PLS-BETA K-NN (VIP>2, K=3)	90.7%	<0.001
RFE & PLS-BETA K-NN (VIP>1, K=5)	94.3%	<0.001
RFE & PLS-BETA K-NN (VIP>1.5, K=5)	91.6%	<0.001
RFE & PLS-BETA K-NN (VIP>2, K=5)	89.7%	<0.001
RFE & PLS-VIP K-NN (VIP>1, K=3)	94.7%	<0.001
RFE & PLS-VIP K-NN (VIP>1.5, K=3)	91.7%	<0.001
RFE & PLS-VIP K-NN (VIP>2, K=3)	90.5%	<0.001

RFE & PLS-VIP K-NN (VIP>1, K=5)	94.3%	<0.001
RFE & PLS-VIP K-NN (VIP>1.5, K=5)	92.3%	<0.001
RFE & PLS-VIP K-NN (VIP>2, K=5)	90.4%	<0.001

Table 4.12 Classification accuracy statistical significance results

4.2.2 Genomic Signature Significance

Unifying The Genomic Signatures: The Common Gene Signature

Instead of assessing the genomic signature of each method separately, a unifying approach was implemented. That is, the common genes existing in the signatures of all methods were selected as the unified common gene signature. Since there are 3 difference cases used for the VIP score threshold of PLS-methods, 3 different unified signatures were extracted, the 77 gene, 16 gene and 5 gene signatures when VIP>1, VIP>1.5 and VIP>2 was used for PLS methods, respectively.

The performance of these signatures was compared to that of random signatures of the same size using a fixed number of 1000 bootstrap datasets, as noted in section 3.3. The classification accuracy on test data was extracted by using these signatures in conjunction with a K-NN classifier for K=3 (noted as 3-NN) and an SVM. The common gene signatures reached up to 95% classification accuracy with the 3-NN classifier and up to 83% with the SVM. The 77 gene signature maximizes the performance of 3-NN, while the 16 gene signature maximizes the performance of the SVM. Finally, as reported in [4] and [5], the performance of random signatures was comparable to that of the genomic signatures extracted by SBV and performed better with a probability Prs that even reached 25%. However, assessing the same test in a different perspective, reveals the signatures produced by SBV still perform better at least 75% of the time.

Signature	G: Signature Size	Classification Accuracy	p-value	Mean classification accuracy of G random genes	Probability Prs
VIP>1	77	95.2%	<0.001	92.7%	25.5%
VIP>1.5	16	92.8%	<0.001	88.2%	11.4%
VIP>2	5	88%	<0.001	84.2%	23.6%

Table 4.13 “Common gene” signature statistical significance results using a 3-NN classifier.

Signature	G: Signature Size	Classification Accuracy	p-value	Mean classification accuracy of G random genes	Probability Prs
VIP>1	77	81.1%	<0.013	67.5%	23.4%
VIP>1.5	16	83.3%	<0.004	73.6%	6.6%
VIP>2	5	80.8%	<0.043	68%	9.4%

Table 4.14 “Common gene” signature statistical significance results using a SVM classifier.

4.3 Consistency Evaluation of Signature Classification Accuracy

As mentioned in section 3.4 the consistency of a classification method refers to the the ability to yield similar performance when applied on the same test set multiple times, while using different sets are used for training. Similarly to section 4.2.2, the “common gene” signature of all methods was used in order to evaluate the consistency of the corresponding classification accuracy. As presented in 3.4, the proposed methodology generates 30 bootstrap training datasets and only one bootstrap test set. The mean value, variance and standard deviation of the observed accuracy are then extracted. The process is repeated 100 times and the overall results are averaged.

Signature	G: Signature Size	Average Classification Accuracy	Variance	Standard Deviation
VIP>1	77	95%	0.000659	0.025112
VIP>1.5	16	93%	0.000875	0.029008
VIP>2	5	89%	0.001177	0.033744

Table 4.15 Classification accuracy consistency of “Common gene” signature, using a 3-NN Classifier.

Signature	G: Signature Size	Average Classification Accuracy	Variance	Standard Deviation
VIP>1	77	81%	0.010651	0.099798
VIP>1.5	16	83%	0.000586	0.023157
VIP>2	5	81%	0.000105	0.008960

Table 4.16 Classification accuracy consistency of “Common gene” signature, using a SVM Classifier.

According to the above observations, when using a 3-NN classifier all signatures lead to consistent results when applied multiple times on the same test set. Moreover, the 77 gene signature maximizes the classification performance of the 3-NN method, as well as the consistency of the observed accuracy, leading to a minimized variance of the observed classification accuracy on the same test set when several instances of the classifier are trained using different training sets. In the case of the SVM, the 16 gene signature maximizes classification performance while the 5 gene signature maximizes consistency, however the consistency observed is similar to that of the 16 gene signature. Both classification methods lead to the observation that the signature maximizing classification performance tends to maximize the consistency of the classification accuracy observed.

4.4 SBV Results Compared to 10-Fold CV

While the results of both methods are comparable, SBV leads to smaller genomic signatures, when compared to 10-Fold CV. Smaller signatures are preferred, since they are easier to interpret in a biological manner. Moreover, the results across different executions of SBV are very similar, while those of 10-Fold CV tend to vary more. Concerning the assessment of classification accuracy, PLS methods due to their exceptional stability, lead to almost identical performance estimates between the two evaluation methods. On the contrary, 10-Fold CV underestimates the accuracy of the SVM by approximately 14%. Finally 10-Fold CV completely fails to assess the performance of the RLS methods RR and LASSO, resulting in values that are off by 60%. That deviation of performance assessment, especially in the case of RLS methods could be a result of bad class ratio and might be improved if a different CV scheme is implemented, such as stratified CV.

Another important observation is that while 10-Fold CV leads to larger signatures, the common genes of these signatures are fewer than the common genes of SBV. That is probably the effect of larger amounts random noise being included in the signatures extracted by 10-Fold CV than those extracted by SBV. The likely reason behind the reduced stability of 10-Fold CV is the small number of iterations. That is, after only 10 iterations it is unlikely that estimates will have converged, according to the weak law of large numbers. On the contrary, SBV utilizes a criterion that guarantees the stability of performance estimates, which results in a considerably larger number of iterations.

FSS & Classification Method	SBV: Classification Accuracy	10fold-CV: Classification Accuracy	SBV: Genomic Signature Size	10fold-CV: Genomic Signature Size
RFE & Ridge Regression	88%	26.9%	1372	2044
RFE & LASSO	86.4%	27.8%	136	301
RFE & PLS VIP (VIP>2.5)	78.4%	80.0%	3	3
RFE & PLS VIP (VIP>2)	82.2%	82.6%	18	20
RFE & PLS VIP (VIP>1.5)	83.6%	83.2%	88	101
RFE & PLS VIP (VIP>1)	87.6%	87.5%	825	754
RFE & PLS BETA (VIP>2.5)	79.5%	80.3%	4	3
RFE & PLS BETA (VIP>2)	81.2%	81.1%	16	15
RFE & PLS BETA (VIP>1.5)	82.1%	81.5%	92	124
RFE & PLS BETA (VIP>1)	88.8%	87.3%	1159	1083
RFE & SVM	89.9%	75.2%	640	930

Table 4.17 Comparison between classification accuracy and genomic signature size of SBV and 10-Fold CV

	Stable Bootstrap Validation	10fold Cross Validation	Overlap
Common genes of all methods (VIP>1 for PLS methods)	77	59	35% (27 of 77)
Common genes of all methods (VIP>1.5 for PLS methods)	16	17	47% (8 of 17)
Common genes of all methods (VIP>2 for PLS methods)	5	3	60% (3 of 5)

Table 4.18 Comparison of size between SBV and 10-Fold CV common gene signatures

4.5 Biological Evaluation

4.5.1 Gene Signatures

Diseases such as breast cancer are highly complex and characterized by a number of genetic aberrations. Patients associated with similar clinical and pathological features may have very different disease profiles at the molecular level and may respond differently to treatment. Toward this direction genome-wide expression profiling of pathological samples (e.g. tissue) has become an important tool to identify gene sets and gene signatures that can be used to predict disease development and clinical endpoints, such as survival and therapy response [48] [50].

4.5.1.1 Convergence of Gene Signatures in Biological Pathways

Using the Integrated Pathway Analysis Database for Systematic Enrichment Analysis [IPAD] [35], we perform enrichment analysis from genes of all three common gene signatures and from '19 gene signature' from PLS-VIP-3NN* method. IPAD is a comprehensive database covering about 22,498 genes, 25,469 proteins, 1956 pathways, 6704 diseases, 5615 drugs, and 52 organs integrated from databases including the BioCarta, KEGG, NCI-Nature curated, Reactome, CTD, PharmGKB, DrugBank, PharmGKB, and HOMER. The results are illustrated in the following table.

	CONVERGENCE of GENE SIGNATURES			
Pathways	**'77 common gene signature'**	**'16 common gene signature'**	**'5 common gene signature'**	**'19 gene signature'**
ECM-receptor interaction [hsa04512]	**3 genes** (COL11A1; FN1; COMP)	**3 genes** (COL11A1; COMP; FN1)	**2 genes** (COMP; FN1)	**3 genes** (COL11A1; COMP; FN1)
Focal adhesion [hsa04510]	**4 genes** (COMP; EGF; COL11A1; FN1)	**3 genes** (COL11A1; COMP; FN1)	**2 genes** (COMP; FN1)	**3 genes** (COL11A1; COMP; FN1)
Signal transduction [162582]	**11 genes** (PENK; NR4A1; EDN2; EDN3; CCL19; FGF18; FN1; FGFR3; EGF; ATP6V0A4; NRG1)	**3 genes** (EDN2; CCL19; FN1)	**2 genes** (FN1; CCL19)	**4 genes** (CXCL9; OXTR; FN1; CCL19)

Table 4.19 Convergence of gene signatures in key pathways for tumor growth, progression and metastasis. Genes known to be associated with cancer according to G2SBC are underlined.

As shown in Table 4.19, focusing on pathways implied by a minimum of two genes in each signature, we found that all three common gene signatures, as well as the '19 gene signature' from PLS-VIP-3NN* method, converge to the following three key pathways:

- Signal transduction

Signal transduction refers to communication processes used by regulatory molecules to mediate the essential cell processes of growth, differentiation, and survival. Signal transduction elements interact through complex biochemically related networks. Aberrations in signal transduction elements can lead to increased proliferative potential, sustained angiogenesis, tissue invasion and metastasis, and apoptosis inhibition. We know that most human neoplasms, including breast cancer have aberrant signal transduction elements. Several compounds that target aberrant signal transduction elements are in development and others are commercially available (e.g. trastuzumab for the treatment of metastatic breast cancer overexpressing the ErbB-2 receptor) [36].

- The extracellular matrix (ECM)

The extracellular matrix (ECM) is a complex network of macromolecules with distinctive physical, biochemical, and biomechanical properties. The ECM is a major component of the local microenvironment, or niche, of a cancer cell that plays important roles in cancer development. Although tightly controlled during embryonic development and organ homeostasis, the ECM is commonly deregulated and becomes disorganized in diseases such as cancer. Abnormal ECM affects cancer progression by directly promoting cellular transformation and metastasis. Importantly, however, ECM anomalies also deregulate behavior of stromal cells, facilitate tumor-associated angiogenesis and inflammation, and thus lead to generation of a tumorigenic microenvironment [37].

- Focal adhesion

Focal adhesions lie at the convergence of integrin adhesion, signaling and the actin cytoskeleton. Cells modify focal adhesions in response to changes in the molecular composition, two-dimensional (2D) vs. three-dimensional (3D) structure, and physical forces present in their extracellular matrix environment. The components of focal adhesions are diverse and include scaffolding molecules, GTPases, and enzymes such as, lipases, proteases, phosphatases, and kinases. One of the critical tyrosine kinases that are linked to the processes of tumor invasion and survival is the Focal adhesion kinase (FAK). This protein plays a critical role in intracellular processes of cell adhesion, motility, survival, and cell cycle progression. Cancer is often characterized by defects of these processes [51].

Overall, we show that the aforementioned crucial pathways - signal transduction, focal adhesion, ECM-receptor interaction - constitute important components of all three common gene signatures, as well as of the '19 gene signature' from PLS-VIP-3NN* method. Their deregulations have a great impact in tumor development, progression and metastasis. These findings enhance the robustness of the proposed methodology.

4.5.2 Biological Features of Gene Signatures

4.5.2.1 Enrichment Analysis of Molecular Pathways - Biological Processes - Disease

When gene sets of interest share genes, the examination of how they overlap can highlight common processes, pathways, and underlying biological themes. Hanahan and Weinberg in two elegant reviews (2000 and 2011) communicate specific biological capabilities, which include sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, activating invasion and metastasis, reprogramming of energy metabolism and evading immune destruction. These biological entities constitute the hallmarks of cancer and are acquired during multistage tumor development in humans. They form a structured principle to rationalize the complex nature of neoplastic diseases [52] [53], which may be initially explored through observation of a set of perturbed genes in microarray experiments, but must be further confirmed in association to such mechanisms and functionality of oncogenesis. Indeed, several studies suggest that although different molecular signatures may not be composed of the same genes, they are mostly targeting the same number of pathways and processes that show high correlation with risk categories [54] [55]. Toward this direction, many studies attempt to identify a common list of genes in primary tumor tissues from breast cancer patients and subsequently by applying gene-set enrichment analysis to detect key pathways and biological processes that are well known to be implicated in breast cancer [58]; they are biologically meaningful concerning the pathological mechanisms of tumor development, progression, invasion and metastasis. The exploration of common biological processes and pathways in our derived signatures is treated in the next subsections.

The unified '77 common-gene signature' and the '19 gene signature' from PLS-VIP-3NN* method

Following our proposed framework of Stable Bootstrap Validation (SBV), a unified '77 common-gene signature' was selected, which is closely associated with several aspects of breast tumorigenesis and progression, as well as patient-specific molecular and clinical characteristics. A '19 gene signature' obtained from PLS-VIP-3NN* method bears similar properties with the unified '77 common-gene signature'. For the biological interpretation of the selected unified '77 common-gene signature' and of the '19 gene signature' from PLS-VIP-3NN* method, the Genes-to-Systems Breast Cancer (G2SBC) Database [33] and WebGestalt (WEB-based GEne SeT AnaLysis Toolkit) [34] are used.

G2SBC provides literature based evidences that 51.95% of genes of the unified '77 common-gene signature' and 36.84% of genes of the '19 gene signature' are altered in breast cancer cells (Appendices A, B, Supplemental Table I). Thus, the '19 gene signature', which includes only six genes from '77 common-gene signature', shares many similar attributes with the unified '77 common-gene signature'.

The WebGestalt functional analysis in terms of gene ontology (GO) biological processes and KEGG pathways reveals significant ($p<0.05$) enriched processes in both signatures (i.e. cell proliferation, cell development, growth, cell differentiation, cell migration, extracellular matrix organization) and KEGG pathways (Pathways in cancer, ECM-receptor interaction, MAPK signaling pathway, ErbB signaling pathway, Cytokine-cytokine receptor interaction), which are implicated in breast tumorigenesis, progression and metastasis (Appendix A, Supplemental Tables I-IV).

Furthermore, the disease association analysis verified the significant ($p<0.05$) relation of both signatures to breast cancer disease (referred as breast diseases, breast neoplasms) but also creates gene subsets of the signatures that are directly related to breast cancer (involved in neoplastic process, carcinoma *in situ*, disease progression, recurrence, neoplasm invasiveness, neoplasm metastasis, disease susceptibility, genetic predisposition to disease, chromosome aberrations and hypersensitivity). Other subsets are related to a variety of diseases appearing as breast cancer comorbidities (hypertension, diabetes mellitus, osteoarthritis, depression, etc.) or treatment side effects (e.g. musculoskeletal diseases) (Appendix B, Supplemental Tables I-IV). Thus, both signatures provide insights to breast carcinogenesis, as well as patients' clinical and molecular profiles according to the original studies (GEO access numbers: GSE22820, GSE19783, GSE31364). Gene sets of both signatures that are not yet associated with breast cancer but participate in various significant ($p<0.05$) biological processes, pathways and a broad spectrum of diseases (248 diseases for the '77 common-gene signature' and 103 for the '19 gene signature'), should be further explored in order to decode their implications in breast carcinogenesis and clinical outcome of breast cancer. According to this analysis, both signatures appear to be potential useful as clinical signatures.

