TECHNICAL UNIVERSITY OF
CRETE
Department of Electronic &
Computer Engineering

PATTERN RECOGNITION APPROACHES IN DNA MICROARRAY ANALYSIS

MICHAIL E. BLAZADONAKIS

July 2008

TECHNICAL UNIVERSITY OF
CRETE
Department of Electronic &
Computer Engineering

PATTERN RECOGNITION APPROACHES IN DNA MICROARRAY ANALYSIS

by

MICHAIL E. BLAZADONAKIS

A thesis submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy, PhD

Chairperson of the Supervisor Committee: Professor Michalis Zervakis
Professor Stavros Christodoulakis
Professor George Stavrakakis
Professor Nikos Sidiropoulos
Associate Professor Evripidis Petrakis
Associate Professor Konstantinos Balas
Associate Professor Dimitris Fotiadis

CHANIA, GREECE, July 2008

Στην οικογένεια μου
Χαρούλα, Κάλλια, και Μαριώ

Contents

ΠΡΟΛΟΓΟΣ

«Χαράς τονε τον άνθρωπο πού `χει φτερά στον ώμο,
κι όμως γροικάται ταπεινά το ζάλο του στο δρόμο.»

Κρητική μαντινάδα

Το σύγγραμμα αυτό αποτελεί τη διδακτορική μου διατριβή η οποία εκπονήθηκε στο Πολυτεχνείο Κρήτης από τον Οκτώβριο του 2003 έως τον Ιούνιο του 2008, στο Τμήμα Ηλεκτρονικών Μηχανικών και Μηχανικών Ηλεκτρονικών Υπολογιστών. Κίνητρό μου για την προσπάθεια αυτή είναι το μεράκι για έρευνα και μάθηση, μαζί με μια συνεχή ανησυχία και αναζήτηση για κάτι πέρα από τα δεδομένα. Το αποτέλεσμα της δουλειάς αυτής θα το κρίνετε εσείς και οι μελλοντικοί ερευνητές, όμως καταθέτω ότι αποτελεί μια λεπτομερή και ειλικρινή καταγραφή αποτελεσμάτων, μεθόδων και παρατηρήσεων, μιας διαρκούς, αγωνιώδους, αλλά και πολύ ενδιαφέρουσας προσπάθειας.

Θέλω να ευχαριστήσω τον κ. Ζερβάκη Μιχάλη, Καθηγητή του Πολυτεχνείου Κρήτης για τη συνεχή και αδιάλειπτη παρακολούθηση, αλλά και το συντονισμό της εργασίας αυτής. Για τις πολύ εύστοχες παρατηρήσεις, πρωτοποριακές ιδέες και ερευνητικές κατευθύνσεις που μου έδειξε. Τέλος, τον ευχαριστώ σαν άνθρωπο, γιατί πραγματικά βίωσε την όλη προσπάθεια, στηρίζοντάς με τόσο επιστημονικά όσο και ηθικά, ενώ ο χαρακτήρας και η προσωπικότητά του, μου επέτρεψαν να εκφραστώ ελεύθερα και να είμαι ο εαυτός μου.

Ευχαριστώ όλους τους καθηγητές της συμβουλευτικής επιτροπής, για το χρόνο που αφιέρωσαν να μελετήσουν και να αξιολογήσουν την εργασία αυτή, για την υποστήριξη, την αμεσότητα και τη φιλικότητα τους, δημιουργώντας ένα οικείο περιβάλλον συνεργασίας.

Ευχαριστώ όλους τους φίλους-συναδέλφους και ερευνητές του Εργαστηρίου Ψηφιακής Επεξεργασίας Σημάτων και Εικόνας για την άψογη και απρόσκοπτη συνεργασία, ιδιαίτερα ευχαριστώ τον Μιχάλη Κουνελάκη και τον Γιώργο Μανίκη για το χρόνο που αφιέρωσαν, τις ατελείωτες αλλά πολύ χρήσιμες συζητήσεις μας όλα αυτά τα χρόνια, γύρω από ερευνητικά θέματα κοινού ενδιαφέροντος.

Ευχαριστώ τους συναδέλφους ερευνητές, του Ινστιτούτου Μοριακής Βιολογίας και Βιοτεχνολογίας του Ιδρύματος Τεχνολογίας και Έρευνας, για τη συνεργασία μας και για τα πολύ χρήσιμα σχόλια, παρατηρήσεις και κατευθύνσεις που μας πρότειναν.

Ευχαριστώ την Γαλάτεια Μαλανδράκη, Γραμματέα του Τμήματος Αρχιτεκτόνων Μηχανικών του Πολυτεχνείου Κρήτης, οικογενειακή μας φίλη, για την πολύτιμη και άμεση βοήθειά της στη διεκπεραίωση διαφόρων θεμάτων, ιδιαίτερα στα πρώτα στάδια της προσπάθειας αυτής.

Ευχαριστώ τους γονείς μου, που μου έδειξαν το μονοπάτι που βαδίζω σήμερα, που μου έμαθαν ότι ο δρόμος της γνώσης είναι πλούτος και που στηρίζουν με όλες τους τις δυνάμεις την οικογένειά μου. Τα αδέρφια μου Γιώργη και Στεφανή για το ενδιαφέρον, την υποστήριξη και την αμέριστη συμπαράστασή τους σε ότι χρειάστηκα όλα αυτά τα χρόνια.

Ευχαριστώ τα πεθερικά μου, που από την πρώτη στιγμή κατάλαβαν και συμπαραστάθηκαν στην προσπάθειά μου, ενώ η φροντίδα και η στήριξή τους σε ότι χρειαστήκαμε ήταν πάντα δεδομένη.

Αφήνω για το τέλος τις πιο θερμές μου ευχαριστίες για τη σύζυγο μου, Χαρούλα, γιατί υπήρξε ο πραγματικός στυλοβάτης της προσπάθειας αυτής αλλά και της οικογένειάς μας, μεγαλώνοντας τα δύο μωρά κοριτσάκια μας, Κάλλια και Μαρία, ξεπερνώντας πολλές φορές τον εαυτό της. Χωρίς την πολύτιμη συμπαράστασή της δεν θα βρισκόμουν μπροστά σας σήμερα για να παρουσιάσω την εργασία αυτή.

Μιχάλης Μπλαζαντωνάκης
Χανιά, Ιούλιος 2008

Thesis work was supported by Biopattern, IST EU funded project, Proposal/Contract no.: 508803, and "Gonotypos" projects funded by Greek Secretariat for Research and Technology as well as the Hellenic Ministry of Education.

Abstract

The release of the human genome working draft marked the biomedical discipline opening a new era in the fields of biology and medicine with the use of bioinformatics. In combination with the advent of microarray technology, scientists can now derive a vast amount of valuable information but the need still remains to understand and exploit it. DNA microarray technology allows researchers to study the behavior of thousands of genes in a single experiment, exploring and monitoring their expression in various diseases with the aim of understanding or discovering the biological mechanisms involved. Studying simultaneously this massive gene expression information is a difficult and very demanding task in many ways, making the use of computational and pattern recognition approaches a necessity. Among all thousands of genes studied, many might be irrelevant or redundant to a specific class discrimination problem. This vast amount of data which might contain noise along with redundant information needs to be processed in such a way so that the real valuable and useful knowledge is finally distilled. This "distillate" of marker genes then could be used by an expert to search, discover and understand the hidden biological mechanisms involved in the development of cancer.

One may argue that this problem is a typical feature selection paradigm which could be faced efficiently through a number of typical pattern recognition and machine learning approaches. However the problem referred to as 'curse of dimensionality' is intensified, which may block the effectiveness of an approach. Usually, in such an application domain each sample (patient) is described though a set of genes, yielding a huge dimensional space (order of thousands) covered by a few (order of tenths) patients. Such a sparse covered space may befool a method to a random or unrealistic solution.

The problem is mainly tackled by two types of approaches the filter and the wrapper approach, each one facing the problem from a quiet different perspective: the static one of the filter approach, and the dynamic one through the wrapper approach. Hybridization or integration of these two approaches is an interesting concept which is addressed in the thesis.