In addition, G2SBC provides that 50% of genes of the '16 common-gene signature' and 40% of genes of the '5 common-gene signature' are associated with breast cancer (Appendices C, D, Supplemental Table I). The WebGestalt functional analysis reveals significant ($p<0.05$) enriched KEGG pathways (ECM-receptor interaction**,** Focal adhesion) in both signatures, and discloses significant ($p<0.05$) enriched

GO processes such as cellular component movement, locomotion, cell motility, and cell migration in the '16 common-gene signature', while it fails to recover the same processes as statistical significant in the '5 common-gene signature'. These biological processes and pathways are linked to tumor invasion, progression and metastasis (Appendices C, D, Supplemental Tables II and III). Also, according to the disease association analysis the '16 common-gene signature' is significantly ($p<0.05$) associated to breast cancer such as the aforementioned signatures, while the '5 common-gene signature' is significantly ($p<0.05$) related only to a narrow spectrum of breast cancer (e.g. carcinoma, papillary and neoplastic processes) (Appendices C, D, Supplemental Table IV).

Summarizing, we can infer that the unified '77 common-gene signature' and the '19 gene signature' are more significant than the rest, i.e. the '16 common-gene signature' and the '5 common-gene signature', because they enclose a wider range of processes, pathways and disease features that concurrently cover a broader range of the multistep process of human breast carcinogenesis. According to this analysis, both signatures appear to be potential useful as clinical signatures.

4.5.2.2 Gene Families

Gene Set Enrichment Analysis (GSEA) was also performed by using the "Gene Families" tool of Molecular Signatures Database (MSigDB), in order to gain further insight into the biology behind a gene signature. Moreover, this tool is used to retrieve a functional overview of the selected gene signatures by categorizing their genes into a small number (eight) of carefully chosen "gene families", as illustrated in Table 4.20.

	GENE SIGNATURES			
GENE FAMILIES	**'77 common gene signature'**	**'16 common gene signature'**	**'5 common gene signature'**	**'19 gene signature'**
Tumor Suppressors	CDKN2A			
Oncogenes	FGFR3; NTRK3			
Translocated Cancer Genes	FGFR3; NTRK3			
Protein Kinases	FGFR3; NTRK3			
Cell Differentiation Markers	FGFR3; CD19 LAMP3			CEACAM6
Homeodomain Proteins	HOXB13; SIX1			
Transcription Factors	HOXB13;SIX1;ASCL2;FOXJ1; NR4A1;PPARGC1A;SOX11; TFCP2L1;VGLL1;ZNF334; CITED1			ELF5
Cytokines and Growth Factors	**CCL19**;CCL18;EDN2;EDN3; EGF;FGF18;GREM1;NRG1; OGN;PENK	**CCL19;** EDN2	**CCL19**	**CCL19**; CXCL9

Table 4.20 "Gene Families" for all three common gene signatures and the 19 gene signature.

A gene family describes any collection of proteins that share a common feature such as homology or biochemical activity. "Gene Families" are very important for understanding the complex nature of breast cancer, as well as for its clinical evaluation and therapy. They are briefly described as follows:

- **Oncogenes:** (a single mutated allele is sufficient to contribute to oncogenesis)

Oncogenes consist one group of genes implicated in the development of cancer is damaged genes. Oncogenes are related to normal genes called proto-oncogenes that encode components of the cell's normal growth-control pathway; they arise from the mutation of proto-oncogenes. Oncogenes are genes whose presence in certain forms and/or overactivity can stimulate the development of cancer. When oncogenes arise in normal cells, they can contribute to the development of cancer by instructing cells to make proteins

that stimulate excessive cell growth and division
[http://www.cancer.gov/cancertopics/understandingcancer/cancer/].

- **Tumor suppressors:** (both alleles of these genes need to be mutated for oncogenesis)

Tumor suppressor genes are normal genes whose absence can lead to cancer. In other words, if a pair of tumor suppressor genes are either lost from a cell or inactivated by mutation, their functional absence might allow cancer to develop [http://www.cancer.gov/cancertopics/understandingcancer/cancer/].

It is known that breast cancer progression involves multiple genetic events, which can activate dominant acting oncogenes and disrupt the function of specific tumor suppressor genes. Several karyotypic and epidemiological analyses of mammary tumors at various stages suggest that breast carcinomas become increasingly aggressive through the stepwise accumulation of genetic changes. The majority of genetic changes found in human breast cancer fall into two categories: gain-of-function mutations in proto-oncogenes, which stimulate cell growth, division, and survival; and loss-of-function mutations in tumor suppressor genes that normally help prevent unrestrained cellular growth and promote DNA repair and cell cycle checkpoint activation [39].

- **Translocated cancer genes:** (genes mutated by translocation)

The most common class of somatic mutation that is registered in the cancer-gene census involves chromosomal translocations that result in a chimeric transcript or apposition of one gene to the regulatory regions of another gene — usually immunoglobulin or T-cell-receptor genes. This mutation type is common in leukaemias, lymphomas and mesenchymal tumours. However, several examples have now been reported among epithelial neoplasms, including breast secretory carcinomas (ETV6 and NTR3). Because two genes are structurally rearranged in each chromosomal translocation, the number of translocated cancer genes, compared with other types of mutated cancer gene, is exaggerated in the census [44].

- Cell differentiation markers**:** (Human leukocyte and stromal cell molecules: the CD markers)

The leukocyte surface molecules (CD molecules), include a selection of cell surface glycoproteins and glycolipids which are expressed by leukocytes (cells of the immune system) and mediate their interaction with antigen, with other components of the immune system, and with other tissues. CD molecules have provided targets for diagnosis and therapy. Notice that leukocytes are centrally involved in defense against infection, in autoimmune disease, allergy, inflammation, and in organ graft rejection. Lymphomas and leukemias are malignancies of leukocytes, and the immune system is almost certainly involved in most other cancers, including breast cancer [41] [42].

- Protein kinases**:** The protein kinase complement of the human genome

Protein kinases form a vast family of enzymes, encoded by more than 500 genes in human cells 2002); in comparison, human cells possess far fewer phosphatase genes, indicating that protein kinases have higher substrate specificity. Moreover, protein kinases transfer a phosphate, generally from ATP (adenosine triphosphate) to protein substrates; but protein phosphorylation is a reversible reaction, phosphatases being the counterpart enzymes to protein kinases. While not all cellular proteins are phosphoproteins, more than 90% are phosphorylated at some point in their existence, and the phosphorylation status of a protein depends on the balance between protein kinase and phosphatase activities. Perturbation of this fragile equilibrium can often lead to defects in key cellular mechanisms such as signal transduction, cell differentiation, cell proliferation and cell cycle progression. Members of the protein kinase family are amongst the most commonly mutated genes in human cancer, and several pathways are frequently deregulated in breast cancer as a consequence of mutations in these genes. Given this vital role in normal cell function and disease, protein kinases form the second most important group of proteins considered as priority targets by pharmaceutical companies for the development of new anticancer therapies [45] [46].

- Homeodomain proteins**:** (Human homeodomain proteins)

Homeobox genes comprise a large and essential family of developmental regulators that are vital for all aspects of growth and differentiation. Homeobox genes belong to the principal examples in which the anomalous expression of genes that regulate growth and development have been implicated in carcinogenesis. These genes encode transcriptional regulatory proteins (homeoproteins) that are widely

used during development and are often aberrantly expressed in cancer. Homeodomain-containing proteins are transcription factors that play a critical role in various cellular processes, including body plan specification, pattern formation and cell fate determination [40] [43].

- Transcription factors**:** A compilation of human transcription factors

Transcription factors are gene regulatory proteins endowed with sequence-specific DNA recognition and the ability to positively or negatively influence the rate and efficiency of transcript initiation at a gene containing the factor's cognate recognition sequence, or DNA response element. Since transcription factors lie at the heart of almost every fundamental developmental and homeostatic organismal process - including DNA replication and repair, cell growth and division, control of apoptosis and cellular differentiation - it is not surprising that inherited or acquired defects in transcription factor structure and function contribute to human carcinogenesis [47].

- **C**ytokines and growth factors**:** (Human cytokine and growth factor genes)

Cytokines are glycoproteins of low molecular weight, which are rapidly synthesized and usually secreted by different healthy and diseased cells (mainly mononuclear phagocytes and activated T lymphocytes) mainly after stimulation. In multicellular organisms, cytokines are intercellular mediators that regulate survival, growth, differentiation, and the effector functions of cells. Therefore, it is not surprising that cytokines significantly affect the growth of tumours in vivo. On the other hand, they are also produced by cancer cells and represent a network with a large variety of molecularly and functionally different members that may act as tumour growth-promoting or inhibiting factors. As they affect the growth and function of immunocompetent cells, they can activate or modulate specific or non-specific antitumor responses. Furthermore, because cytokines are mediators of the effector response from innate and acquired cellular immunities, they are probably involved in the mechanism from tumour cell evasion of the immunosurveillance system [56]

Growth factors, encoded by growth factor genes, bind to receptors on the cell surface, which activate signaling enzymes inside the cell that, in turn, activate special proteins called transcription factors inside the cell's nucleus. The activated transcription factors "turn on" the genes required for cell growth and proliferation.

Based on the gene family composition illustrated in Table 4.20, we observe that the unified '77 common-gene signature induces many more significant gene families than any other gene signature. Importantly, the chemokine CCL19 is the only common gene across all gene signatures and is assessed as potential biomarker of metastatic dissemination in primary breast cancer [57].

In summary, Figure 4.16 provides the association of the individual gene signatures with diseases, biological processes, pathways and gene families, which have particular importance in a wide spectrum of the multistep process of breast carcinogenesis. We notice that all signatures capture key issues, but the first signature ('77 common-gene signature') embodies a wealth of information that is gradually lost in other common signatures, as well as in the individual signature of 19 genes. Considering all the above aspects, the biological evaluation correlates very well with the methodological outcomes of this study. Key biological processes and pathways that are implicated in breast cancer, as well as the wide variety of disease associations and functional gene families derived, synthesize a robust '77 common-gene signature'.

Figure 4.16 Comparison of gene signatures in relation with breast cancer features, GO biological processes, KEGG pathways and gene families.

Conclusion

This thesis aimed at introducing a framework for selection of stable genomic signatures that maximize classifier performance. By performing multivariate feature subset selection (FSS) an a dataset preprocessed by univariate FSS a two-step feature selection scheme was achieved, aiming to utilize the advantages of both univariate and multivariate FSS methods, leading to small genomic signatures while being computationally efficient.

Furthermore, an evaluation method called Stable Bootstrap Validation (SBV) was proposed that employs bootstrap resampling of the original dataset and an explicit stability assessment criterion in order to extract stable estimates of the classification accuracy as well as the genomic signature size; the number of genes selected in the signature. Under the assumption that the mean classification accuracy and the mean signature size extracted from each bootstrap dataset are independent identically distributed (i.i.d.) random variables, SBV is guaranteed to lead to stable estimates according to the Law of Large Numbers (LLN). Experimental results confirm that SBV always achieves stability of results after a sufficient number of bootstrap datasets have been evaluated. The results also confirm that estimates for the classification accuracy and genomic signature are very similar for independent executions of SBV. Compared to similar evaluation methods that utilize resampling or random splitting of the original dataset, SBV requires fewer bootstrap datasets to achieve stability, since it utilizes and explicit stability criterion. Thus, SBV is a more computationally efficient approach. Moreover, while similar methods only extract a stable estimate for the classification accuracy and select a number of genes based on their selection frequency, SBV extracts stable estimates for the classification accuracy as well as the genomic signature size and then proceeds to select the genes having the largest selection frequency.

The statistical significance of the observed classification accuracy and genomic signature was also evaluated. The process of significance evaluation determines to what extent the observed results reflect random noise, or the underlying biological model. To determine the statistical significance of the observed classification accuracy, permutation tests are performed to calculate the corresponding p-value. The significance of the genomic signature is assessed by comparing it's performance to random signatures of the same size. The final step of the proposed methodology is assessing the consistency of the observed classification accuracy. The consistency of a classification method refers to the ability of yielding similar classification results on the same test set, while different sets of samples are used for training. In that manner, a fixed genomic signature is selected and a fixed test set is generated by bootstrap resampling. Then, the classification method is trained on multiple bootstrap training sets and performance is assessed using the same test set, leading to measurement of mean classification accuracy, variance and standard deviation. The process is repeated multiple times and the results are averaged.

The above methodology is performed on a breast cancer dataset. Recursive Feature Elimination (RFE) is the multivariate FSS method used, while several categories of classification methods are implemented: Regularized Least Squares (RLS) Classifiers, Ridge Regression (RR) and Least Absolute Shrinkage and Selection Operator (LASSO). Partial Least Squares (PLS) Classifiers PLS-VIP and PLS-BETA. Support Vector Machines (SVMs) and K-Nearest Neighbor (K-NN) classifiers are also implemented. Since K-NN does not provide feature weights that are necessary for RFE, it was used in conjunction with I-RELIEF feature weighting. The high computational cost of I-RELIEF and the good classification accuracy of K-NN lead to pairing the K-NN classifier with the computationally efficient PLS feature weighting, resulting in the PLS-VIP K-NN and PLS-BETA K-NN classification methods.

Experimental results proved that SBV reached stable results after a maximum of 200 iterations on a worst case scenario, which is half what the number of iterations required by the similar evaluation methodology of Davis et al. [1]. Moreover, observed estimates for the classification accuracy and the genomic signature were consistent across different and independent executions of SBV. According to the SBV results, PLS K-NN outperformed all other methods, reaching accuracy close to, or greater than 90% while keeping as few as 16 genes. Even, greater classification accuracy of 94% was achieved when the methods were tuned to be less selective, resulting in larger signatures. Both PLS-VIP K-NN and PLS-BETA K-NN yielded almost identical results, while the choice of K=3 or 5 neighbors proved to have no significant effect on the resulting accuracy. Due to PLS being used for feature weighting, they were the most computationally efficient methods along with PLS-VIP and PLS-BETA. The SVM classifier was second in terms of classification accuracy, but lead to a large signature of 640 genes. RR achieved a considerable 88% accuracy but resulted in the largest signature of all methods, 1372 genes. Compared to RR, LASSO sacrificed a small amount of accuracy reaching 86.4% but resulted in a small signature of 136 genes. PLS reached accuracies of 82.2% for 18 genes, 83.6% for 88 genes and 87.6% for 825 genes selected in the signature in the case of PLS-VIP, while PLS BETA yielded similar results. I-RELIEF proved to be computationally expensive and was evaluated using bootstrap datasets of smaller size. Even on the reduced dataset, I-RELIEF was 1000 times slower than PLS methods. In the case of K=5 it reached 79.6% accuracy

for only 15 genes, while results were similar for K=5. The low classification accuracy may be a result of the smaller size of the training set. However, the very long execution time made the I-RELIEF method practically unusable on a full size dataset.