In assessing the validity of the derived results we, as many others use various stringent statistical criteria. We take the statistical framework one step further, by studying several issues related to the stability and generalization of algorithmic performance. In addition, we expand our evaluation in assessing the biological significance of results achieved through the use of publicly available and widely accepted biological knowledge. This effort contributes in the establishment of an evaluation framework where the various examined methodologies could be tested in a subjective and fair manner leading to both statistical significant and biologically valid results.

Integration of biological knowledge is an open problem in the field of marker gene selection. Based on the fact that different research teams propose different solutions, with minimal or no overlap at all among them, we propose an evolutionary process of assembling the biological knowledge contributed by different efforts towards a more unified and global approach.

CHAPTER 1

Introduction

Ever since Hippocrates, medicine is in a continuous search path for understanding and revealing the various "mechanisms" that trigger specific diseases, aiming in prognosis, early diagnosis and treatment. Such an aggressive disease which has long before challenged the medical community is cancer. Medical doctors have realized that patient's differ in response to the followed treatment protocols; what is effective for the many may not be effective for the few or vise versa. This well verified and established fact is actually a foreboder of the so called personalized treatment. Having little understanding on what causes such differences and how to best account for them, genomic data/knowledge could be of significant importance towards the aim of understanding and revealing the biological mechanisms hidden behind cancer. Another important issue that motivated research on genomic data is the fact that in many types of the disease, i.e. breast cancer, even though chemotherapy or hormonal therapy reduces the risk of distant metastasis by approximately 1/3, 70-80% of patients receiving this treatment would have survived without it [1]. In addition, despite recommendations by the college of American Pathologists that tumor grade should be used as a prognostic factor in breast cancer, the latest Breast Task Force of the American Joint Committee on Cancer did not include histological tumor grade in its staging criteria, because of insurmountable inconsistencies in histological grading between institutions. Concordance between two pathologists has been investigated and found to range from 50% to 85% [2]. Recent research has demonstrated that gene expression profiling could be much more effective in prognosis of breast cancer than

classic standard grading criteria [3] and also contributes to overcoming such inconsistencies that may exist among doctors or institutions.

1.1 Studying the Genome

With a few exceptions, every cell contains a copy of each of our 30000 genes or so which are expressed (turned on) or unexpressed (turned off) in different cell types. A gene is expressed when it is churning out molecules of messenger RNA (m-RNA).

Thus, we would except a muscle cell to make m-RNA for the various muscle proteins such as actin and myosin but a muscle cell shouldn't produce m-RNA for the pigment melanin or the hormone insulin for instance. Measuring the amount of m-RNA produced by every gene that is expressed in a tissue sample, we could make an 'expression profile' of the biological processes that are triggered in that cell. Then by comparing the different expression profiles among different tissues we may discover the biological mechanisms that make those cells different from each other. DNA microarray technology helps towards this direction by pin pointing all the differences in gene expression between two different cell types.

The surface of DNA-microarray (Figure 1, Figure 2) is divided into thousands of spots, where each spot contains multiple copies of a unique DNA sequence which corresponds to a single gene. Using such a technology we can measure differences between healthy and cancerous cells. Cancer is basically a 'gene disease', many genes control the way cells grow divide and eventually die, when these genes stop working properly cell growth may spin out of control leading to tumor formation and cancer. In order to be able to prognose, diagnose, understand and treat cancer we should identify these genes. The so-called DNA microarray experiment assists researchers towards this task, and enables them to locate differences on the expression levels

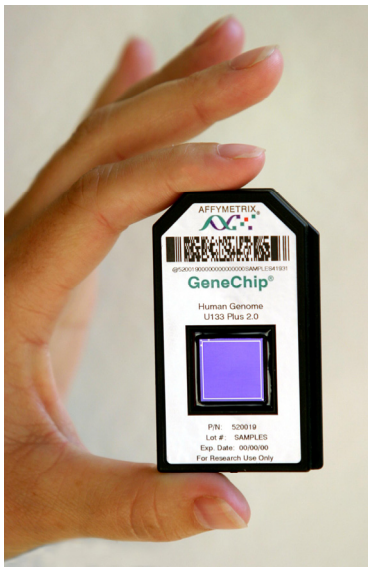

Figure 1: The DNA microarray measures the expression of thousands of genes in a single experiment.

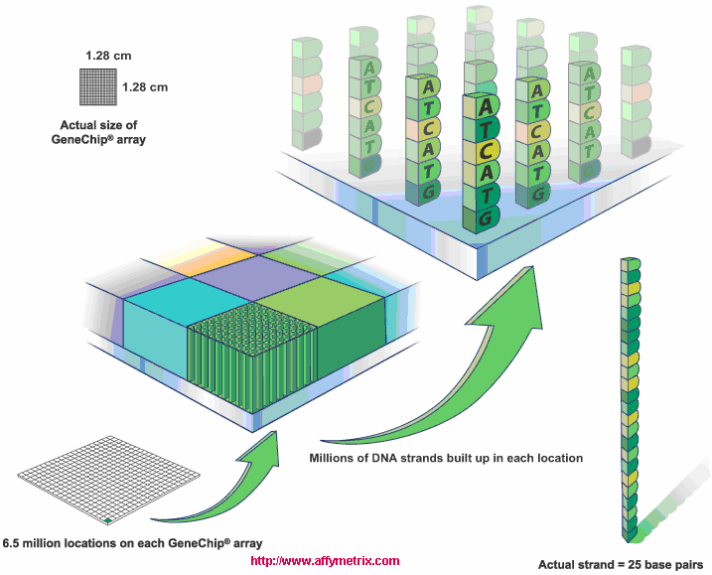

Figure 2: The surface of a DNA microarray is divided into thousands of spots, each spot contains multiple copies of a single gene

between cancerous and healthy cells. This is done by measuring the types of m-RNA found in both types of cells. To accomplish this, healthy and cancerous tissue samples are collected and isolated from the same patient. m-RNA is extracted from both tissue samples by dissolving them in a mixture of various organic solvents. The two samples are labeled with different colors; a red one (cy5) is used to label the cancerous tissue while a green color (cy3) is used for the healthy tissue. A biological process in which complementary part of the m-RNA strand is isolated and degraded into a labeled c-RNA molecule takes place; this is known as the reverse transcriptase process. Next, the hybridization process helps in the completion of the experiment. Through such a

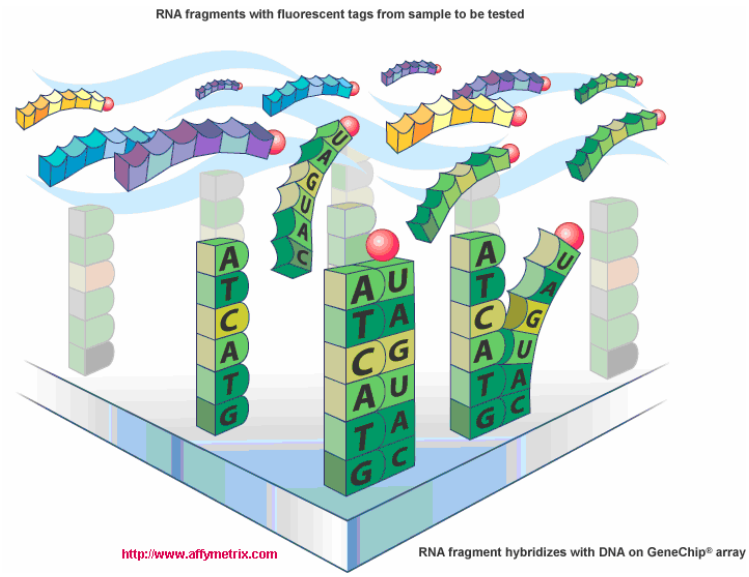

Figure 3: Hybridization process of complementary DNA strands on the microarray surface.

process when two complementary DNA strands are mixed together, they will soon find a correspondence base and pair with each other, it doesn't matter where they come from they will do that even if they come from different sources (Figure 3). The labeled healthy tissue sample is fused onto the surface of the microarray (Figure 3), the c-DNA sequences then hybridize specifically with their corresponding gene sequences in the array. The same process is repeated on a different chip for the

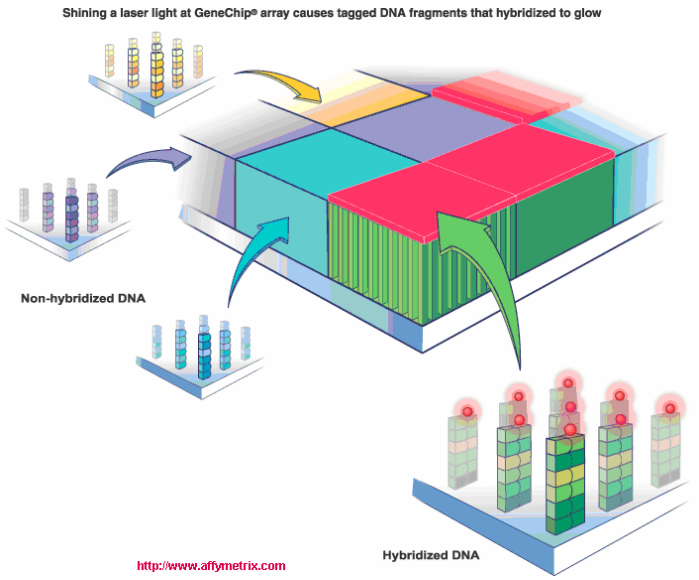

Figure 4: Hybridized DNA microarray, red spots correspond to hybridized cancerous genes; an analogous hybridization scheme for the healthy tissue produces a green spotted microarray.

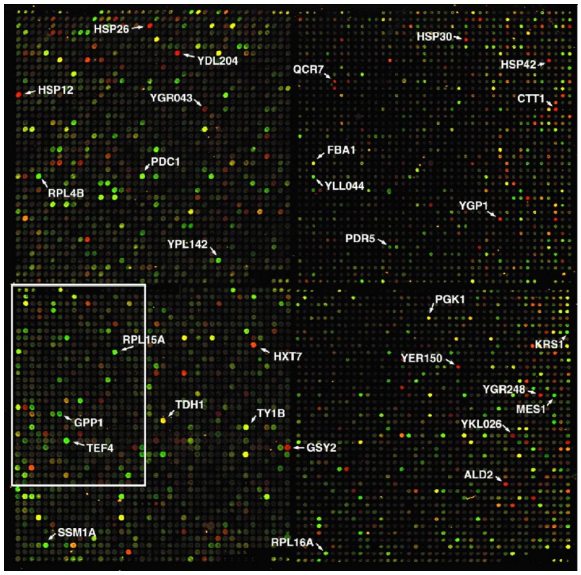

Figure 5: The result of the DNA microarray experiment after the mixture of a double microarray experiment; we can visualize the green, red, yellow and black spots.

cancerous tissue sample, laser light is emitted to both microarrays which makes the hybridized areas to glow.

After the completion of the experiment we have a red spotted array, i.e., an array with red color spots over the area that correspond to hybridized cancerous genes (Figure 4) and a corresponding green spotted array. Overlaying the two images into a single one (Figure 5), we can visualize green spots indicating that the specific genes have been expressed more in the healthy tissue than in the cancerous. Similarly we can visualize red spots which are interpreted in exactly the opposite way, or we can visualize yellow spots implying that the specific genes have expressed themselves in

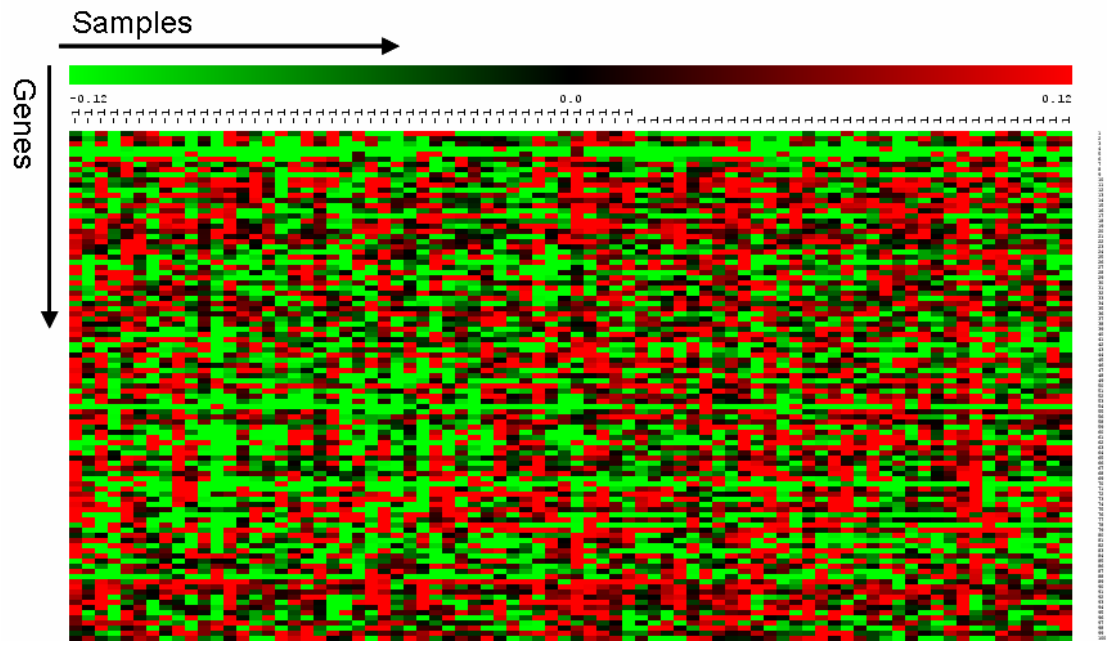

Figure 6: Expression profile analysis of genes after completion of the microarray experiment using log transformations.

approximately the same manner in both situations; black spots correspond to non hybridized genes. Taking $\log_{10}\left(\frac{cy5}{cy3}\right)$ or $\log_2\left(\frac{cy5}{cy3}\right)$ of the intensities of the two color channels, we get a numeric representation of the expression level of each gene, where a negative value indicates a higher expression on the healthy tissue, positive values indicates a higher expression on the cancerous tissue while zero values indicate

approximately the same expression in both tissues. In a post experimental step the expression of each gene (after the log transformation) is kept in a separate cell in an m by n expression matrix M, where each row corresponds to the expression levels of a single gene, while each column corresponds to a different patient.The expression level of each gene can be visualized in terms of a color map varying in a range from green to red for instance, with the mid-point being represented by black (Figure 6). A green colored cell (corresponds to a negative log ratio value) manifests that the specific gene has expressed itself more in the normal than in the pathological state, a red color in a cell (corresponds to a positive log ratio value) implies exactly the opposite, while a black color (corresponds to a zero log ratio value) means that the specific gene has expressed itself in exactly the same way in both situations. Colors are translated into numbers on a closed interval, [-3, +3] for instance, where -3, 0 and +3 indicate green, black and red, respectively.

1.2 Selection of Genomic Markers

The release of the human genome working draft [4] marked the biomedical discipline opening a new era in the fields of biology and medicine with the use of bioinformatics. In combination with the advent of microarray technology, scientists can now derive a vast amount of valuable information but the need still remains to understand and exploit it. DNA microarray technology allows researchers to study the behavior of thousands of genes in a single experiment, exploring and monitoring their expression in various diseases with the aim of understanding or discovering the biological mechanisms involved.

Studying simultaneously the massive gene expression information could be a difficult and very demanding task in many ways. Among all thousands of genes studied, many might be irrelevant or redundant to a specific class discrimination

problem. Many studies have shown that by significantly reducing the number of genes, the generalization performance in classification can be increased; we selectively refer to [1], [5], [6], [7]. This leads to lowering significantly the cost of the experiment, without compromising its value. Furthermore, by focusing on a smaller but representative number of genes, scientists are highly assisted in their task to understand or discover the mechanisms involved in a specific disease, facilitating to early diagnosis, prognosis and drug discovery.