The statistical significance of the classification accuracy of each method was assessed using permutation tests to calculate the corresponding p-value. According to the results, the classification accuracy observed was significant for all methods, leading to a p-values smaller than 0.05. The common genes of all signatures were then extracted , resulting into three “common gene” signatures of 77, 16 and 5 genes, according to different values being used for the VIP score threshold of the PLS methods. The accuracy of the “common gene” signatures was evaluated using a 3-NN, as well as a SVM classifier. The 3-NN classifier performed best, leading to 95.2%, 92.8% and 88% accuracy for the 77, 16 and 5 “common gene” signatures, respectively. All observed accuracies were proved to be statistically significant, using permutation tests. Finally, the performance of the “common gene” signatures was better but comparable to that of random signatures of the same size, a phenomenon also observed in [4] and [5].

Moreover, the consistency of the classification accuracy achieved by the “common gene” signature was assessed by training a classifier on multiple bootstrap training sets while using the same bootstrap test set for evaluation. The classification methods used were 3-NN and SVM, while both methods lead to consistent results. The “77 common gene” signature lead to the most consistent performance of the 3-NN classifier, reaching 95% with 0.025 standard deviation. On the other hand, the SVM classifier reached the most consistent 81% accuracy with 0.009 standard deviation, using the “5 common gene” signature.

Next, the performance of SBV was compared to that of standard 10-Fold Cross Validation (CV) across a set of different types of FSS & classification methods: LASSO, RR, SVM, PLS K-NN on the same breast cancer dataset. It is observed that even though the derived results are comparable, SBV generally leads to smaller signatures that have more genes in common compared to those extracted by 10-Fold CV. As such, the signatures extracted by SBV reflect the biological model to a greater extent and include less random noise.

Finally, according to the biological evaluation the unified “77 common-gene signature” that was derived by our proposed model paying attention to statistical significance, stability and repeatability (using SBV) is highly associated with breast cancer and its clinical features, suggesting that it could be potentially viewed as promising clinical signature.

References

[1] C. A. Davis, F. Gerick, V. Hintermair, C. C. Friedel, K. Fundel, R. Küffner, R. Zimmer, "Reliable gene signatures for microarray classification: assessment of stability and performance," Bioinformatics., vol. 22, no. 19, pp. 2356–2363, 2006.

[2] S. Y. Neo, C. K. Leow, V. B. Vega, P. M. Long, A. F.M. Islam, P. B.S. Lai, E. T. Liu, E. C. Ren, "Identification of Discriminators of Hepatoma by Gene Expression Profiling Using a Minimal Dataset Approach," HEPATOLOGY., vol. 39, pp. 944-953, 2004.

[3] I. Suzuki, T. Takenouchi, M. Ohira, S. Oba, S. Ishii, "Robust Model Selection for Classification of Microarrays," Cancer Informatics., vol.7, pp. 141–157, 2009.

[4] A.C. Haury, P. Gestraud, J.P. Vert , "The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures,", PLoS ONE 6(12): e28210. doi:10.1371/journal.pone.0028210, 2011.

[5] L. Ein-Dor, I. Kela, G. Getz, D. Givol, E. Domany, "Outcome signature genes in breast cancer: is there a unique set?," Bioinformatics., vol. 21, no. 2, pp. 171–178, 2005

[6] R. Armañanzas, I. Inza, P. Larrañaga , "Detecting reliable gene interactions by a hierarchy of Bayesian network classifiers," Computer Methods and Programs in Biomedicine., vol. 91, pp.110-121, 2008.

[7] A. García-Bilbao, R. Armañanzas, Z. Ispizua, B. Calvo, A. Alonso-Varona, I. Inza, P. Larrañaga, G. López-Vivanco, B.Suárez-Merino, M. Betanzos, "Identification of a biomarker panel for colorectar cancer diagnosis," BMC Cancer 2012, 12:43

[8] A. Barrier, P. Boelle, F. Roser, J. Gregg, C. Tse, D. Brault, F. Lacaine,
S. Houry, M. Huguier, B. Franc, A. Flahault, A. Lemoine, S. Dudoit, "Stage II Colon Cancer Prognosis Prediction by Tumor Gene Expression Profiling," Journal of Clinical Oncology., vol. 24, no. 29, pp. 4665-4691, 2006.

[9] B.Efron, "Bootstrap Methods: Another Look at the Jackknife," The Annals of Statistics., vol. 7, no. 1, pp. 1-26, 1979.

[10] M. Kathleen Kerr, G. A. Churchill, "Bootstrapping cluster analysis: Assessing the
reliability of conclusions from microarray experiments," PNAS., vol. 98, no. 16, pp. 8961-8965, 2001.

[11] N. Friedman, M. Goldszmidt, A. Wyner, "Data Analysis with Bayesian Networks: A Bootstrap Approach," ., Proc. Fifteenth Conf. on Uncertainty in Artificial Intelligence (UAI), 1999., 1999

[12] National Human Genome Research Institute, http://www.genome.gov/

[13] Danh V. Nguyen, A. Bulak Arpat, Naisyin Wang, Raymond J. Carroll, "DNA Microarray Experiments: Biological and Technological Aspects," BIOMETRICS., vol. 58, pp. 701-717, 2002.

[14] Musa H. Asyali, Dilek Colak, Omer Demirkaya, Mehmet S. Inan, "Gene Expression Profile Classification: A Review," Current Bioinformatics., vol. 1, no. 1, pp. 55-73, 2006.

[15] Christopher M. Bishop, "Pattern Recognition and Machine Learning," Springer, 2006.

[16] M. E. Blazadonakis , M. Zervakis, "The linear neuron as marker selector and clinical predictor in cancer gene analysis," Computer methods and programs in biomedicine., vol. 91, pp. 22–35, 2008.

[17] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics," Bioinformatics., vol. 23, no. 19, pp. 2507–2517, 2007. doi:10.1093/bioinformatics/btm344

[18] I. Guyon, A. Elisseeff, "An Introduction to Variable and Feature Selection," Journal of Machine Learning Research., vol. 3, pp. 1157-1182, 2003.

[19] R. Maglietta, A. D'Addabbo, A. Piepoli, F. Perri, S. Liuni, G. Pesole, N. Ancona, "Selection of relevant genes in cancer diagnosis based on their prediction accuracy," Artificial Intelligence in Medicine., vol. 40, pp. 29-44, 2007.

[20] C. Saunders, A. Gammerman, V. Vovk, "Ridge Regression Learning Algorithm in Dual Variables," Royal Holloway, University of London Proceedings of the 15th International Conference on Machine Learning, ICML '98, 1998.

[21] R. Tibshirani "Regression, Shrinkage and Selection via the Lasso," Journal of the Royal Statistical Society. Series B (Methodological) vol. 58, no. 1, pp. 267-288, 1996.

[22] I.G. Chong, C. H. Jun, "Performance of some variable selection methods when multicollinearity is present," Chemometrics and Intelligent Laboratory Systems., vol 78, pp. 103-112, 2005.

[23] Svante Wold a, Michael Sjostrom, Lennart Eriksson, "PLS-regression: a basic tool of chemometrics," Chemometrics and Intelligent Laboratory Systems., vol. 58, pp. 109-130, 2001.

[24] P. Geladi, B.R. Kowalski, "Partial Least-Squares Regression: A Tutorial," Analytica Chimica Acta., vol. 185, pp. 1-17, 1986.

[25] C. Cortes, V. Vapnik,, "Support-Vector Networks," Machine Learning., vol. 20, pp. 273-297, 1995.

[26] K. Kira, L. A. Rendell, "A Pratical Approach to Feature Selection," the 9th International Conference on Machine Learning, pp. 249-256, Morgan Kaufmann.

[27] Yijun Sun, Jian Li, "Iterative RELIEF for Feature Weighting," IEEE Transactions on Pattern Analysis and Machine Intelligence., vol. 29, no. 6, pp. 1035-1051, 2007.

[28] Gilad-Bachrach, R., Navot, A., & Tishby, N. (2004). Margin based feature selection - theory and algorithms. the 21st International Conference on Machine Learning.

[29] "Wolfram MathWorld - Weak Law of Large Numbers"
http://mathworld.wolfram.com/WeakLawofLargeNumbers.html

[30] Athanasios Papoulis, S. Unnikrishna Pillai, "Probability, Random Variables and Stochastic Processes", McGraw-Hill Education.

[31] V.G. Tusher, R. Tibshirani, G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response,"
Proceedings of the National Academy of Sciences., vol. 98, no. 9, pp. 5116–5121,
2001. doi:10.1073/pnas.091062498

[32] Sayan Mukherjee, Polina Golland, Dmitry Panchenko, "Permutation Tests for Classification," AI Memo 2003-019, 2003.

[33] Mosca E, Alfieri R, Merelli I, Viti F, Calabria A, Milanesi L. A multilevel data integration resource for breast cancer study. BMC Syst Biol. 2010 Jun 3; 4:76.

[34] Wang J, Duncan D, Shi Z, Zhang B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Res. 2013 Jul; 41(Web Server issue):W77-83.

[35] Zhang F, Drabier R. IPAD: the Integrated Pathway Analysis Database for Systematic Enrichment Analysis. BMC Bioinformatics. 2012; 13 Suppl 15:S7.

[36] Rowinsky EK. Signal events: Cell signal transduction and its inhibition in cancer. Oncologist. 2003; 8 Suppl 3:5-17.

[37] Lu P, Weaver VM, Werb Z. The extracellular matrix: a dynamic niche in cancer progression. J Cell Biol. 2012 Feb 20; 196(4):395-406.

[38] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25; 102(43):15545-50.

[39] Lee EY, Muller WJ. Oncogenes and tumor suppressor genes. Cold Spring Harb Perspect Biol. 2010 Oct; 2(10):a003236.

[40] Abate-Shen C. Deregulated homeobox gene expression in cancer: cause or consequence? Nat Rev Cancer. 2002 Oct; 2 (10):777-85.

[41] Zola H. Medical applications of leukocyte surface molecules--the CD molecules. Mol Med. 2006 Nov-Dec; 12(11-12):312-6.

[42] Zola H, Swart B, Banham A, Barry S, Beare A, Bensussan A, Boumsell L, D Buckley C, Bühring HJ, Clark G, Engel P, Fox D, Jin BQ, Macardle PJ, Malavasi F, Mason D, Stockinger H, Yang X. CD molecules 2006--human cell differentiation molecules. J Immunol Methods. 2007 Jan 30; 319(1-2):1-5. Epub 2006 Dec 4.

[43] Moreland RT, Ryan JF, Pan C, Baxevanis AD. The Homeodomain Resource: a comprehensive collection of sequence, structure, interaction, genomic and functional information on the homeodomain protein family. Database (Oxford). 2009; 2009:bap004.

[44] Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. Nat Rev Cancer. 2004 Mar;4(3):177-83.

[45] Richardson CJ, Gao Q, Mitsopoulous C, Zvelebil M, Pearl LH, Pearl FM. MoKCa database--mutations of kinases in cancer. Nucleic Acids Res. 2009 Jan; 37(Database issue):D824-31.

[46] Petretti C, Prigent C. The Protein Kinase Resource: everything you always wanted to know about protein kinases but were afraid to ask. Biol Cell. 2005 Feb;97(2):113-8.

[47] Benz CC.Transcription factors and breast cancer. Endocrine-Related Cancer. 1998; 5:271-282.

[48] Segal E, Friedman N, Kaminski N, Regev A, Koller D. From signatures to models: understanding cancer using microarrays. Nat Genet. 2005 Jun; 37 Suppl:S38-45.

[49] Wu JM, Fackler MJ, Halushka MK, Molavi DW, Taylor ME, Teo WW, Griffin C, Fetting J, Davidson NE, De Marzo AM, Hicks JL, Chitale D, Ladanyi M, Sukumar S, Argani P. Heterogeneity of breast cancer metastases: comparison of therapeutic target expression and promoter methylation between primary tumors and their multifocal metastases. Clin Cancer Res. 2008 Apr 1; 14(7):1938-46.

[50] Grant GM, Fortney A, Gorreta F, Estep M, Del Giacco L, Van Meter A, Christensen A, Appalla L, Naouar C, Jamison C, Al-Timimi A, Donovan J, Cooper J, Garrett C, Chandhoke V. Microarrays in cancer research. Anticancer Res. 2004 Mar-Apr; 24(2A):441-8.

[51] Wozniak MA, Modzelewska K, Kwong L, Keely PJ. Focal adhesion regulation of cell behavior. Biochim Biophys Acta. 2004 Jul 5; 1692(2-3):103-19.

[52] Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000 Jan 7; 100(1):57-70.

[53] Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011 Mar 4; 144(5):646-74.

[54] Wirapati P, Sotiriou C, Kunkel S, Farmer P, Pradervand S, Haibe-Kains B, Desmedt C, Ignatiadis M, Sengstag T, Schütz F, Goldstein DR, Piccart M, Delorenzi M. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. Breast Cancer Res. 2008;10(4):R65.

[55] Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DS, Nobel AB, van't Veer LJ, Perou CM. Concordance among gene-expression-based predictors for breast cancer. N Engl J Med. 2006 Aug 10; 355(6):560-9.

[56] Nicolini A, Carpi A, Rossi G. Cytokines in breast cancer. Cytokine Growth Factor Rev. 2006 Oct; 17(5):325-37.

[57] Cassier PA, Treilleux I, Bachelot T, Ray-Coquard I, Bendriss-Vermare N, Ménétrier-Caux C, Trédan O, Goddard-Léon S, Pin JJ, Mignotte H, Bathélémy-Dubois C, Caux C, Lebecque S, Blay JY. Prognostic value of the expression of C-Chemokine Receptor 6 and 7 and their ligands in non-metastatic breast cancer. BMC Cancer. 2011 May 30; 11:213.

[58] Hung JH, “Gene Set/Pathway enrichment analysis.,” Methods Mol Biol. 2013;939:201-13. doi: 10.1007/978-1-62703-107-3_13.

[59] Dennis Kostka , Rainer Spang , “Microarray Based Diagnosis Profits from Better Documentation of Gene Expression Signatures ,” PLoS Comput Biol 4(2): e22. doi:10.1371/ journal.pcbi.0040022 , 2008.