From a pattern recognition point of view, one might argue that the above mentioned problem is a standard feature selection paradigm. This is not a simple task, however, since in such problems we have to face the so called 'curse of dimensionality'. We are provided with a small number of samples (order of some tenths) compared to a very large number of features (genes at the order of thousands), rendering the solution quite ill posed and necessitating the use of prior information in the form of constraints for its regularization. This algorithmic issue provides an extra motive to significantly reduce the number of important genes, since the performance of any data mining procedure depends on the ratio between the number of training samples and the number of features.

The primary goal of any gene selection method is to find a set of genes (markers) with size much smaller than the initial, which is able to describe the data set of interest fairly well both in terms of classification accuracy and quality. The first attribute (accuracy) relates to the ability of the selected genes to successfully classify samples into their correct class, whereas the quality attribute reflects the ability of each gene to clearly differentiate its expression between the states of interest. This fact has been implicitly implied in almost every gene selection study (we selectively refer to [14] - [18]) and has also been explicitly stated in [6], [19]-[21]. The concise

study of a small number of genes can help biologists to get significant insight into the genetic structure and mechanisms involved in a specific disease, which may lead to drug discovery and early diagnosis.

The most important advantages of marker selection are summarized as:

1. Classification accuracy can be improved by selecting a small but representative number of genes; we selectively refer to [6], [7].
2. Expression arrays recording the behavior of thousands of genes are impractical to be used and studied by biologists. The experts would prefer to monitor and study changes in a smaller set of genes [22], which will assist their task to discover the biological processes involved in the development of the disease.
3. By reducing the number of genes to be studied, the cost of the examination is also significantly reduced.

Marker selection studies usually address three commonly encountered types of objectives [23].

1. Class comparison: Is the comparison of gene expression in different groups of specimens. The specific objective of such a study is to determine whether the expression profile of the derived gene signature is different between the classes.
2. Class prediction studies give emphasis on developing a gene signature that actually predicts class membership of new samples on the basis of the expression levels of the genes in the derived gene signature. Finally,
3. Class discovery is fundamentally different from class comparison or class prediction. In such studies the classes are not predefined and usually clustering methodologies are used to reveal possible discrete subsets of disease entities

which possible define different disease subgroups. Using such an approach Golub et al. in [1] managed to discover the two basic types of leukaemia i.e., Acute Lymphoplastic Leukemia (ALL) and Acute Myeloid Leukemia (AML) but additionally it also discovered the two subtypes B-Cell and T-Cell of the ALL type.

1.3 Algorithmic Approaches to the Problem

Feature selection methods are divided into two categories [25], i.e. the so called Filter and Wrapper methods. Filter approaches give emphasis and focus on intrinsic characteristics of data neglecting however gene interactions [26], they rank genes according to how they score on various stochastic measures such as Fisher's ratio, T-statistics, χ^2 statistic, information gain, Pearson correlation and many others, the highest rank genes that give the maximum classification performance are the genes that constitute the final derived gene signature. A very close alternative to the Fisher's coefficient that has been widely applied in the field of marker gene selection is given below:

$$f(g_i) = \frac{\left|\mu_+(g_i) - \mu_-(g_i)\right|}{\sigma_+(g_i) + \sigma_-(g_i)} \tag{1.1}$$

where, $\mu_+(g_i)$, $\mu_-(g_i)$, $\sigma_+(g_i)$, $\sigma_-(g_i)$ is the mean and standard deviation values of gene i in positive and negative class respectively. Such a criterion aims at genes which differentiate their expression more in the two classes and hence it is searching for a gene signature that shares the intrinsic property of low intra-class but high inert-class distance. Wrapper methods on the other hand use a classifier to assign scores and rank genes; classifiers usually generate weight vectors which are used as gene scores. The genes are ranked according to such scores, the lowest rank gene(s) is (are) eliminated and the process continues in an iterative manner. In filter methods the

feature ranking criterion remains stable along the gene selection process while wrapper methods re-evaluate and dynamically update the criterion from iteration to iteration, thus an insignificant gene in one iteration, may become significant in the next or vise versa.. Another fundamental difference which should be emphasized is that filter methods focus on intrinsic data characteristics neglecting gene interactions while wrapper methods behave in exactly the opposite way underlying a major gap in the 'philosophy' of the two approaches. Although it has been demonstrated by various studies that wrapper methods outperform their filter alternatives accuracy-wise [6], [27], a little attention has been paid to the 'quality' aspect of the derived result. By the term quality we address the ability of a gene signature to significantly differentiate its expression from one class to the other; this aspect refers to the class comparison criterion addressed earlier in section 1.2. One aim of this study is to propose a methodological platform serving as a vehicle of bridging the gap between the two quite different philosophies but also provide an evaluation framework which takes into account not only the accuracy criterion but also additional quality aspects of the derived result.

1.4 Background knowledge in terms of gene ontologies and pathways

Gene Ontology project [28] provides a controlled vocabulary to describe gene and gene product attributes in any organisms. The Gene Ontology Biological Processes (GOBP) constitute basic background knowledge organizing genes into ensembles according to the biological process they participate [28]. Using ontologies we can find a collection of genes that are involved in a specific biological process or find the biological processes that a specific gene is involved to. Notice that while there are uniquely identified biological processes, a specific gene may be involved in more than one of them. Hence, we are given background biological knowledge on the genes that

constitute specific biological ‘mechanisms’ but also the mechanisms associated with a specific gene.

GOBPs are structured as directed acyclic graphs, similar to hierarchies where a more specialized term (child) can be related to more than one less specialized terms (parents). Graph nodes are connected together through two relationships, the ‘is a’ and ‘part of’ relationships. Using such relationships biologists can build up a biological process as the immune system process for instance, depicted in Figure 7. The GO consortium serves a broad variety of needs, assessing the problem of consistent descriptions of gene products in different databases. In this thesis we focus only on the GOBP components and primarily on the genes that constitute them, while a deeper and thorough analysis of the underlined processes and mechanisms used to assess project’s aim is beyond the scope of this work. For a more detail description on such aspects the interested reader may refer to [28].

A pathway on the other hand is a series of biochemical reactions occurring within a cell. Such processes are usually rapid, lasting on the order of milliseconds in the case of ion flux, or minutes for the activation of protein, but some can take hours and even days (as is the case with gene expression) to complete. To better organize the various biological concepts included in the pathway formation process, we present the hierarchical structure in Figure 8. Chromosomes induce genes that produce proteins used in chemical reactions occurring within a cell. A set of such chemical reactions serving a specific purpose constitutes a pathway [28]. For instance, the IL-10 anti-inflammatory pathway is depicted in Figure 9, where we notice that additional pathways could contribute to the formation of a new one, while many pathways could react within a cell. Hence, GOBPs constitute a conceptual network of biological processes dynamically adapted and updated with evolving knowledge, while a

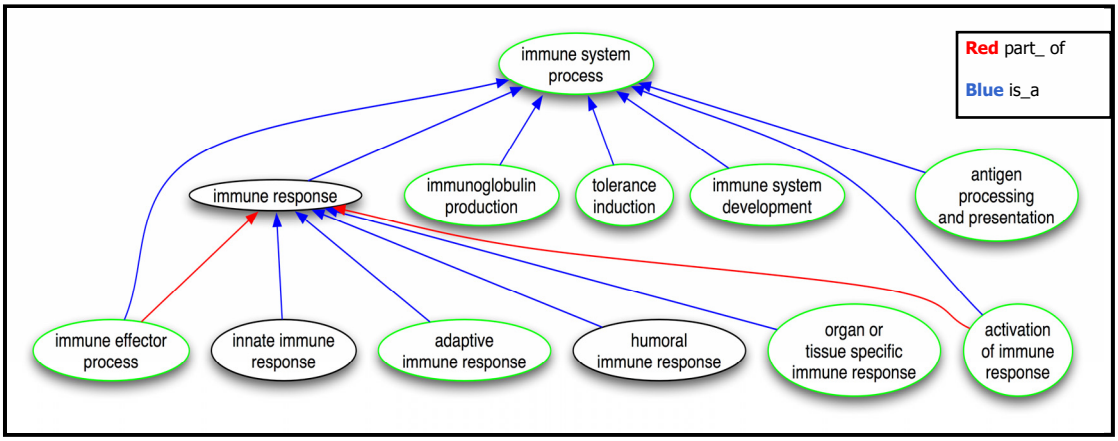

Figure 7: Bulding the gene ontology immune system process in temrs of is_a and part_of relationships.