APPENDIX A. SUPPLEMENTARY TABLE I. GENE LIST - DESCRIPTION		
gene id	**gene symbol**	**description**
1029	CDKN2A	cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)
10331	B3GNT3	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 3
10381	TUBB3	tubulin, beta 3
1047	CLGN	calmegin
10481	HOXB13	homeobox B13
10562	OLFM4	olfactomedin 4
10568	SLC34A2	solute carrier family 34 (sodium phosphate), member 2
10648	SCGB1D1	secretoglobin, family 1D, member 1
10891	PPARGC1A	peroxisome proliferator-activated receptor gamma, coactivator 1 alpha
11005	SPINK5	serine peptidase inhibitor, Kazal type 5
11012	KLK11	kallikrein-related peptidase 11
1118	CHIT1	chitinase 1 (chitotriosidase)
11197	WIF1	WNT inhibitory factor 1
1301	COL11A1	collagen, type XI, alpha 1
1308	COL17A1	collagen, type XVII, alpha 1
1311	COMP	cartilage oligomeric matrix protein
1359	CPA3	carboxypeptidase A3 (mast cell)
1475	CSTA	cystatin A (stefin A)
1811	SLC26A3	solute carrier family 26, member 3
1907	EDN2	endothelin 2
1908	EDN3	endothelin 3
1950	EGF	epidermal growth factor (beta-urogastrone)
2173	FABP7	fatty acid binding protein 7, brain
2261	FGFR3	fibroblast growth factor receptor 3
2302	FOXJ1	forkhead box J1
2335	FN1	fibronectin 1
23532	PRAME	preferentially expressed antigen in melanoma
26585	GREM1	gremlin 1, cysteine knot superfamily, homolog (Xenopus laevis)
27074	LAMP3	lysosomal-associated membrane protein 3
2938	GSTA1	glutathione S-transferase alpha 1
2940	GSTA3	glutathione S-transferase alpha 3
29842	TFCP2L1	transcription factor CP2-like 1
3084	NRG1	neuregulin 1
3164	NR4A1	nuclear receptor subfamily 4, group A, member 1
APPENDIX A. SUPPLEMENTARY TABLE I. GENE LIST - DESCRIPTION		
3294	HSD17B2	hydroxysteroid (17-beta) dehydrogenase 2
347902	AMIGO2	adhesion molecule with Ig-like domain 2
3500	IGHG1	immunoglobulin heavy constant gamma 1 (G1m marker)
3773	KCNJ16	potassium inwardly-rectifying channel, subfamily J, member 16
3851	KRT4	keratin 4
3868	KRT16	keratin 16
4069	LYZ	lysozyme (renal amyloidosis)
430	ASCL2	achaete-scute complex homolog 2 (Drosophila)
4321	MMP12	matrix metallopeptidase 12 (macrophage elastase)
4435	CITED1	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 1
4477	MSMB	microseminoprotein, beta-
4916	NTRK3	neurotrophic tyrosine kinase, receptor, type 3
4969	OGN	osteoglycin
50617	ATP6V0A4	ATPase, H+ transporting, lysosomal V0 subunit a4
51442	VGLL1	vestigial like 1 (Drosophila)
5179	PENK	proenkephalin
5304	PIP	prolactin-induced protein
54829	ASPN	asporin
55273	TMEM100	transmembrane protein 100
55713	ZNF334	zinc finger protein 334

57348	TTYH1	tweety homolog 1 (Drosophila)
57586	SYT13	synaptotagmin XIII
58	ACTA1	actin, alpha 1, skeletal muscle
6278	S100A7	S100 calcium binding protein A7
6280	S100A9	S100 calcium binding protein A9
6286	S100P	S100 calcium binding protein P
6362	CCL18	chemokine (C-C motif) ligand 18 (pulmonary and activation-regulated)
6363	CCL19	chemokine (C-C motif) ligand 19
6495	SIX1	SIX homeobox 1
6505	SLC1A1	solute carrier family 1 (neuronal/epithelial high affinity glutamate transporter, system Xag), member 1
6664	SOX11	SRY (sex determining region Y)-box 11
6898	TAT	tyrosine aminotransferase
7031	TFF1	trefoil factor 1
7136	TNNI2	troponin I type 2 (skeletal, fast)
79785	RERGL	RERG/RAS-like
7980	TFPI2	tissue factor pathway inhibitor 2
79919	C2orf54	chromosome 2 open reading frame 54
APPENDIX A. SUPPLEMENTARY TABLE I. GENE LIST - DESCRIPTION		
8483	CILP	cartilage intermediate layer protein, nucleotide pyrophosphohydrolase
8788	DLK1	delta-like 1 homolog (Drosophila)
8817	FGF18	fibroblast growth factor 18
9073	CLDN8	claudin 8
9185	REPS2	RALBP1 associated Eps domain containing 2
930	CD19	CD19 molecule

GENENAME: associated with breast cancer

GENENAME: not associated yet with breast cancer

KEGG PATHWAYS
'77 Common-Gene Signature' (p≤0.05)

APPENDIX A. SUPPLEMENTARY TABLE II			
KEGG Pathways	**P**	**adjP**	**Genes participated in KEGG Pathways**
1.Protein digestion and absorption	1.41e-05	0.0004	COL11A1 | COL17A1 | CPA3 | SLC1A1
2.Bladder cancer	5.97e-05	0.0007	FGFR3 | CDKN2A | EGF
3.Melanoma	0.0003	0.0019	CDKN2A | FGF18 | EGF
4.Pathways in cancer	0.0003	0.0019	FGFR3 | FN1 | CDKN2A | FGF18 | EGF
5.ECM-receptor interaction	0.0005	0.0021	COL11A1 | FN1 | COMP
6.Focal adhesion	0.0005	0.0021	COL11A1 | FN1 | COMP | EGF
7.Regulation of actin cytoskeleton	0.0006	0.0021	FGFR3 | FN1 | FGF18 | EGF
8.**MAPK signaling pathway**	0.0014	0.0044	FGFR3 | NR4A1 | FGF18 | EGF
9.Phagosome	0.0026	0.0072	COMP | TUBB3 | ATP6V0A4
10.Glutathione metabolism	0.0036	0.0090	GSTA3 | GSTA1
11.Non-small cell lung cancer	0.0042	0.0095	CDKN2A | EGF
12.Pancreatic cancer	0.0070	0.0127	CDKN2A | EGF
13.Glioma	0.0061	0.0127	CDKN2A | EGF
14.Drug metabolism - cytochrome P450	0.0076	0.0127	GSTA3 | GSTA1
15.Metabolism of xenobiotics by cytochrome P450	0.0072	0.0127	GSTA3 | GSTA1
16.**ErbB signaling pathway**	0.0107	0.0167	NRG1 | EGF
17.Cytokine-cytokine receptor interaction	0.0120	0.0167	CCL19 | CCL18 | EGF
18.Gap junction	0.0114	0.0167	EGF | TUBB3
19.Pancreatic secretion	0.0142	0.0187	SLC26A3 | CPA3
20.Amoebiasis	0.0155	0.0194	COL11A1 | FN1
21.Lysosome	0.0199	0.0237	LAMP3 | ATP6V0A4
22.Hepatitis C	0.0241	0.0274	CLDN8 | EGF
23.Chemokine signaling pathway	0.0451	0.0490	CCL19 | CCL18

GENE ONTOLOGY ENRICHMENT ANALYSIS in terms of BIOLOGICAL PROCESS

'77 Common-Gene Signature' (p≤0.05)

APPENDIX A. SUPPLEMENTARY TABLE III

Biological Process	P	adjP	Genes participated in Biological Process
1.epithelial cell differentiation	4.72e-08	4.96e-05	TFCP2L1 | GREM1 | FGFR3 | S100A7 | CITED1 | SIX1 | FOXJ1 | KRT4 | NRG1 | HOXB13 | CSTA | SPINK5
2.tissue development	5.02e-07	0.0002	MMP12 | TFCP2L1 | GREM1 | COL17A1 | EDN3 | KRT4 | COL11A1 | SPINK5 | CSTA | KRT16 | FGFR3 | S100A7 | CITED1 | ASPN | SIX1 | FOXJ1 | FGF18 | ACTA1 | NRG1 | HOXB13 | COMP | SOX11 | CDKN2A
3.epithelium development	3.32e-07	0.0002	MMP12 | TFCP2L1 | GREM1 | FGFR3 | S100A7 | CITED1 | SIX1 | FOXJ1 | KRT4 | NRG1 | HOXB13 | CSTA | SPINK5 | SOX11 | CDKN2A
4.cell proliferation	6.07e-06	0.0013	MMP12 | GREM1 | PRAME | EDN3 | CCL19 | NR4A1 | KRT4 | EGF | KRT16 | FABP7 | EDN2 | FGFR3 | CITED1 | SIX1 | FOXJ1 | FGF18 | OGN | ASCL2 | NRG1 | SOX11 | CDKN2A | LAMP3
5.regulation of cell proliferation	5.13e-06	0.0013	MMP12 | GREM1 | PRAME | EDN3 | CCL19 | NR4A1 | KRT4 | EGF | FABP7 | EDN2 | FGFR3 | SIX1 | FOXJ1 | FGF18 | OGN | ASCL2 | NRG1 | SOX11 | CDKN2A
6.cell chemotaxis	8.39e-06	0.0015	GREM1 | EDN2 | S100A7 | EDN3 | CCL19 | S100A9 | NR4A1
7.regulation of leukocyte chemotaxis	1.15e-05	0.0017	GREM1 | EDN2 | S100A7 | EDN3 | CCL19
8.negative regulation of gliogenesis	1.80e-05	0.0021	FGFR3 | NTRK3 | SOX11 | ASCL2
9.negative regulation of developmental process	1.74e-05	0.0021	GREM1 | FGFR3 | PRAME | CITED1 | SIX1 | ASPN | FOXJ1 | ASCL2 | NTRK3 | SPINK5 | SOX11 | CDKN2A
10.circulatory system development	2.47e-05	0.0022	GREM1 | S100A7 | CITED1 | SIX1 | FN1 | FOXJ1 | NR4A1 | FGF18 | EGF | NRG1 | HOXB13 | COL11A1 | SPINK5 | SOX11
11.cardiovascular system development	2.47e-05	0.0022	GREM1 | S100A7 | CITED1 | SIX1 | FN1 | FOXJ1 | NR4A1 | FGF18 | EGF | NRG1 | HOXB13 | COL11A1 | SPINK5 | SOX11
12.positive regulation of cell proliferation	2.52e-05	0.0022	MMP12 | GREM1 | FGFR3 | EDN2 | PRAME | SIX1 | EDN3 | CCL19 | NR4A1 | FGF18 | EGF | NRG1 | SOX11
13.leukocyte chemotaxis	2.81e-05	0.0023	GREM1 | EDN2 | S100A7 | EDN3 | CCL19 | S100A9
14.skeletal system morphogenesis	3.59e-05	0.0025	GREM1 | FGFR3 | SIX1 | COL11A1 | COMP | SOX11 | FGF18
15.leukocyte migration	3.86e-05	0.0025	GREM1 | EDN2 | S100A7 | FN1 | EDN3 | CCL19 | FOXJ1 | S100A9
16.anatomical structure formation involved in morphogenesis	3.36e-05	0.0025	GREM1 | FN1 | NR4A1 | EGF | NTRK3 | COL11A1 | SPINK5 | FGFR3 | S100A7 | ASPN | SIX1 | CITED1 | FOXJ1 | FGF18 | ACTA1 | NRG1 | HOXB13 | COMP | SOX11 | CDKN2A | TUBB3
17.single-multicellular organism process	4.52e-05	0.0028	TFCP2L1 | WIF1 | FN1 | EDN3 | CCL19 | CLGN | EGF | NTRK3 | COL11A1 | SPINK5 | CSTA | SLC1A1 | SLC34A2 | KRT16 | FABP7 | FGFR3 | SIX1 | ASPN | ATP6V0A4 | ASCL2 | S100A9 | COMP | SOX11 | KCNJ16 | TUBB3 | DLK1 | GREM1 | TFPI2 | TNNI2 | COL17A1 | CPA3 | NR4A1 | HSD17B2 | PENK | TFF1 | SLC26A3 | EDN2 | S100A7 | CITED1 | FOXJ1 | FGF18 | ACTA1 | NRG1 | HOXB13 | PPARGC1A | CDKN2A
18.multicellular organismal process	5.33e-05	0.0031	TFCP2L1 | WIF1 | FN1 | EDN3 | CCL19 | CLGN | EGF | NTRK3 | COL11A1 | SPINK5 | CSTA | SLC1A1 | SLC34A2 | KRT16 | FABP7 | FGFR3 | SIX1 | ASPN | ATP6V0A4 | ASCL2 | S100A9 | COMP | SOX11 | KCNJ16 | TUBB3 | DLK1 | GREM1 | TFPI2 | TNNI2 | COL17A1 | CPA3 | NR4A1 | HSD17B2 | PENK | TFF1 | SLC26A3 | EDN2 | S100A7 | CITED1 | FOXJ1 | FGF18 | ACTA1 | NRG1 | HOXB13 | PPARGC1A | CDKN2A
19.positive regulation of response to external stimulus	0.0001	0.0032	EDN2 | NTRK3 | S100A7 | EDN3 | CCL19 | S100A9
20.kidney morphogenesis	0.0001	0.0032	GREM1 | CITED1 | SIX1 | FOXJ1

APPENDIX A. SUPPLEMENTARY TABLE III

Biological Process	P	adjP	Genes participated in Biological Process
21.positive regulation of leukocyte chemotaxis	0.0001	0.0032	EDN2 | S100A7 | EDN3 | CCL19
22.cellular component movement	8.41e-05	0.0032	MMP12 | GREM1 | TNNI2 | FN1 | EDN3 | CCL19 | NR4A1 | NTRK3 | EDN2 | SIX1 | S100A7 | FOXJ1 | ACTA1 | S100P | NRG1 | S100A9 | TUBB3
23.blood vessel morphogenesis	6.90e-05	0.0032	GREM1 | S100A7 | CITED1 | SIX1 | FN1 | NR4A1 | FGF18 | EGF | HOXB13 | SPINK5

24.organ development	9.14e-05	0.0032	TFCP2L1 | GREM1 | COL17A1 | EDN3 | CCL19 | EGF | NTRK3 | COL11A1 | SPINK5 | CSTA | KRT16 | FABP7 | FGFR3 | EDN2 | S100A7 | CITED1 | SIX1 | ASPN | FOXJ1 | FGF18 | ACTA1 | ASCL2 | NRG1 | HOXB13 | COMP | SOX11 | CDKN2A
25.cell development	0.0001	0.0032	TFCP2L1 | GREM1 | FN1 | EDN3 | CCL19 | NTRK3 | COL11A1 | FGFR3 | SIX1 | CITED1 | FOXJ1 | ACTA1 | FGF18 | ASCL2 | NRG1 | HOXB13 | SOX11 | CDKN2A | TUBB3
26.ossification	9.52e-05	0.0032	GREM1 | FGFR3 | CITED1 | ASPN | COL11A1 | SOX11 | FGF18 | ATP6V0A4
27.cell migration	8.13e-05	0.0032	MMP12 | GREM1 | EDN2 | S100A7 | SIX1 | FN1 | EDN3 | CCL19 | FOXJ1 | NR4A1 | S100P | NTRK3 | NRG1 | S100A9
28.neutrophil chemotaxis	0.0001	0.0032	EDN2 | EDN3 | CCL19 | S100A9
29.tissue morphogenesis	0.0001	0.0032	MMP12 | GREM1 | FGFR3 | CITED1 | SIX1 | FOXJ1 | NRG1 | HOXB13 | COL11A1 | SOX11
30.regulation of developmental process	8.37e-05	0.0032	GREM1 | PRAME | WIF1 | FN1 | EDN3 | CCL19 | EGF | NTRK3 | SPINK5 | FGFR3 | ASPN | SIX1 | CITED1 | FOXJ1 | FGF18 | ASCL2 | NRG1 | SOX11 | CDKN2A
31.bone development	0.0001	0.0032	GREM1 | FGFR3 | COMP | FGF18 | HSD17B2
32.epithelial cell development	6.21e-05	0.0032	TFCP2L1 | GREM1 | HOXB13 | CITED1 | FOXJ1
33.regulation of leukocyte migration	6.58e-05	0.0032	GREM1 | EDN2 | S100A7 | EDN3 | CCL19
34.cell differentiation	0.0002	0.0044	TFCP2L1 | GREM1 | PRAME | WIF1 | FN1 | EDN3 | CCL19 | KRT4 | NTRK3 | COL11A1 | SPINK5 | CSTA | FABP7 | FGFR3 | S100A7 | CITED1 | SIX1 | FOXJ1 | FGF18 | ACTA1 | ASCL2 | NRG1 | HOXB13 | PPARGC1A | SOX11 | CDKN2A | TUBB3
35.growth	0.0002	0.0044	TFCP2L1 | GREM1 | FGFR3 | PRAME | SIX1 | ACTA1 | NTRK3 | NRG1 | HOXB13 | S100A9 | COMP | SPINK5 | CDKN2A

APPENDIX A. SUPPLEMENTARY TABLE III

Biological Process	**P**	**adjP**	**Genes participated in Biological Process**
36.blood vessel development	0.0002	0.0044	GREM1 | S100A7 | CITED1 | SIX1 | FN1 | NR4A1 | FGF18 | EGF | HOXB13 | SPINK5
37.regulation of chemotaxis	0.0002	0.0044	GREM1 | EDN2 | S100A7 | EDN3 | CCL19
38.negative regulation of leukocyte proliferation	0.0002	0.0044	GREM1 | FOXJ1 | SOX11 | CDKN2A
39.anatomical structure morphogenesis	0.0002	0.0044	MMP12 | TFCP2L1 | GREM1 | FN1 | NR4A1 | EGF | NTRK3 | COL11A1 | SPINK5 | FGFR3 | S100A7 | CITED1 | ASPN | SIX1 | FOXJ1 | FGF18 | ACTA1 | NRG1 | HOXB13 | COMP | SOX11 | CDKN2A | TUBB3
40.regulation of cell differentiation	0.0002	0.0044	GREM1 | FGFR3 | PRAME | CITED1 | SIX1 | WIF1 | CCL19 | FOXJ1 | FGF18 | ASCL2 | NTRK3 | NRG1 | SPINK5 | SOX11 | CDKN2A

DISEASE ASSOCIATION ANALYSIS
'77 Common-Gene Signature' (p≤0.05)

APPENDIX A. SUPPLEMENTARY TABLE IV			
DISEASE	**P**	**adjP**	**Genes involved in Disease**
1.Neoplasms	1.01e-14	2.83e-12	PRAME | TFPI2 | WIF1 | EGF | HSD17B2 | MSMB | KLK11 | FABP 7 | TFF1 | FGFR3 | OLFM4 | S100A7 | S100P | HOXB13 | SOX11 | C DKN2A | TUBB3 | GSTA1
2.Skin Diseases	5.63e-13	7.88e-11	TFF1 | FABP7 | FGFR3 | S100A7 | COL17A1 | EGF | HSD17B2 | NR G1 | HOXB13 | SPINK5 | CDKN2A | KRT16 | PIP
3.cancer or viral infections	8.51e-13	7.94e-11	PRAME | TFPI2 | WIF1 | EGF | HSD17B2 | MSMB | KLK11 | FABP 7 | TFF1 | FGFR3 | OLFM4 | CITED1 | S100A7 | S100P | HOXB13 | CDKN2A | TUBB3
4.Skin and Connective Tissue Diseases	3.38e-12	2.37e-10	TFF1 | FABP7 | FGFR3 | S100A7 | COL17A1 | HSD17B2 | HOXB13 | SPINK5 | S100A9 | CCL18 | CDKN2A | KRT16 | PIP
5.Breast Neoplasms	7.86e-11	4.40e-09	TFF1 | FABP7 | S100A7 | SIX1 | EGF | HSD17B2 | NRG1 | HOXB13 | PIP | KLK11 | GSTA1
6.Carcinoma	2.39e-09	1.12e-07	TFF1 | FGFR3 | TFPI2 | S100A7 | SIX1 | WIF1 | EGF | S100P | CDK N2A | KLK11 | TUBB3
7.Neoplastic Processes	3.44e-09	1.38e-07	FABP7 | FGFR3 | PRAME | TFPI2 | SIX1 | WIF1 | FN1 | EGF | S100 P | CDKN2A
8.Keratosis	4.85e-09	1.70e-07	TAT | FGFR3 | SPINK5 | CSTA | CDKN2A | KRT16
9.Urogenital Neoplasms	5.53e-09	1.72e-07	FGFR3 | WIF1 | REPS2 | EGF | HSD17B2 | HOXB13 | CDKN2A | M SMB | KLK11 | TUBB3
10.Breast Diseases	1.37e-08	3.84e-07	TFF1 | FABP7 | S100A7 | EGF | HSD17B2 | NRG1 | HOXB13 | PIP | GSTA1
11.Nevus	4.27e-08	9.58e-07	FABP7 | FGFR3 | CITED1 | KRT4 | CDKN2A
12.Urologic Neoplasms	4.45e-08	9.58e-07	FABP7 | FGFR3 | EDN2 | CITED1 | WIF1 | CDKN2A | DLK1
13.Growth Disorders	4.13e-08	9.58e-07	FGFR3 | NRG1 | NTRK3 | CILP | COMP | SLC1A1 | EGF
14.Gastrointestinal Diseases	5.64e-08	1.13e-06	TFF1 | SLC26A3 | TFPI2 | OLFM4 | EDN3 | EGF | S100A9 | CDKN2 A | GSTA1
15.Brain Neoplasms	7.08e-08	1.32e-06	FABP7 | NTRK3 | SOX11 | CDKN2A | EGF | DLK1 | KLK11
16.Cartilage Diseases	8.04e-08	1.41e-06	FGFR3 | CILP | ASPN | COL11A1 | COMP
17. Skin Diseases, Genetic	8.91e-08	1.47e-06	FGFR3 | COL17A1 | HSD17B2 | HOXB13 | SPINK5 | KRT16 | CDK N2A | GSTA1
18.Neuroectodermal Tumors	9.59e-08	1.49e-06	FABP7 | PRAME | TFPI2 | CITED1 | EGF | NTRK3 | CDKN2A | DL K1
19.Disease Progression	1.45e-07	2.14e-06	TFF1 | MMP12 | FGFR3 | PRAME | S100A7 | CDKN2A | EGF
20.Genetic Predisposition to Disease	2.03e-07	2.84e-06	MMP12 | ASPN | EGF | HSD17B2 | NRG1 | PPARGC1A | SPINK5 | SLC1A1 | CDKN2A | MSMB | GSTA1
21.Skin Neoplasms	2.54e-07	3.39e-06	FABP7 | FGFR3 | S100A7 | CDKN2A | EGF | KLK11
22.Recurrence	5.82e-07	7.41e-06	FABP7 | FGFR3 | PRAME | HOXB13 | CDKN2A | MSMB
23.Central Nervous System Neoplasms	7.15e-07	8.39e-06	FABP7 | NTRK3 | SOX11 | CDKN2A | EGF | DLK1
24.Colonic Diseases	7.19e-07	8.39e-06	SLC26A3 | OLFM4 | EDN3 | S100A9 | CDKN2A | GSTA1 | ASCL2
25.Fibrosis	9.93e-07	1.11e-05	SLC26A3 | GREM1 | FN1 | S100A9 | COMP | CCL18
26.Neoplasm of unspecified nature of digestive system	1.36e-06	1.44e-05	TFF1 | TFPI2 | OLFM4 | WIF1 | EGF | S100P | CDKN2A | GSTA1
27.Nervous System Neoplasms	1.39e-06	1.44e-05	FABP7 | NTRK3 | SOX11 | CDKN2A | EGF | DLK1
28.Epithelial cancers	1.64e-06	1.64e-05	TFF1 | S100A7 | SIX1 | KRT4 | CDKN2A | EGF | S100P
29.Musculoskeletal Diseases	1.80e-06	1.74e-05	FGFR3 | TNNI2 | CILP | ASPN | ACTA1 | COL11A1 | COMP | S100 A9
30.Glioma	2.03e-06	1.89e-05	FABP7 | TFPI2 | NTRK3 | SOX11 | CDKN2A | EGF
31.**Disease Susceptibility**	2.15e-06	1.94e-05	MMP12 | ASPN | EGF | HSD17B2 | NRG1 | CHIT1 | SLC1A1 | MSM B | CDKN2A | GSTA1
32.Intestinal Diseases	2.23e-06	1.95e-05	TFF1 | SLC26A3 | OLFM4 | EDN3 | S100A9 | CDKN2A | GSTA1
APPENDIX A. SUPPLEMENTARY TABLE IV			
DISEASE	**P**	**adjP**	**Genes involved in Disease**
33.Gastrointestinal Neoplasms	3.48e-06	2.87e-05	TFF1 | TFPI2 | OLFM4 | S100A9 | CDKN2A | EGF | GSTA1
34.Lung Diseases	3.48e-06	2.87e-05	MMP12 | CHIT1 | SPINK5 | CCL18 | CDKN2A | SLC34A2 | TUBB3
35.Lymphatic Diseases	4.10e-06	3.19e-05	PRAME | CHIT1 | CCL19 | CCL18 | CD19 | SOX11 | CDKN2A
36.Adenocarcinoma	4.02e-06	3.19e-05	TFF1 | FABP7 | OLFM4 | CDKN2A | EGF | TUBB3 | S100P
37.Kidney Neoplasms	4.60e-06	3.48e-05	FABP7 | EDN2 | CITED1 | SIX1 | DLK1
38.Respiratory Tract Diseases	4.89e-06	3.49e-05	MMP12 | CHIT1 | SPINK5 | CCL18 | CDKN2A | SLC34A2 | TUBB3
39.Papulosquamous dermatosis	4.77e-06	3.49e-05	S100A7 | S100A9 | CSTA | KRT16 | LAMP3
40.Lymphoproliferative Disorders	4.98e-06	3.49e-05	FGFR3 | PRAME | CCL19 | CCL18 | CD19 | SOX11 | CDKN2A

DISEASE ASSOCIATION ANALYSIS
'77 Common-Gene Signature' (p≤0.05)

41.Bone Diseases	5.35e-06	3.65e-05	GREM1 | FGFR3 | CILP | ASPN | COL11A1 | COMP
42.Psoriasis	5.69e-06	3.79e-05	S100A7 | S100A9 | CSTA | KRT16 | LAMP3
43.Polyps	6.22e-06	4.05e-05	TFF1 | SLC26A3 | S100A9 | SPINK5
44.Sinusitis	6.61e-06	4.21e-05	S100A7 | FOXJ1 | S100A9 | SPINK5
45.Head and Neck Neoplasms	7.84e-06	4.50e-05	S100A7 | WIF1 | FN1 | KRT4 | CDKN2A | EGF
46.Pathologic Processes	7.47e-06	4.50e-05	MMP12 | FGFR3 | ASPN | EGF | NRG1 | SPINK5 | MSMB | CDKN2A
47.Colorectal Neoplasms	7.51e-06	4.50e-05	SLC26A3 | TFPI2 | OLFM4 | CDKN2A | GSTA1 | ASCL2
48.Osteoarthritis, Knee	7.87e-06	4.50e-05	CILP | ASPN | COL11A1 | COMP
49.Prostatic Neoplasms	7.84e-06	4.50e-05	HOXB13 | REPS2 | MSMB | EGF | KLK11 | HSD17B2
50.Intestinal Neoplasms	8.92e-06	5.00e-05	TFF1 | SLC26A3 | OLFM4 | CDKN2A | GSTA1 | ASCL2
51.Male Urogenital Diseases	1.11e-05	6.09e-05	GREM1 | FGFR3 | HOXB13 | REPS2 | MSMB | KLK11 | ATP6V0A4
52.Pulmonary Fibrosis	1.21e-05	6.52e-05	MMP12 | GREM1 | S100A9 | CCL18
53.Nevus, Pigmented	1.55e-05	8.04e-05	FGFR3 | CITED1 | CDKN2A
54.Halo nevus	1.55e-05	8.04e-05	FGFR3 | CITED1 | CDKN2A
55.Neoplasm Invasiveness	1.63e-05	8.30e-05	MMP12 | FABP7 | TFPI2 | CDKN2A | EGF | S100P
56.Degeneration of lumbar intervertebral disc	1.74e-05	8.70e-05	CILP | ASPN | COL11A1
57.Congenital Abnormalities	1.99e-05	9.38e-05	FGFR3 | TNNI2 | COL17A1 | SIX1 | EDN3 | COL11A1 | SPINK5 | KRT16
58.Stomach Neoplasms	2.01e-05	9.38e-05	TFF1 | OLFM4 | NR4A1 | CDKN2A | KLK11
59.Rhinitis	1.95e-05	9.38e-05	S100A7 | CHIT1 | FOXJ1 | SPINK5
60.Bronchial Diseases	1.92e-05	9.38e-05	MMP12 | CHIT1 | CCL19 | FOXJ1 | SPINK5 | CCL18
61.Hirschsprung Disease	2.37e-05	0.0001	NRG1 | NTRK3 | EDN3
62.Megacolon	2.87e-05	0.0001	NRG1 | NTRK3 | EDN3
63.Lumbar Disc Herniation	2.37e-05	0.0001	CILP | ASPN | COL11A1
64.Immune System Diseases	2.96e-05	0.0001	PRAME | COL17A1 | CCL19 | CD19 | CHIT1 | SPINK5 | SOX11 | CCL18
65.Collagen Diseases	2.75e-05	0.0001	COL17A1 | COL11A1 | FN1 | COMP

APPENDIX A. SUPPLEMENTARY TABLE IV

DISEASE	P	adjP	Genes involved in Disease
66.Lymphoma, Low-Grade	4.73e-05	0.0002	PRAME | CCL19 | CD19 | SOX11 | CDKN2A
67.Ovarian Diseases	4.15e-05	0.0002	CDKN2A | SLC34A2 | TUBB3 | KLK11 | HSD17B2
68.Leukoplakia	3.74e-05	0.0002	S100A7 | KRT4 | CDKN2A
69.Nevi and Melanomas	3.88e-05	0.0002	FABP7 | FGFR3 | PRAME | CITED1 | CDKN2A
70.Lymphoma, B-Cell	6.09e-05	0.0002	PRAME | CCL19 | CD19 | SOX11 | CDKN2A
71.Sarcoidosis	0.0001	0.0003	CHIT1 | S100A9 | CCL18
72.Neoplasms, Squamous Cell	6.74e-05	0.0003	S100A7 | WIF1 | KRT4 | CDKN2A | EGF
73.Inflammation	0.0001	0.0003	MMP12 | EDN2 | S100A7 | CHIT1 | S100A9 | CCL18
74.Pancreatic Diseases	0.0001	0.0003	TFF1 | TFPI2 | CDKN2A | S100P
75.Esophageal Diseases	0.0001	0.0003	WIF1 | KRT4 | CDKN2A | EGF
76.Myosarcoma	0.0001	0.0003	TFF1 | FGFR3 | TNNI2 | CDKN2A
77.Osteoarthritis	0.0001	0.0003	CILP | ASPN | COL11A1 | COMP
78.Joint Diseases	0.0001	0.0003	TNNI2 | CILP | ASPN | S100A9 | COMP
79.Muscle Neoplasms	0.0001	0.0003	TFF1 | FGFR3 | TNNI2 | CDKN2A
80.Nasal Polyps	0.0001	0.0003	S100A7 | S100A9 | SPINK5
81.Mouth Neoplasms	8.47e-05	0.0003	S100A7 | WIF1 | KRT4 | CDKN2A
82.Adenocarcinoma, Mucinous	0.0001	0.0003	TFF1 | CDKN2A | S100P
83.Ichthyosis Vulgaris	0.0001	0.0003	SPINK5 | KRT16
84.Lung Diseases, Interstitial	0.0002	0.0006	CHIT1 | S100A9 | CCL18
85.Wilms Tumor	0.0002	0.0006	CITED1 | SIX1 | DLK1
86.Arthritis	0.0002	0.0006	CILP | ASPN | COL11A1 | S100A9 | COMP
87.Skeletal Dysplasia	0.0002	0.0006	FGFR3 | COL11A1 | COMP
88.Achondroplasia	0.0002	0.0006	FGFR3 | COMP
89.Carcinoma, Transitional Cell	0.0002	0.0006	FGFR3 | WIF1 | CDKN2A
90.Connective Tissue Diseases	0.0002	0.0006	CILP | S100A9 | COMP | CCL18 | CD19
91.Adenoma	0.0002	0.0006	SLC26A3 | WIF1 | CDKN2A | DLK1
92.Fibrosarcoma	0.0003	0.0009	TFPI2 | NTRK3 | FN1
93.Carcinoma, Pancreatic Ductal	0.0003	0.0009	PENK | CDKN2A | S100P
94.Scleroderma, Systemic	0.0003	0.0009	COMP | CCL18 | CD19
95.Esophageal Neoplasms	0.0003	0.0009	WIF1 | KRT4 | CDKN2A | EGF