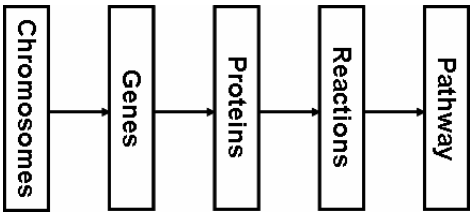

Figure 8: Pathway-Genome hierarchical structure.

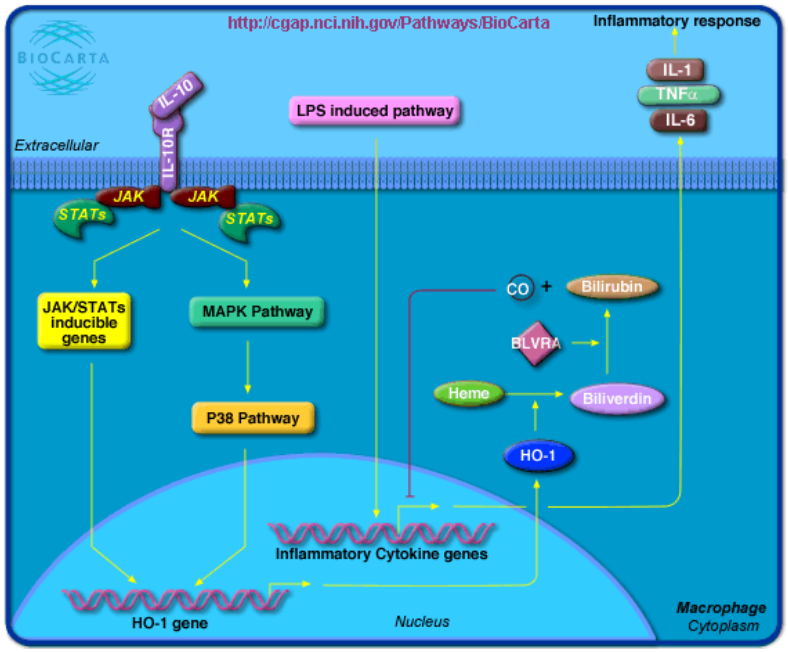

Figure 9: IL-10 Anti-Inflammatory signaling pathway.

pathway refers to very specific and strict biological functions accomplished through a series of bioechemical reactions. GOBPs could be seen as candidates for constituting future pathways.

In this thesis we focus on the 20 pathways published by NetPath [29], 10 of which are related to immune system (immune signaling pathways) and 10 related to cancer (cancer signaling pathways). In subsequent chapters we are using background knowledge provided through GOBPs and NetPath to either validate or enhance derived results.

1.5 State of the art in Gene Selection for Disease Diagnosis

In this section we focus on two application domains, namely breast cancer [1] and leukemia [5], we overview benchmark results derived in those domains, constituting a reference base line.

1.5.1 Breast Cancer

Laura Van't Veer and colleagues in [1] derived a 70 gene signature, using a filter approach, able to discriminate between the two prognostic groups in breast cancer. The two prognostic groups correspond to those patients that after treatment or operation, a relapse didn't occur for a period of at least 5 years and belong to the good prognosis group, while patients for whom a relapse occurred within a 5 year period correspond to the poor prognosis group.

Classification on the training set of 78 patients is depicted in Figure 10 (panel b). Patients below the dashed line have a good prognosis profile while the prognosis is poor for patients above the dashed line. Classification on a test set of 19 new patients is depicted in Figure 10 (panel c) classifying correctly 19/17 patients (89.47%) success rate. The same gene signature was further validated on 295 patients [3] (234 new patients + 61 patients that were included in the previous study [1]), Figure 11.

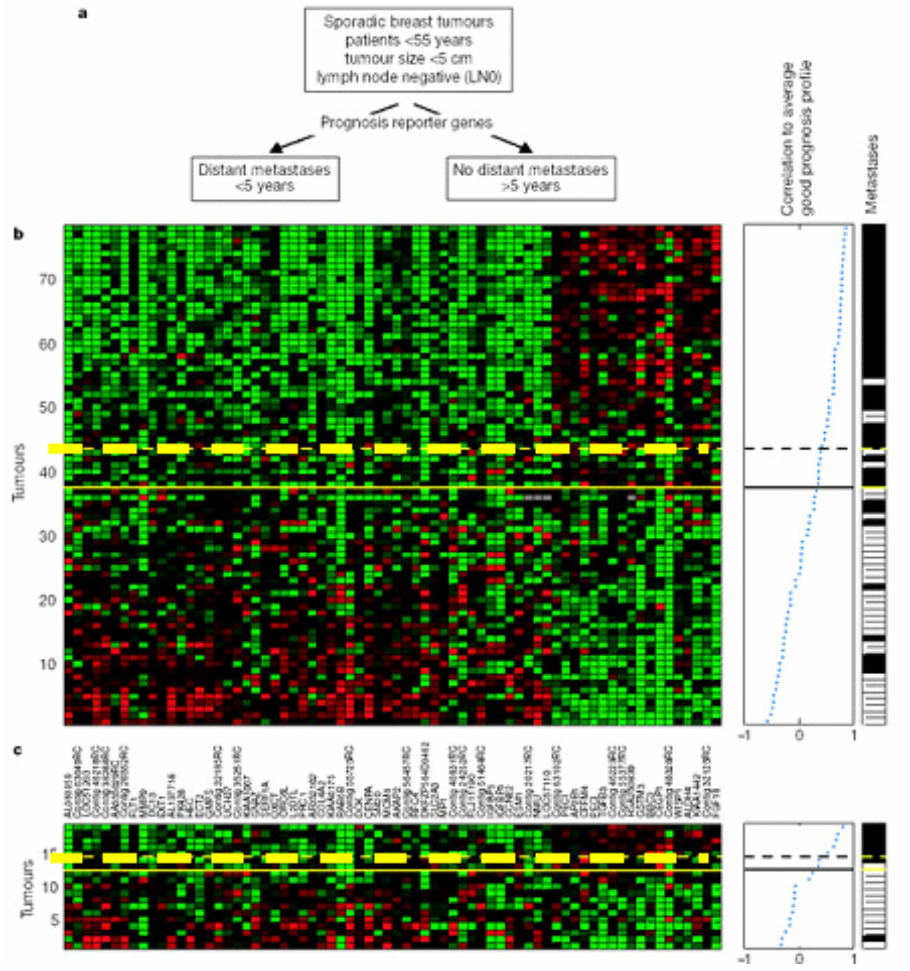

Figure 10: Supervised classification, patients correspond to rows, genes to columns. Genes are ordered according to their correlation coefficient with the two prognostic groups. Patients are ordered by the correlation coefficient to the average profile of the good prognosis group. Above the dashed line patients correspond to the good prognosis signature; below the dashed line the prognosis signature is poor.