96.Eating Disorders	0.0003	0.0009	PENK | NTRK3 | SLC1A1 | DLK1
97.Musculoskeletal Abnormalities	0.0003	0.0009	FGFR3 | TNNI2 | COL11A1 | COMP | FGF18
98.Neoplasm Metastasis	0.0003	0.0009	TFPI2 | SIX1 | WIF1 | CDKN2A | S100P
99.Obsessive-Compulsive Disorder	0.0004	0.0011	PENK | NTRK3 | SLC1A1
100.Epidermolysis Bullosa Simplex	0.0004	0.0011	COL17A1 | KRT16
APPENDIX A. SUPPLEMENTARY TABLE IV			
DISEASE	**P**	**adjP**	**Genes involved in Disease**
101. Scoliosis	0.0004	0.0011	COMP | ACTA1 | DLK1
102.Pancreatic Neoplasms	0.0004	0.0011	TFF1 | TFPI2 | CDKN2A | S100P
103.Burkitt Lymphoma	0.0004	0.0011	CCL19 | CD19 | SOX11
104.Epidermodysplasia Verruciformis	0.0005	0.0013	KRT4 | KRT16
105.Precancerous Conditions	0.0005	0.0013	TFF1 | S100A7 | CDKN2A
106.Hearing Disorders	0.0006	0.0015	SIX1 | COL11A1 | OGN | ATP6V0A4
107.Neuroendocrine Tumors	0.0006	0.0015	PRAME | CITED1 | CDKN2A | DLK1
108. Gaucher Disease	0.0006	0.0015	CHIT1 | CCL18
109.Ovarian Neoplasms	0.0006	0.0015	CDKN2A | SLC34A2 | TUBB3 | KLK11
110.Condylomata Acuminata	0.0006	0.0015	S100A7 | CDKN2A
111.Epstein-Barr Virus Infections	0.0007	0.0018	CCL19 | CD19 | CDKN2A
112.Osteochondrodysplasias	0.0008	0.0019	FGFR3 | COL11A1 | COMP
113. Carcinoma, Adenosquamous	0.0008	0.0019	OLFM4 | CDKN2A
114.Ichthyosis, X-Linked	0.0008	0.0019	SPINK5 | KRT16
115.Carcinoma, Papillary	0.0008	0.0019	FGFR3 | FN1 | S100P
116.Epidermolysis Bullosa	0.0008	0.0019	COL17A1 | KRT16
117.Polyhydramnios	0.0008	0.0019	SLC26A3 | FGFR3
118.Carcinoma, Squamous Cell	0.0009	0.0021	S100A7 | WIF1 | CDKN2A | EGF
119.Hypersensitivity	0.0009	0.0021	S100A7 | CHIT1 | SPINK5 | CCL18
120.Melanoma	0.0010	0.0023	FABP7 | PRAME | CITED1 | CDKN2A
121.Syndrome	0.0011	0.0025	TAT | FGFR3 | SIX1 | COL11A1 | EDN3 | SPINK5
122.Robin sequence	0.0011	0.0025	COL11A1 | KCNJ16
123.Adhesion	0.0011	0.0025	COL17A1 | OLFM4 | CCL19 | S100A9 | ACTA1
124.Retinal Dysplasia	0.0013	0.0029	COL11A1 | SPINK5
125.Dermatitis	0.0013	0.0029	S100A7 | SPINK5 | CCL18
126.Carcinoma, Renal Cell	0.0014	0.0031	FABP7 | FN1 | CLDN8
127.Lymphoid leukemia NOS	0.0015	0.0033	PRAME | CCL18 | CD19 | CDKN2A
128.Asthma	0.0015	0.0033	MMP12 | CHIT1 | SPINK5 | CCL18
129.Leukemia, Lymphoid	0.0015	0.0033	PRAME | CCL18 | CD19 | CDKN2A
130.Carcinoma in Situ	0.0016	0.0034	S100A7 | SIX1 | CDKN2A
131.Glioblastoma	0.0017	0.0036	FABP7 | TFPI2 | EGF
132. Multiple synostosis syndrome	0.0017	0.0036	GREM1 | FGFR3
133.Tumor Virus Infections	0.0017	0.0036	CCL19 | CD19 | CDKN2A
134.Virus Diseases	0.0018	0.0038	CCL19 | CD19 | CDKN2A | ACTA1 | TUBB3
APPENDIX A. SUPPLEMENTARY TABLE IV			
DISEASE	**P**	**adjP**	**Genes involved in Disease**
135.Bone Neoplasms	0.0019	0.0039	GREM1 | WIF1 | CDKN2A
136.Rheumatic Diseases	0.0020	0.0041	CILP | ASPN | S100A9 | COMP
137.Precursor Cell Lymphoblastic Leukemia-Lymphoma	0.0020	0.0041	PRAME | CCL18 | CDKN2A
138.Urinary Bladder Neoplasms	0.0020	0.0041	FGFR3 | WIF1 | CDKN2A
139.Cholangiocarcinoma	0.0021	0.0042	TFF1 | TMEM100 | S100P
140.Hearing Loss, Sensorineural	0.0021	0.0042	FGFR3 | COL11A1 | ATP6V0A4
141.Lung Neoplasms	0.0022	0.0044	MMP12 | WIF1 | CDKN2A | TUBB3
142.Chronic Disease	0.0023	0.0045	MMP12 | PRAME | CHIT1 | S100A9
143.Spinal Diseases	0.0023	0.0045	CILP | ASPN | COL11A1
144.Rupture	0.0024	0.0047	MMP12 | FN1 | LYZ
145.Bile Duct Neoplasms	0.0025	0.0048	TFF1 | TMEM100 | S100P
146.Pleural Diseases	0.0025	0.0048	WIF1 | CDKN2A
147.Barrett Esophagus	0.0026	0.0049	TFF1 | CDKN2A
148.Cervical Intraepithelial Neoplasia	0.0026	0.0049	SIX1 | CDKN2A
149.Arthrogryposis	0.0026	0.0049	TNNI2 | ACTA1
150.Anxiety Disorders	0.0027	0.0050	PENK | NTRK3 | SLC1A1

151.Cystadenocarcinoma, Serous	0.0030	0.0055	PRAME | CDKN2A
152.Preterm rupture of membranes	0.0030	0.0055	MMP12 | FN1 | LYZ
153.Arthritis, Reactive	0.0031	0.0056	S100A9 | COMP
154.Calculi	0.0031	0.0056	TFF1 | SLC34A2
155.Cystadenocarcinoma	0.0031	0.0056	PRAME | CDKN2A
156.Warts	0.0032	0.0057	KRT4 | CDKN2A
157.Skin Abnormalities	0.0033	0.0058	COL17A1 | SPINK5 | KRT16
158.Lymphoma	0.0033	0.0058	PRAME | CD19 | SOX11 | CDKN2A
159.Abnormalities, Multiple	0.0033	0.0058	FGFR3 | COL11A1 | SPINK5 | COMP
160.Chorioamnionitis	0.0034	0.0059	MMP12 | FN1 | LYZ
161.Leukemia, B-Cell	0.0034	0.0059	PRAME | CD19 | SOX11
162.Bipolar Disorder	0.0034	0.0059	FABP7 | NRG1 | NTRK3 | LAMP3
163.Metabolism, Inborn Errors	0.0035	0.0060	TAT | SLC26A3 | CHIT1 | ATP6V0A4
164.Osteoarthritis, Hip	0.0035	0.0060	ASPN | COMP
165.Mesothelioma	0.0036	0.0061	WIF1 | CDKN2A
166.Alopecia	0.0041	0.0068	COL17A1 | KRT16
167.Neoplasm, Residual	0.0041	0.0068	PRAME | CD19
168.Rhinitis, Allergic, Seasonal	0.0041	0.0068	S100A7 | FOXJ1
APPENDIX A. SUPPLEMENTARY TABLE IV			
DISEASE	**P**	**adjP**	**Genes involved in Disease**
169.Hearing Loss, Bilateral	0.0046	0.0076	SIX1 | ATP6V0A4
170.Infertility, Male	0.0047	0.0077	SLC26A3 | CLGN | SOX11
171.Pheochromocytoma	0.0047	0.0077	PENK | DLK1
172.Metabolic Diseases	0.0048	0.0078	TAT | CHIT1 | PPARGC1A | LYZ | ATP6V0A4
173.Adenocarcinoma, Papillary	0.0050	0.0081	TFPI2 | S100P
174.Precursor T-Cell Lymphoblastic Leukemia-Lymphoma	0.0054	0.0086	PRAME | CDKN2A
175.Melanoma, Experimental	0.0054	0.0086	NRG1 | CDKN2A
176.Ichthyosis	0.0054	0.0086	SPINK5 | CSTA
177.Gastroenteritis	0.0056	0.0089	TFF1 | SLC26A3 | S100A9
178.Premature Birth	0.0057	0.0089	EDN2 | FN1 | CCL18
179.Immunologic Deficiency Syndromes	0.0057	0.0089	CCL19 | CD19 | ACTA1 | TUBB3
180.Optic Nerve Diseases	0.0059	0.0092	SIX1 | TUBB3
181.Arthritis, Juvenile Rheumatoid	0.0061	0.0094	S100A9 | COMP
182.Carcinoma, Hepatocellular	0.0062	0.0095	WIF1 | CDKN2A | DLK1
183.Infertility	0.0063	0.0096	SLC26A3 | CLGN | HSD17B2
184.Obstetric Labor Complications	0.0066	0.0100	EDN2 | FN1 | LYZ
185.Prostatic Hyperplasia	0.0066	0.0100	MSMB | KLK11
186.Ear Diseases	0.0070	0.0105	SIX1 | COL11A1 | ATP6V0A4
187.Hernia	0.0070	0.0105	CILP | COL11A1
188.Arterial Occlusive Diseases	0.0072	0.0107	MMP12 | CHIT1 | NR4A1
189.Endocrine System Diseases	0.0074	0.0108	PPARGC1A | CDKN2A | SLC34A2 | S100P
190.Endocrine disturbance NOS	0.0074	0.0108	PPARGC1A | CDKN2A | SLC34A2 | S100P
191.Endocrine disorder NOS	0.0074	0.0108	PPARGC1A | CDKN2A | SLC34A2 | S100P
192.Choriocarcinoma	0.0078	0.0114	TFPI2 | ASCL2
193.Hypertension	0.0079	0.0115	EDN2 | EDN3 | PPARGC1A
194.Sensation Disorders	0.0083	0.0120	SIX1 | COL11A1 | ATP6V0A4
195.Dwarfism	0.0084	0.0121	FGFR3 | COMP
196.Bronchitis	0.0089	0.0127	MMP12 | CCL19 | CCL18
197.Endometriosis	0.0095	0.0135	EGF | HSD17B2
198.Respiratory Tract Infections	0.0096	0.0136	CCL19 | S100A9 | CCL18
199.Hearing Loss, Conductive	0.0098	0.0137	SIX1 | COL11A1
200.Common Cold	0.0098	0.0137	CCL19 | S100A9 | CCL18
201.Mood Disorders	0.0112	0.0156	NRG1 | NTRK3 | LAMP3
202.Burns	0.0114	0.0156	FGFR3 | COMP
APPENDIX A. SUPPLEMENTARY TABLE IV			
DISEASE	**P**	**adjP**	**Genes involved in Disease**
203.Skin Manifestations	0.0114	0.0156	COL17A1 | SPINK5
204.Renal Disease, Pediatric	0.0114	0.0156	CITED1 | SIX1
205.Hyperplasia	0.0126	0.0172	TFF1 | KLK11
206.Kidney Diseases	0.0131	0.0178	GREM1 | OLFM4 | ATP6V0A4

207.Dermatitis, Atopic	0.0134	0.0181	SPINK5 | CCL18
208.Carcinoma, Small Cell	0.0136	0.0183	WIF1 | CDKN2A | TUBB3
209.Infection	0.0138	0.0185	CHIT1 | CCL19 | S100A9 | CDKN2A
210.Vasculitis	0.0139	0.0185	MMP12 | S100A9
211.Craniofacial Abnormalities	0.0147	0.0194	FGFR3 | COL11A1 | FGF18
212.Urologic Diseases	0.0146	0.0194	GREM1 | FGFR3 | ATP6V0A4
213.Asperger's disorder	0.0150	0.0197	NTRK3 | HSD17B2
214.Pharyngeal Neoplasms	0.0155	0.0203	WIF1 | CDKN2A
215.Thyroid Neoplasms	0.0158	0.0206	FN1 | NR4A1
216.Oral Manifestations	0.0167	0.0215	COL17A1 | KRT16
217.Carcinoma, Large Cell	0.0167	0.0215	CDKN2A | TUBB3
218.Mental Disorders	0.0185	0.0238	NRG1 | PENK | NTRK3 | SLC1A1
219.Uterine Cervical Neoplasms	0.0199	0.0254	SIX1 | CDKN2A
220.Brain Death	0.0202	0.0257	NRG1 | NTRK3
221.Glomerular disease	0.0218	0.0276	GREM1 | FN1
222.Diabetes Mellitus	0.0224	0.0283	GREM1 | PPARGC1A | CDKN2A
223.Cysts	0.0228	0.0286	KRT16 | HSD17B2
224.Multiple Myeloma	0.0234	0.0292	FGFR3 | PRAME
225.Disorder of uterus NOS	0.0244	0.0304	CDKN2A | EGF
226.Drug interaction with drug	0.0248	0.0306	FN1 | ACTA1 | EGF
227.Uterine Neoplasms	0.0248	0.0306	CDKN2A | EGF
228.Menopause, Premature	0.0254	0.0312	PENK | HSD17B2
229.Fatty Liver	0.0282	0.0345	FABP7 | PPARGC1A
230.Chromosome Aberrations	0.0289	0.0352	FGFR3 | CDKN2A | DLK1
231.Lentivirus Infections	0.0297	0.0360	CCL19 | ACTA1 | TUBB3
232.Sexually Transmitted Diseases	0.0301	0.0363	CCL19 | ACTA1 | TUBB3
233.Colitis	0.0308	0.0370	SLC26A3 | S100A9
234.Retroviridae Infections	0.0312	0.0372	CCL19 | ACTA1 | TUBB3
235.HIV Infections	0.0312	0.0372	CCL19 | ACTA1 | TUBB3
236.Cystic Fibrosis	0.0315	0.0374	SLC26A3 | S100A9
237.Peripheral neuroepithelioma	0.0349	0.0412	NTRK3 | DLK1
APPENDIX A. SUPPLEMENTARY TABLE IV			
DISEASE	**P**	**adjP**	**Genes involved in Disease**
238.Leukemia, Lymphocytic, Chronic, B-Cell	0.0357	0.0420	PRAME | SOX11
239.Eye Abnormalities	0.0365	0.0428	COL11A1 | OGN
240.Li-Fraumeni syndrome	0.0377	0.0438	FGFR3 | CDKN2A
241.Colonic Neoplasms	0.0377	0.0438	SLC26A3 | OLFM4
242.Muscle Weakness	0.0389	0.0450	TNNI2 | ACTA1
243.Bacterial Infections	0.0393	0.0453	S100A7 | CHIT1
244.Carcinoma, Non-Small-Cell Lung	0.0401	0.0460	CDKN2A | TUBB3
245.Oat cell carcinoma of lung	0.0417	0.0477	CDKN2A | TUBB3
246.Cleft Lip	0.0422	0.0478	COL11A1 | FGF18
247.Translocation, Genetic	0.0421	0.0478	SLC26A3 | FGFR3 | ATP6V0A4
248.Depression	0.0430	0.0485	NRG1 | NTRK3