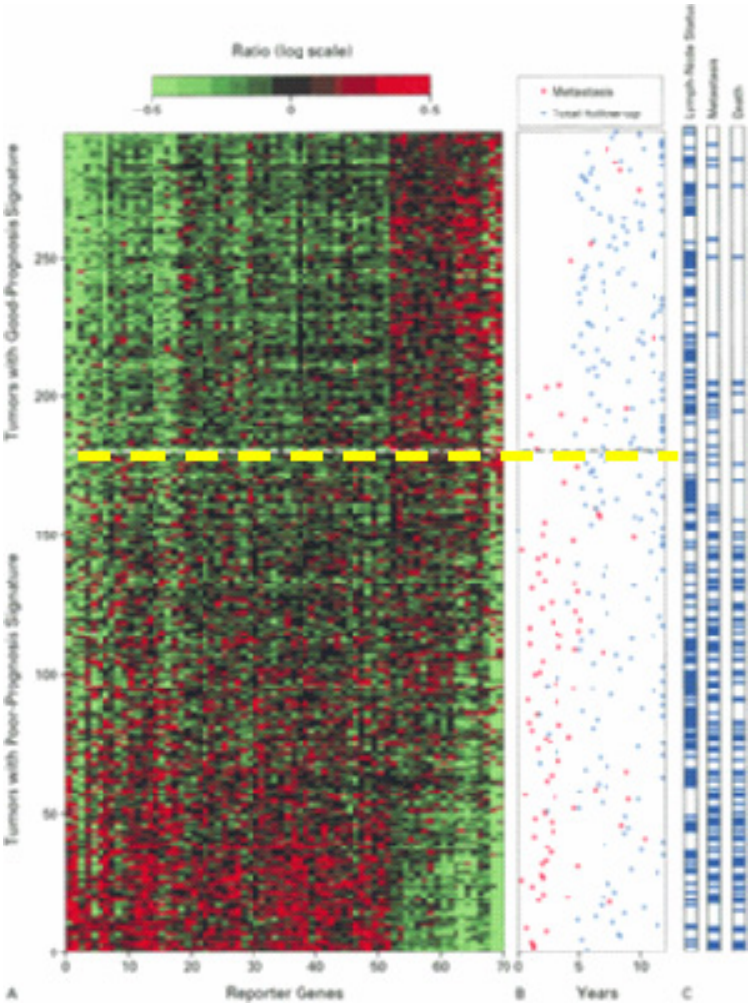

Figure 11: Panel A shows the expression profile of the 70 marker genes (columns) in a series of 295 patients (rows). Tumors are rank ordered according to their correlation with the previously determined average profile, while genes are ordered according to their correlation with the two prognostic groups. Panel B shows the time in years to distant metastasis as a first event for those in whom this occurred, and the total duration of follow up for all other patients.

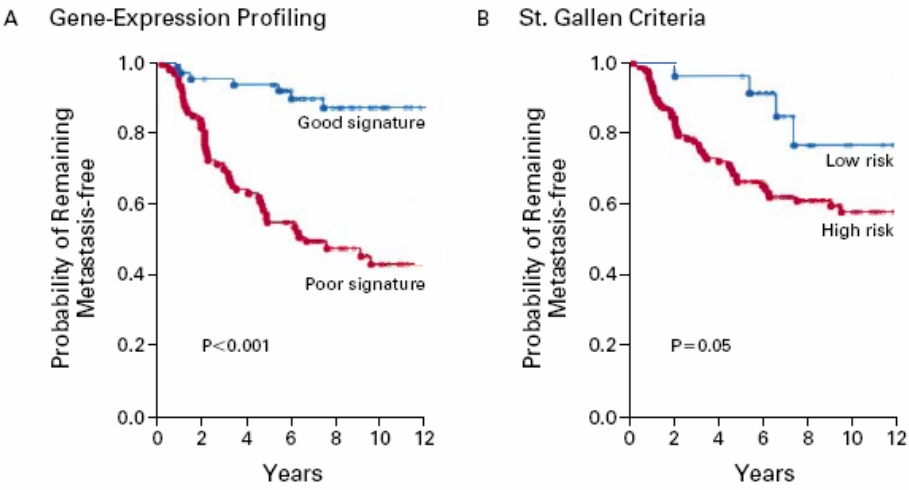

Figure 12: Survival prediction on the classification result of the 70-gene signature on a series of 295 patients. The gene expression profiling (Panel A) produced more significantly differentiated survival curves corresponding to the two prognostic groups than the standard St. Gallen Criteria.

Taking follow up times of the 294 patients the classification result produces the survival curves depicted in Figure 12 (panel A), where we observe that the probability of reaching a 12 year survival for the good prognosis group is approximately 0.9, while the corresponding probability for the poor prognosis falls below 0.5, indicating that the expression profile analysis of the 70-gene signature could be applied as a reliable clinical outcome predictor. Additionally the survival prediction derived through the 70-gene signature outperforms significantly the one derived by the standard St. Gallen risk criteria, indicating that the derived gene signature can help doctors in deciding for the treatment protocol in a more effective and reliable approach.

1.5.2 Leukemia

Leukemia is basically distinguished in two types AML (Acute Myeloid Leukemia) and ALL (Acute Lymphoblastic Leukemia). Although the distinction between ALL and AML has been well established, no single test is currently available to establish the diagnosis. Rather, current clinical practice involves an experienced hematopathologist's interpretation of the tumor's morphology, histochemistry, immunophenotyping, and cytogenetic analysis, each performed in a separate, highly specialized laboratory. Although usually accurate, leukemia classification remains imperfect and errors do occur.

Distinguishing ALL from AML is critical for successful treatment; chemotherapy regimens for ALL generally contain corticosteroids, vincristine, methotrexate, and L-asparaginase, whereas most AML regimens rely on a backbone of daunorubicin and cytarabine. Although remissions can be achieved using ALL therapy for AML (and vice versa), cure rates are markedly diminished, and unwarranted toxicities are encountered. Golub et al. [5], using a filter method in combination with a variation of