19 Gene Signature (PLS-VIP-3NN)

APPENDIX B. SUPPLEMENTARY TABLE I. GENE LIST - DESCRIPTION		
gene id	**gene symbol**	**description**
1300	COL10A1	collagen, type X, alpha 1
6363	**CCL19**	chemokine (C-C motif) ligand 19
4320	MMP11	matrix metallopeptidase 11 (stromelysin 3)
6286	S100P	S100 calcium binding protein P
1311	COMP	cartilage oligomeric matrix protein
8483	CILP	cartilage intermediate layer protein, nucleotide pyrophosphohydrolase
2001	ELF5	E74-like factor 5 (ets domain transcription factor)
4680	CEACAM6	carcinoembryonic antigen-related cell adhesion molecule 6 (non-specific cross reacting antigen)
5284	PIGR	polymeric immunoglobulin receptor
9	**NAT1**	N-acetyltransferase 1 (arylamine N-acetyltransferase)
5021	OXTR	oxytocin receptor
3866	KRT15	keratin 15
2335	**FN1**	fibronectin 1
4312	**MMP1**	matrix metallopeptidase 1 (interstitial collagenase)
362	AQP5	aquaporin 5
1301	COL11A1	collagen, type XI, alpha 1
4283	**CXCL9**	chemokine (C-X-C motif) ligand 9
2191	**FAP**	fibroblast activation protein, alpha
11202	**KLK8**	kallikrein-related peptidase 8

GENENAME: associated with breast cancer

GENENAME: not associated yet with breast cancer

KEGG PATHWAY ANALYSIS
19 PLS-VIP-3NN Gene Signature (p≤0.05)

APPENDIX B. SUPPLEMENTARY TABLE II			
KEGG Pathway	**P**	**adjP**	**Genes participated in KEGG Pathway**
1.ECM-receptor interaction	7.00e-06	4.20e-05	COL11A1; FN1; COMP
2.Focal adhesion	9.01e-05	0.0003	COL11A1; FN1; COMP
3.Amoebiasis	0.0010	0.0020	COL11A1; FN1
4.Chemokine signaling pathway	0.0031	0.0046	CXCL9; CCL19
5.Cytokine-cytokine receptor interaction	0.0060	0.0072	CXCL9; CCL19
6.Pathways in cancer	0.0089	0.0089	MMP1; FN1

GENE ONTOLOGY ENRICHMENT ANALYSIS in terms of BIOLOGICAL PROCESS

19 PLS-VIP-3NN Gene Signature (p≤0.05)

APPENDIX B. SUPPLEMENTARY TABLE III			
Biological Process	**P**	**adjP**	**Genes participated in Biological Process**
1.extracellular matrix organization	1.35e-07	1.57e-05	MMP1; FAP; COL11A1; MMP11; FN1; COL10A1
2.extracellular structure organization	1.38e-07	1.57e-05	MMP1; FAP; COL11A1; MMP11; FN1; COL10A1
3.extracellular matrix disassembly	1.18e-05	0.0009	MMP1; FAP; MMP11
4.collagen catabolic process	0.0003	0.0171	MMP1; MMP11
5.multicellular organismal catabolic process	0.0006	0.0228	MMP1; MMP11
6.regulation of body fluid levels	0.0006	0.0228	MMP1; FAP; AQP5; OXTR; FN1
7.multicellular organismal metabolic process	0.0038	0.0365	MMP1; MMP11
8.digestive system process	0.0020	0.0365	AQP5; OXTR
9.body fluid secretion	0.0036	0.0365	AQP5; OXTR
10.regulation of synapse organization	0.0014	0.0365	OXTR; KLK8
11.multicellular organismal development	0.0032	0.0365	KRT15; MMP11; KLK8; CCL19; FN1; ELF5; AQP5; COL11A1; OXTR; COMP; COL10A1
12.system development	0.0040	0.0365	KRT15; KLK8; CCL19; FN1; ELF5; AQP5; COL11A1; OXTR; COMP; COL10A1
13.response to wounding	0.0016	0.0365	CXCL9; MMP1; FAP; KLK8; CCL19; FN1
14.cell migration	0.0028	0.0365	MMP1; FAP; CCL19; FN1; S100P
15.single-organism process	0.0013	0.0365	CXCL9; KRT15; FAP; CILP; CEACAM6; MMP11; KLK8; CCL19; FN1; ELF5; MMP1; AQP5; OXTR; COL11A1; COMP; COL10A1
16.leukocyte migration	0.0033	0.0365	MMP1; CCL19; FN1
17.cell motility	0.0040	0.0365	MMP1; FAP; CCL19; FN1; S100P
18.locomotion	0.0028	0.0365	CXCL9; MMP1; FAP; CCL19; FN1; S100P
19.memory	0.0034	0.0365	OXTR; KLK8
20.single-multicellular organism process	0.0037	0.0365	**KRT15; FAP; MMP11; KLK8; CCL19; FN1; ELF5; MMP1; AQP5; OXTR; COL11A1; COMP; COL10A1**
21.multicellular organismal process	0.0040	0.0365	**KRT15; FAP; MMP11; KLK8; CCL19; FN1; ELF5; MMP1; AQP5; OXTR; COL11A1; COMP; COL10A1**
22.multicellular organismal macromolecule metabolic process	0.0028	0.0365	MMP1; MMP11
23.localization of cell	0.0040	0.0365	MMP1; FAP; CCL19; FN1; S100P
24.collagen metabolic process	0.0023	0.0365	MMP1; MMP11
25.regulation of synapse structure and activity	0.0018	0.0365	OXTR; KLK8
26.cellular component disassembly at cellular level	0.0050	0.0438	MMP1; FAP; MMP11
27.cellular component disassembly	0.0052	0.0439	MMP1; FAP; MMP11

APPENDIX B. SUPPLEMENTARY TABLE IV			
DISEASE	**P**	**adjP**	**Genes involved in Disease**
1.Cartilage Diseases	5.03e-11	5.18e-09	MMP1 | CILP | COL11A1 | COMP | COL10A1
2.Fibrosis	1.39e-10	7.16e-09	CXCL9 | MMP1 | FAP | AQP5 | FN1 | COMP
3.Collagen Diseases	5.58e-10	1.92e-08	MMP1 | COL11A1 | FN1 | COMP | COL10A1
4.Osteoarthritis	4.64e-09	1.19e-07	MMP1 | CILP | COL11A1 | COMP | COL10A1
5.Neoplastic Processes	1.76e-08	3.63e-07	MMP1 | FAP | CEACAM6 | MMP11 | FN1 | S100P
6.Osteoarthritis, Knee	2.42e-08	4.15e-07	MMP1 | CILP | COL11A1 | COMP
7.Obstetric Labor Complications	3.01e-08	4.43e-07	MMP1 | OXTR | MMP11 | FN1 | NAT1
8.Carcinoma	7.25e-08	9.33e-07	MMP1 | FAP | CEACAM6 | MMP11 | KLK8 | S100P
9.Degeneration of lumbar intervertebral disc	2.36e-07	2.70e-06	CILP | COL11A1 | COL10A1
10.Lumbar Disc Herniation	3.23e-07	2.77e-06	CILP | COL11A1 | COL10A1
11.Dermatitis, Allergic Contact	3.23e-07	2.77e-06	CXCL9 | CCL19 | NAT1
12.Dermatitis, Contact	2.92e-07	2.77e-06	CXCL9 | CCL19 | NAT1
13.Adenocarcinoma	4.28e-07	3.39e-06	AQP5 | CEACAM6 | MMP11 | NAT1 | S100P
14.Preterm rupture of membranes	6.77e-07	4.98e-06	MMP1 | OXTR | MMP11 | FN1
15.Neoplasm of unspecified nature of digestive system	1.18e-06	8.10e-06	FAP | CEACAM6 | MMP11 | NAT1 | S100P
16.Neoplasms	1.29e-06	8.30e-06	MMP1 | FAP | CEACAM6 | MMP11 | NAT1 | S100P
17.Musculoskeletal Diseases	1.42e-06	8.60e-06	MMP1 | CILP | COL11A1 | COMP | COL10A1
18.Premature Birth	1.68e-06	9.61e-06	MMP1 | OXTR | FN1 | NAT1
19.Cancer or Viral infections	2.40e-06	1.30e-05	MMP1 | FAP | CEACAM6 | MMP11 | NAT1 | S100P
20.Skeletal Dysplasia	2.97e-06	1.53e-05	COL11A1 | COMP | COL10A1
21.Bone Diseases	3.68e-06	1.80e-05	CILP | COL11A1 | COMP | COL10A1
22.Colorectal Neoplasms	4.66e-06	2.18e-05	MMP1 | CEACAM6 | MMP11 | NAT1
23.Neoplasm Invasiveness	7.98e-06	3.57e-05	MMP1 | FAP | MMP11 | S100P
24.Arthritis	8.41e-06	3.61e-05	MMP1 | CILP | COL11A1 | COMP
25.Epithelial cancers	1.00e-05	3.96e-05	FAP | CEACAM6 | MMP11 | S100P
26.**Neoplasm Metastasis**	9.93e-06	3.96e-05	MMP1 | FAP | MMP11 | S100P
27.Osteochondrodysplasias	1.18e-05	4.50e-05	COL11A1 | COMP | COL10A1
28.Gastrointestinal Neoplasms	1.57e-05	5.78e-05	MMP1 | CEACAM6 | MMP11 | NAT1
29.Oral submucosal fibrosis	1.67e-05	5.93e-05	MMP1 | COMP
APPENDIX B. SUPPLEMENTARY TABLE IV			
DISEASE	**P**	**adjP**	**Genes involved in Disease**
30.Gastrointestinal Diseases	2.87e-05	9.54e-05	MMP1 | CEACAM6 | MMP11 | NAT1
31.Sweat gland disorder NOS	2.80e-05	9.54e-05	KRT15 | AQP5
32.Spinal Diseases	3.47e-05	0.0001	CILP | COL11A1 | COL10A1
33.Rupture	3.69e-05	0.0001	MMP1 | MMP11 | FN1
34.Chorioamnionitis	5.28e-05	0.0002	OXTR | MMP11 | FN1
35.Bronchitis	0.0001	0.0003	CXCL9 | MMP1 | CCL19
36.Carcinoma, Acinar Cell	0.0001	0.0003	KLK8 | FN1
37.Intestinal Neoplasms	0.0002	0.0005	CEACAM6 | MMP11 | NAT1
38.Colonic Diseases	0.0002	0.0005	MMP1 | CEACAM6 | NAT1
39.Polymyositis	0.0002	0.0005	CXCL9 | CCL19
40.Adenocarcinoma, Mucinous	0.0002	0.0005	CEACAM6 | S100P
41.Myositis	0.0002	0.0005	CXCL9 | CCL19
42.Osteoarthritis, Hip	0.0002	0.0005	MMP1 | COMP
43.Joint Diseases	0.0002	0.0005	MMP1 | CILP | COMP
44.Rheumatic Diseases	0.0003	0.0007	MMP1 | CILP | COMP
45.Connective Tissue Diseases	0.0003	0.0007	MMP1 | CILP | COMP
46.Hernia	0.0004	0.0008	CILP | COL11A1
47.Polyps	0.0004	0.0008	AQP5 | CEACAM6
48.Fibrosarcoma	0.0004	0.0008	MMP1 | FN1
49.Chondrosarcoma	0.0004	0.0008	MMP1 | COMP

50.Colonic Diseases, Functional	0.0004	0.0008	MMP1 | CEACAM6
51.Intestinal Diseases	0.0004	0.0008	MMP1 | CEACAM6 | NAT1
52.Musculoskeletal Abnormalities	0.0004	0.0008	COL11A1 | COMP | NAT1
53.Lung Diseases	0.0005	0.0009	MMP1 | AQP5 | PIGR
54.Breast Diseases	0.0005	0.0009	MMP1 | MMP11 | NAT1
55.Scleroderma, Systemic	0.0005	0.0009	MMP1 | COMP
56.Endometriosis	0.0006	0.0011	MMP1 | OXTR
57.Nasopharyngeal Neoplasms	0.0006	0.0011	MMP1 | PIGR
58.Refractive Errors	0.0008	0.0014	MMP1 | COL11A1
59.Skin Diseases	0.0008	0.0014	KRT15 | MMP1 | MMP11
60.Carcinoma, Papillary	0.0009	0.0015	FN1 | S100P
APPENDIX B. SUPPLEMENTARY TABLE IV			
Genes involved in Disease			
61.Thyroid Neoplasms	0.0010	0.0017	MMP11 | FN1
62.Pharyngeal Neoplasms	0.0010	0.0017	MMP1 | PIGR
63.Skin and Connective Tissue Diseases	0.0012	0.0020	KRT15 | MMP1 | MMP11
64.Thyroid Diseases	0.0014	0.0022	CXCL9 | FN1
65.Mouth Neoplasms	0.0014	0.0022	MMP1 | KLK8
66.Lung Diseases, Obstructive	0.0016	0.0025	MMP1 | AQP5
67.Cholangiocarcinoma	0.0017	0.0026	NAT1 | S100P
68.Pulmonary Disease, Chronic Obstructive	0.0018	0.0027	MMP1 | AQP5
69.Pathologic Processes	0.0018	0.0027	MMP1 | MMP11 | NAT1
70.Pancreatic Diseases	0.0019	0.0028	CEACAM6 | S100P
71.Pain	0.0019	0.0028	OXTR | COMP
72.Recurrence	0.0024	0.0034	MMP1 | MMP11
73.Bacterial Infections	0.0027	0.0037	CXCL9 | PIGR
74.Adhesion	0.0027	0.0037	CEACAM6 | CCL19 | FN1
75.Colonic Neoplasms	0.0026	0.0037	CEACAM6 | PIGR
76.Growth Disorders	0.0029	0.0039	CILP | COMP
77.Cleft Lip	0.0029	0.0039	COL11A1 | NAT1
78.Cleft Palate	0.0035	0.0045	COL11A1 | NAT1
79.Gastroenteritis	0.0035	0.0045	CXCL9 | CEACAM6
80.Pancreatic Neoplasms	0.0034	0.0045	CEACAM6 | S100P
81.Transplantation	0.0036	0.0046	CXCL9 | FAP
82.Bronchiolitis	0.0041	0.0051	CXCL9 | CCL19
83.Lymphoma, Low-Grade	0.0041	0.0051	CCL19 | NAT1
84.Disease Progression	0.0042	0.0052	MMP1 | MMP11
85.Genetic Predisposition to Disease	0.0051	0.0061	MMP1 | OXTR | NAT1
86.Respiratory Tract Infections	0.0051	0.0061	CXCL9 | CCL19
87.Common Cold	0.0052	0.0062	CXCL9 | CCL19
88.Disease Susceptibility	0.0054	0.0063	MMP1 | OXTR | NAT1
89.Head and Neck Neoplasms	0.0059	0.0068	MMP1 | FN1
APPENDIX B. SUPPLEMENTARY TABLE IV			
Genes involved in Disease			
90.Arthritis, Rheumatoid	0.0060	0.0069	MMP1 | COMP
91.Craniofacial Abnormalities	0.0070	0.0079	COL11A1 | NAT1
92.Chronic Disease	0.0080	0.0090	CXCL9 | MMP1
93.Skin Diseases, Genetic	0.0081	0.0090	MMP1 | NAT1
94.Abnormalities, Multiple	0.0097	0.0106	COL11A1 | COMP
95.Lymphatic Diseases	0.0110	0.0119	CCL19 | NAT1
96.Lymphoproliferative Disorders	0.0116	0.0123	CCL19 | NAT1
97.Respiratory Tract Diseases	0.0116	0.0123	MMP1 | AQP5
98.Breast Neoplasms	0.0118	0.0124	MMP11 | NAT1

99.Urogenital Neoplasms	0.0153	0.0159	KLK8	NAT1
100.Inflammation	0.0155	0.0160	CXCL9	MMP1
101.Virus Diseases	0.0192	0.0196	CXCL9	CCL19
102.Infection	0.0214	0.0216	CXCL9	CCL19
103.Immune System Diseases	0.0356	0.0356	CXCL9	CCL19