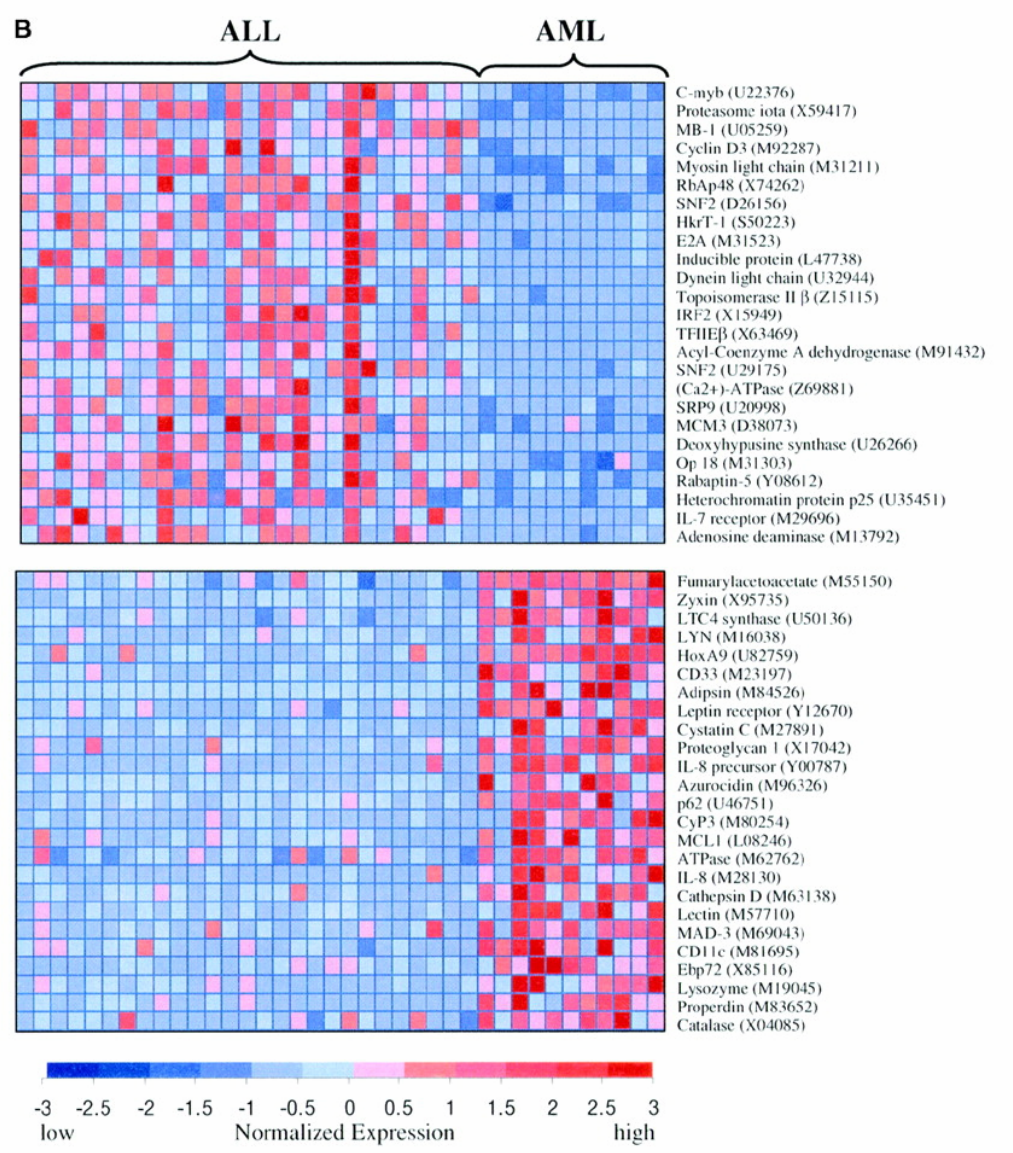

Figure 13: The 50 genes mostly correlated with AML and ALL, each row corresponds to a gene and each column corresponds to a patient. Top panel shows genes more highly expressed in ALL , bottom panel shows genes more highly expressed in AML.

Fisher's ratio derived a 50 gene signature that was able to discriminate between the two types of leukaemia. The 50-gene predictor in a leave one out cross validation procedure classified correctly 36/38 samples of the training set, while it classified perfectly 34 samples that were used as an independent test set. The expression of the 50-gene signature on the training set is depicted in Figure 13. Applying a wrapper approach in combination with an SVM classifier Guyon et al. in [6] improved even further the result by deriving an 8-gene signature able to distinguish perfectly the two types of leukaemia. An even better result is reported in [7], where a wrapper method in combination with a Ridge Regression Classifier derived a 3-gene signature able to distinguish also perfectly between the AML and ALL type. Even though authors of these two studies did not report the gene names, still the results are very impressive opening the road to wrapper methods and pattern recognition approaches.

1.5.3 Related Work

To effectively address the problem of gene marker selection one has to take into account its various peculiarities. In this section we point out some important aspects of the problem along with related work. One such aspect that needs special attention is the existence of multiple equivalent solutions due to 'curse of dimensionality' which leads to a huge and sparsely covered search space; this issue has been addressed by Ein-Dor et al. in [8].

Some studies distinguish the task of classification from that of feature selection, but many published results contradict each other. We refer to the study of Niijima and Kuhara [9], in which by applying a filter method along with a nearest mean classifier their results contradict the findings of Guyon et al. [6] in colon cancer. The selection of the training set could play a crucial and catalytic role on the selection of the final gene signature. This is issue has been addressed by Michiels et al.[10] showing the

strong dependence of a molecular signature on the patients selected to constitute the training set. The authors propose a strict approach for assessing the statistical validity of a gene signature, by appropriately derived confidence intervals.

The aspect of bias in the gene selection problem is an issue that needs special attention and has been effectively addressed by R. Simon et al. [11] showing that wrapper methods introduce a large amount of bias when using internal evaluation criteria. Ambroise and McLachlan [12] verify this fact but also demonstrate that the bias is corrected when external evaluation criteria are used.

Evaluating cross platform performance of a gene signature is still an open issue on the problem of marker gene selection. Experiments are conducted using a) different microarray platforms, b) different protocols, c) different populations and d) different experimental set up; some research groups use a double array while other groups use a single array experiment. Such differences raise limitations on the cross platform evaluation of results Yu et al. [13] address such issues and derive a 62-gene expression signature that could predict effectively the prognosis group of estrogen receptor positive patients in two different cross platform evaluation sets.

1.6 Our contribution

In this study we address the problem of marker gene selection, mostly from the view point of wrapper methods, revealing their advantages and disadvantages on various domains (diseases) of interest. It is evident to the reader by now that there are two quite different approaches addressing the problem of marker gene selection (section 1.3), the Filter and the Wrapper one, each demonstrating its advantages and disadvantages depending on the application domain. Beyond results, comparison criteria, effectiveness or reliability of each approach, it becomes apparent that integration or hybridization of those two quite different approaches into a single task

could be an effective alternative to address the problem. Furthermore, integration of statistical results with biological knowledge is essential for validating any gene signature on a meaningful clinical basis. Thus, we propose

1. An integration platform where filter and wrapper methods could be hybridized with one another aiming in the improvement of the produce result.
2. Application of such an integration platform in breast cancer to derive comparable to the state of the art gene signatures. In addition we suggest an evaluation framework where the various methods could be tested in an objective, reliable and fair manner, using stringent statistic criteria.
3. Besides the proposed statistical criteria which address the significance of results from a statistical point of view, we also propose a framework for biological validation and integration through appropriate use of GOBPs and Pathways.

1.7 Thesis Overview

The thesis is organized into 5 chapters following a conceptual evolution of results. Each chapter is accompanied with an abstract and an introduction section to assist reader's focus, along with partial conclusion at the end of each chapter. In Chapter 2 we proceed by proposing a framework where Filter and Wrapper approaches could be effectively hybridized with each other. The proposed approach is tested in comparison with a representative wrapper method based on support vector machines. We alternatively refer to our publications relative to the work assessed in this chapter [21], [30], [31], [32], [33], and [34]. In Chapter 3 we further validate our approach on additional data sets, but we also establish an evaluation framework where various methodologies could be evaluated in a fair and subjective manner. Through such an evaluation we derive at a promising 57-gene breast cancer signature [35], giving

significant statistical results compared to other bench mark studies. In Chapter 4, we assess the biological significance of the derived gene signature, demonstrating that besides its statistical importance it is also compliant with valid biological knowledge [36]. In Chapter 5 we investigate the biological knowledge hidden behind the 70-gene Van't Veer's signature (section 1.5.1). We proceed by integrating it with the knowledge hidden behind our 57-gene signature derived in Chapter 3, proposing a process of biological knowledge evolution, leading to an approach of unfolding the biological mechanisms which might be involved in breast cancer [37].

References

[1] L. J. Van't Veer, H. Dai, M. J. Van de Vijver, Y. D. He, et al., Gene expression profiling predicts clinical outcome of breast cancer. *Letters to Nature* 415 (2002) 530-536.

[2] C. Sotiriou, P. Wirapati, S. Loi, A. Harris et al. Gene Expression Profiling in Breast Cancer: Understanding the Molecular Bias of Histologic Grade to Improve Prognosis. Journal the National Cancer Institute 98 (2006) 262-272.

[3] M. J. Van De Vijver, Y. D. He, L. J. Van't Veer, et al., A gene expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine,* 347(2002), 1999-2009.

[4] University of California Santa Cruz Genome Bioinformatics, Human Genome Working Draft, http://genome.ucsc.edu, (2001).

[5] R. T. Golub, K. D. Slonim, P. Tamayo, C. Huard, et al., Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, *Science*, 286 (1999) 531-536.

[6] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, Gene selection for cancer classification using Support vector machines, *machine learning*, 46 (2002) 389-422.

[7] F. Li and Y. Yang, Analysis of recursive gene selection approaches from microarray data. *Bioinformatics*, 21 (2005), 3741-3747.

[8] L. Ein-Dor, I. Kela, G. Getz, D. Givol and Eytan Domany, Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* 21 (2005) , pp 171-178.

[9] S. Niijima, S. Kuhara, Recursive gene selection based on maximum margin criterion: a comparison with SVM-RFE, *BMC Bioinformatics* 7:543 (2006).