APPENDIX C. SUPPLEMENTARY TABLE I. GENE LIST - DESCRIPTION		
gene id	**gene symbol**	**description**
10381	TUBB3	tubulin, beta 3
1301	COL11A1	collagen, type XI, alpha 1
1311	COMP	cartilage oligomeric matrix protein
1907	EDN2	endothelin 2
2335	FN1	fibronectin 1
23532	PRAME	preferentially expressed antigen in melanoma
2938	GSTA1	glutathione S-transferase alpha 1
3500	IGHG1	immunoglobulin heavy constant gamma 1 (G1m marker)
4321	MMP12	matrix metallopeptidase 12 (macrophage elastase)
6278	S100A7	S100 calcium binding protein A7
6286	S100P	S100 calcium binding protein P
6363	CCL19	chemokine (C-C motif) ligand 19
7031	TFF1	trefoil factor 1
7980	TFPI2	tissue factor pathway inhibitor 2
8483	CILP	cartilage intermediate layer protein, nucleotide pyrophosphohydrolase
8788	DLK1	delta-like 1 homolog (Drosophila)

GENENAME: associated with breast cancer

GENENAME: not associated yet with breast cancer

KEGG PATHWAYS
'16 Common-Gene Signature' (p≤0.05)

APPENDIX C. SUPPLEMENTARY TABLE II			
KEGG Pathways	**P**	**adjP**	**Genes participated in KEGG Pathways**
1.ECM-receptor interaction	4.06e-06	1.62e-05	COL11A1 | FN1 | COMP
2.Focal adhesion	5.26e-05	0.0001	COL11A1 | FN1 | COMP
3.Amoebiasis	0.0007	0.0009	COL11A1 | FN1
4.Phagosome	0.0015	0.0015	COMP | TUBB3

GENE ONTOLOGY ENRICHMENT ANALYSIS in terms of BIOLOGICAL PROCESS

'16 Common-Gene Signature' (p≤0.05)

APPENDIX C. SUPPLEMENTARY TABLE III

Biological Process	P	adjP	Genes participated in Biological Process
1.cellular component movement	0.0001	0.0027	MMP12 | EDN2 | S100A7 | FN1 | CCL19 | TUBB3 | S100P
2.leukocyte migration	0.0001	0.0027	EDN2 | S100A7 | FN1 | CCL19
3.positive regulation of leukocyte chemotaxis	2.04e-05	0.0027	EDN2 | S100A7 | CCL19
4.positive regulation of behavior	0.0001	0.0027	EDN2 | S100A7 | CCL19
5.response to stimulus	6.46e-05	0.0027	TFF1 | MMP12 | EDN2 | PRAME | TFPI2 | CILP | S100A7 | FN1 | CCL19 | S100P | COL11A1 | IGHG1 | DLK1 | TUBB3 | GSTA1
6.regulation of leukocyte migration	9.35e-05	0.0027	EDN2 | S100A7 | CCL19
7.positive regulation of leukocyte migration	3.90e-05	0.0027	EDN2 | S100A7 | CCL19
8.positive regulation of chemotaxis	6.37e-05	0.0027	EDN2 | S100A7 | CCL19
9.regulation of leukocyte chemotaxis	3.19e-05	0.0027	EDN2 | S100A7 | CCL19
10.locomotion	0.0002	0.0041	MMP12 | EDN2 | S100A7 | FN1 | CCL19 | TUBB3 | S100P
11.regulation of chemotaxis	0.0002	0.0041	EDN2 | S100A7 | CCL19
12.cell migration	0.0002	0.0041	MMP12 | EDN2 | S100A7 | FN1 | CCL19 | S100P
13.cell motility	0.0003	0.0049	MMP12 | EDN2 | S100A7 | FN1 | CCL19 | S100P
14.localization of cell	0.0003	0.0049	MMP12 | EDN2 | S100A7 | FN1 | CCL19 | S100P
15.leukocyte chemotaxis	0.0003	0.0049	EDN2 | S100A7 | CCL19
16.regulation of behavior	0.0004	0.0061	EDN2 | S100A7 | CCL19
17.cell chemotaxis	0.0005	0.0068	EDN2 | S100A7 | CCL19
18.positive regulation of response to external stimulus	0.0005	0.0068	EDN2 | S100A7 | CCL19
19.neutrophil chemotaxis	0.0013	0.0167	EDN2 | CCL19
20.positive regulation of cell motility	0.0018	0.0209	EDN2 | S100A7 | CCL19
21.positive regulation of cell migration	0.0018	0.0209	EDN2 | S100A7 | CCL19
22.positive regulation of cellular component movement	0.0020	0.0212	EDN2 | S100A7 | CCL19
23.positive regulation of locomotion	0.0020	0.0212	EDN2 | S100A7 | CCL19
24.positive regulation of immune system process	0.0025	0.0254	EDN2 | S100A7 | CCL19 | IGHG1
25.sequestering of metal ion	0.0027	0.0264	S100A7 | CCL19
26.taxis	0.0033	0.0298	EDN2 | S100A7 | CCL19 | TUBB3
27.chemotaxis	0.0033	0.0298	EDN2 | S100A7 | CCL19 | TUBB3
28.positive regulation of ERK1 and ERK2 cascade	0.0040	0.0349	S100A7 | CCL19
29.positive regulation of cell proliferation	0.0047	0.0395	MMP12 | EDN2 | PRAME | CCL19

DISEASE ASSOCIATION ANALYSIS
'16 Common-Gene Signature' (p≤0.05)

APPENDIX C. SUPPLEMENTARY TABLE IV			
DISEASE	**P**	**adjP**	**Genes involved in Disease**
1.Neoplasms	1.14e-08	9.35e-07	TFF1 | PRAME | TFPI2 | S100A7 | TUBB3 | GSTA1 | S100P
2.Disease Progression	1.14e-06	2.33e-05	TFF1 | MMP12 | PRAME | S100A7
3.Carcinoma	9.97e-07	2.33e-05	TFF1 | TFPI2 | S100A7 | TUBB3 | S100P
4.cancer or viral infections	7.50e-07	2.33e-05	TFF1 | PRAME | TFPI2 | S100A7 | TUBB3 | S100P
5.Cartilage Diseases	1.42e-06	2.33e-05	CILP | COL11A1 | COMP
6.Osteoarthritis, Knee	2.26e-06	3.09e-05	CILP | COL11A1 | COMP
7.Collagen Diseases	5.86e-06	6.86e-05	COL11A1 | FN1 | COMP
8.Neoplastic Processes	1.35e-05	0.0001	PRAME | TFPI2 | FN1 | S100P
9.Neoplasm of unspecified nature of digestive system	1.84e-05	0.0002	TFF1 | TFPI2 | GSTA1 | S100P
10.Osteoarthritis	2.06e-05	0.0002	CILP | COL11A1 | COMP
11.Pancreatic Diseases	2.10e-05	0.0002	TFF1 | TFPI2 | S100P
12.Pancreatic Neoplasms	5.03e-05	0.0003	TFF1 | TFPI2 | S100P
13.Lumbar Disc Herniation	5.96e-05	0.0003	CILP | COL11A1
14.Degeneration of lumbar intervertebral disc	4.85e-05	0.0003	CILP | COL11A1
15.Bone Diseases	9.60e-05	0.0005	CILP | COL11A1 | COMP
16.Adenocarcinoma, Papillary	0.0002	0.0008	TFPI2 | S100P
17.Adenocarcinoma, Mucinous	0.0002	0.0008	TFF1 | S100P
18.Arthritis	0.0002	0.0008	CILP | COL11A1 | COMP
19.Epithelial cancers	0.0002	0.0008	TFF1 | S100A7 | S100P
20.Neuroectodermal Tumors	0.0002	0.0008	PRAME | TFPI2 | DLK1
21.Neoplasm Invasiveness	0.0002	0.0008	MMP12 | TFPI2 | S100P
22.Breast Neoplasms	0.0003	0.0009	TFF1 | S100A7 | GSTA1
23.Breast Diseases	0.0003	0.0009	TFF1 | S100A7 | GSTA1
24.Hernia	0.0003	0.0009	CILP | COL11A1
25.Gastrointestinal Neoplasms	0.0003	0.0009	TFF1 | TFPI2 | GSTA1
26.Skeletal Dysplasia	0.0003	0.0009	COL11A1 | COMP
27.Fibrosarcoma	0.0003	0.0009	TFPI2 | FN1
28.Adenocarcinoma	0.0003	0.0009	TFF1 | TUBB3 | S100P
29. Scoliosis	0.0004	0.0011	COMP | DLK1
30.Gastrointestinal Diseases	0.0004	0.0011	TFF1 | TFPI2 | GSTA1
31.Precancerous Conditions	0.0005	0.0013	TFF1 | S100A7
32.Inflammation	0.0005	0.0013	MMP12 | EDN2 | S100A7
33.Osteochondrodysplasias	0.0006	0.0014	COL11A1 | COMP
APPENDIX C. SUPPLEMENTARY TABLE IV			
DISEASE	**P**	**adjP**	**Genes involved in Disease**
34.Carcinoma, Papillary	0.0006	0.0014	FN1 | S100P
35.Musculoskeletal Diseases	0.0006	0.0014	CILP | COL11A1 | COMP
36.Kidney Neoplasms	0.0011	0.0025	EDN2 | DLK1
37.Cholangiocarcinoma	0.0012	0.0027	TFF1 | S100P
38.Spinal Diseases	0.0013	0.0028	CILP | COL11A1
39.Bile Duct Neoplasms	0.0014	0.0029	TFF1 | S100P
40.Rupture	0.0014	0.0029	MMP12 | FN1
41.Preterm rupture of membranes	0.0016	0.0032	MMP12 | FN1
42.Chorioamnionitis	0.0017	0.0033	MMP12 | FN1
43.Fibrosis	0.0021	0.0038	FN1 | COMP
44.Urologic Neoplasms	0.0021	0.0038	EDN2 | DLK1
45.Growth Disorders	0.0021	0.0038	CILP | COMP
46.Premature Birth	0.0025	0.0045	EDN2 | FN1
47.Neuroendocrine Tumors	0.0027	0.0047	PRAME | DLK1
48.Obstetric Labor Complications	0.0028	0.0048	EDN2 | FN1
49.Lymphoma, Low-Grade	0.0029	0.0049	PRAME | CCL19
50.Lymphoma, B-Cell	0.0033	0.0054	PRAME | CCL19
51.Bronchitis	0.0034	0.0055	MMP12 | CCL19
52.Pregnancy Complications	0.0035	0.0055	EDN2 | TFPI2
53.Colorectal Neoplasms	0.0041	0.0063	TFPI2 | GSTA1
54.Joint Diseases	0.0043	0.0064	CILP | COMP

55.Intestinal Neoplasms	0.0044	0.0064	TFF1 | GSTA1
56.Head and Neck Neoplasms	0.0042	0.0064	S100A7 | FN1
57.Connective Tissue Diseases	0.0051	0.0073	CILP | COMP
58.Rheumatic Diseases	0.0053	0.0075	CILP | COMP
59.Chronic Disease	0.0057	0.0077	MMP12 | PRAME
60.Lung Neoplasms	0.0056	0.0077	MMP12 | TUBB3
61.Bronchial Diseases	0.0057	0.0077	MMP12 | CCL19
62.Neoplasm Metastasis	0.0060	0.0079	TFPI2 | S100P
63.Musculoskeletal Abnormalities	0.0064	0.0083	COL11A1 | COMP
64.Intestinal Diseases	0.0066	0.0085	TFF1 | GSTA1
65.Abnormalities, Multiple	0.0069	0.0087	COL11A1 | COMP
66.Lung Diseases	0.0075	0.0093	MMP12 | TUBB3
67.Lymphatic Diseases	0.0078	0.0095	PRAME | CCL19
APPENDIX C. SUPPLEMENTARY TABLE IV			
DISEASE	**P**	**adjP**	**Genes involved in Disease**
68.Lentivirus Infections	0.0083	0.0097	CCL19 | TUBB3
69.Sexually Transmitted Diseases	0.0084	0.0097	CCL19 | TUBB3
70.Lymphoproliferative Disorders	0.0083	0.0097	PRAME | CCL19
71.Respiratory Tract Diseases	0.0083	0.0097	MMP12 | TUBB3
72.HIV Infections	0.0087	0.0098	CCL19 | TUBB3
73.Retroviridae Infections	0.0087	0.0098	CCL19 | TUBB3
74.Immunologic Deficiency Syndromes	0.0094	0.0104	CCL19 | TUBB3
75.Skin Diseases	0.0102	0.0112	TFF1 | S100A7
76.Skin and Connective Tissue Diseases	0.0134	0.0145	TFF1 | S100A7
77.Virus Diseases	0.0138	0.0147	CCL19 | TUBB3
78.Adhesion	0.0235	0.0247	FN1 | CCL19
79.Immune System Diseases	0.0257	0.0267	PRAME | CCL19
80.HIV	0.0312	0.0320	FN1 | TUBB3
81.Genetic Predisposition to Disease	0.0354	0.0358	MMP12 | GSTA1
82.**Disease Susceptibility**	0.0367	0.0367	MMP12 | GSTA1

APPENDIX D. SUPPLEMENTARY TABLE I. GENE LIST - DESCRIPTION		
gene id	**gene symbol**	**description**
1311	COMP	cartilage oligomeric matrix protein
2335	FN1	fibronectin 1
6286	S100P	S100 calcium binding protein P
6363	CCL19	chemokine (C-C motif) ligand 19
8483	CILP	cartilage intermediate layer protein, nucleotide pyrophosphohydrolase
GENENAME: associated with breast cancer *GENENAME*: not associated yet with breast cancer		

KEGG PATHWAYS
'5 Common-Gene Signature' (p≤0.05)

APPENDIX D. SUPPLEMENTARY TABLE II				
KEGG Pathways	**P**	**adjP**	**Genes participated in KEGG Pathways**	
1.ECM-receptor interaction	3.82e-05	7.64e-05	FN1	COMP
2.Focal adhesion	0.0002	0.0002	FN1	COMP

GENE ONTOLOGY ENRICHMENT ANALYSIS in terms of BIOLOGICAL PROCESS

'5 Common-Gene Signature' (p≤0.1)

APPENDIX D. SUPPLEMENTARY TABLE III					
Biological Process	**P**	**adjP**	**Genes participated in Biological Process**		
1.localization of cell	0.0022	0.0602	FN1	CCL19	S100P
2.cell migration	0.0017	0.0602	FN1	CCL19	S100P
3.leukocyte migration	0.0028	0.0602	FN1	CCL19	
4.cell motility	0.0022	0.0602	FN1	CCL19	S100P
5.locomotion	0.0054	0.0774	FN1	CCL19	S100P
6.cellular component movement	0.0047	0.0774	FN1	CCL19	S100P

DISEASE ASSOCIATION ANALYSIS
'5 Common-Gene Signature' (p≤0.05)

APPENDIX D. SUPPLEMENTARY TABLE IV			
DISEASE	**P**	**adjP**	**Genes involved in Disease**
1.Osteoarthritis, Knee	2.59e-05	0.0002	CILP | COMP
2.Collagen Diseases	4.88e-05	0.0002	FN1 | COMP
3.Carcinoma, Papillary	5.41e-05	0.0002	FN1 | S100P
4.Cartilage Diseases	1.90e-05	0.0002	CILP | COMP
5.Osteoarthritis	0.0001	0.0003	CILP | COMP
6.Fibrosis	0.0002	0.0004	FN1 | COMP
7.Growth Disorders	0.0002	0.0004	CILP | COMP
8.Rheumatic Diseases	0.0005	0.0006	CILP | COMP
9.Arthritis	0.0005	0.0006	CILP | COMP
10.Joint Diseases	0.0004	0.0006	CILP | COMP
11.Connective Tissue Diseases	0.0004	0.0006	CILP | COMP
12.Bone Diseases	0.0003	0.0006	CILP | COMP
13.Neoplastic Processes	0.0009	0.0010	FN1 | S100P
14.Musculoskeletal Diseases	0.0011	0.0012	CILP | COMP
15.Adhesion	0.0022	0.0022	FN1 | CCL19