[10] S. Michiels, S. Koscielny and C. Hill C., "Prediction of cancer outcome with microarrays: a multiple random validation strategy", *Lancet*, 365 (2005) pp. 488-492, 2005.

[11] R. Simon., M. D. Radmache, K. Dobbin. and L. M. McShane, "Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification", *J. Natl Cancer Inst.*95 (2003), pp. 14-18.

[12] C. Ambroise, G. J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data", *Proc. Natl. Acad. Sci. USA,* 99 (2002), pp. 6562-6566.

[13] K. Yu, C. H.Lee, P.H. Tan, G. S. Hong, S. B. Wee, C. Y. Wong, and P. Tan, "A Molecular Signature of the Nottingham Prognostic Index in Breast Cancer", Cancer Research 64 (2004), pp.2962-2968.

[14] S. Armstrong, J. Staunton, L. Silverman, R. Pieters, et al., MLL translocations specify a distinct gene expression profile that distinguishes a unique leukaemia, *nature genetics*, 30 (2002), 41-47.

[15] C. Nutt, D. Mani, R. Betensky, P. Tamayo et al., (2003), Gene Expression-based Classification of Malignant Gliomas Correlates Better with Survival than Histological Classification. *Cancer Research*, 63 (2003) 1602-1607.

[16] S. Ramaswamy, K. Ross, E. Lander and T. Golub, A molecular signature of metastasis in primary solid tumors, *nature genetics* 33 (2003) 49-54.

[17] D. Singh, P. Febbo, K. Ross, J. Donald, et al., Gene expression correlates of clinical prostate cancer behaviour. *Cancer Cell* 1 (2002) 203-209.

[18] R. Simon, M. D. Radmacher, K. Dobbin, L. M. McShane, Pitfalls in the use of DNA Microarray Data for Diagnostic and Prognostic Classification, *Journal of the National Cancer Institute* 95 **(**2003) 14–18.

[19] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, et al., Gene expression profiles of herediatary breast cancer, *The New England Journal of Medicine*, 344 (2001) 539-548.

[20] T. Hastie, R. Tibshirani, B. M. Eisen, A. Alizadeh, et al., "Gene shaving" as a method for identifying distinct sets of genes with similar expression patterns, *Genome Biology*, 1 (2000), 1-21.

[21] M. E. Blazadonakis, M. Zervakis, Wrapper Filter Criteria Via Linear Neuron and Kernel Approaches, *Computers in Biology and Medicine*, To appear.

[22] Roth, F. P. Bringing out the best features of expression data. *Genome Research* , 11 (2001) 1801-1802.

[23] Simon, R., Radmacher, M. D., Dobbin, K. and McShane, L. M., "Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification", *J. Natl Cancer Inst.* 95 (2003), 14-18, 2003.

[24] Y.Wang, J. G. M. Klijn, Y. Zhang, A. Sieuerts, M. P. Look, Fei Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, T. Jaktoe, E. MJJ Berns, D. Atkins and J. A. Foekens, Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet* 365 (2005) 671-679.

[25] R. Kohavi and G. H. John, Wrappers for feature subset selection, *Artificial Intelligence*, 97 (1997), 273-324.

[26] S. G. Baker and B. S. Kramer, Identifying genes that contribute more to good classification in microarrays, *BMC Bioinformatics*, 7 (2006), 407.

[27] I. Inza, P. Larranaga, R. Blanco and A. J. Cerrolaza, Filter versus wrapper gene selection approaches in DNA microarray domains, Artificial Intelligence in Medicine 31 (2004), 91-103.

[28] http://www.geneontology.org/

[29] http://www.netpath.org/

[30] M. E. Blazadonakis and M.Zervakis, "Support Vector Machines and Neural Networks as Marker Selectors in Cancer Gene Analysis", book chapter in "Intelligent Techniques and Tools for Novel System Architectures", Series of Computational Intelligence, Springer Verlag, to be published end of July 2008.

[31] M. E. Blazadonakis, A. Perperoglou and M. Zervakis, "Using a Single Neuron as a Marker Selector – A Breast Cancer Case Study", *"Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society"*. Lyon France Aug. 23-26 2007, pp. 4219-4222

[32] M. E. Blazadonakis and M. Zervakis, "Support Vector Machines and Neural Networks as Marker Selectors for Cancer Gene Analysis", *Proceeding of 3rd International IEEE Conference on Intelligent Systems*. London Sept. 4-6 2006, pp. 626-630.

[33] M. E. Blazadonakis and M. Zervakis, "Polynomial and RBF Kernels as Marker Selection Tools – A Breast Cancer Case Study", *"Proceedings of the 6th International Conference on Machine Learning and Applications"*. Cincinnati Ohio 13-15 Dec. 2007, pp. 488-493.

[34] M E. Blazadonakis and M Zervakis, "Improving Filter Methods Through Wrapper Approaches - A Breast Cancer Case Study", *"Proceedings of the 3rd International Conference on Computational Intelligence in Medicine"*. Plymouth England 25-27 July 2007.

[35] M. E. Blazadonakis and M.Zervakis, "The Linear Neuron as Marker Selector and Clinical Predictor". *Computer Methods and Programs in Biomedicine*, Vol 91/1 pp. 22-35

[36] M. E. Blazadonakis and M. Zervakis, "Revealing Significant Biological Knowledge via Gene Ontologies and Pathways", *2008 International Conference in Biomedical Engineering and Informatics (BMEI 2008)*, Sanya, Hainan , China, to appear.

[37] M. E. Blazadonakis and M. Zervakis, "Integrating Biological Knowledge for Marker Gene Selection in Breast Cancer" under submission in BMC-Cancer.

CHAPTER 2

Wrapper Filtering Criteria Via a Linear Neuron and Kernel Approaches

2.1 Abstract

Objective: In this chapter we aim at integrating the filter with the wrapper approaches by applying filter criteria in a recursive fashion, where weights are potentially adjusted from iteration to iteration, producing noticeable improvement on the generalization performance measured on independent test sets.

Methods and Materials: Towards this direction we explore the behavior of two well known and broadly accepted pattern recognition approaches namely the Support Vector Machines (SVM) and a *single* Linear Neuron (LN), properly adapted to the problem of marker selection. Within this context we also show how the kernel ability of SVM could be employed in a practical manner to provide alternative ways to approach the problem of reliable marker selection.

Results: We explore how the proposed approaches behave in two application domains (breast cancer and leukemia), achieving comparable with, or even better results than those reported in the related bibliography. An important advantage of these approaches is their ability to derive stable performance without deteriorating due to the complexity of the application domain. Validation is performed using Internal Leave One Out (ILOO) and 10-fold cross validation as well as independent test set evaluation. Results show that the proposed methodologies achieve remarkable performance and indicate that applying filter criteria in a wrapper fashion ('wrapper filtering criteria') provides a useful tool for marker selection. The contribution of this study is three-fold. First it provides a methodology for integrating the filter with the wrapper approach, second it introduces a fundamental pattern recognition component

namely the single neuron (which is a linear estimator) and explores its behavior on marker selection and third, it demonstrates an approach to exploit the kernel ability of SVMs in a practical and effective manner.

2.2 Introduction

Although it has been demonstrated that performance of wrapper methods is superior to those of filter [1], [2], many experts prefer using filter methods [3], [4], [5], [6]. In these studies mostly variations of Fisher's ratio are used as the basic tool for marker selection, relying very much on desirable intrinsic characteristics of selected features, i.e. the differential expression of genes in the two classes of interest. This aspect is not addressed by wrapper methods that focus only on classification neglecting such characteristics. Even though a number of wrapper approaches based mostly on support vector machines have been introduced in the recent years [1], [7], [8], no attempt to addresses the concept of integrating the wrapper with the filter approaches has been addressed so far. One of the closest approaches to this direction is the Gene Expression Model Selector (GEMS) [9], [10]. It uses a filter criterion to rank genes, while the selection process prefers the top most genes that maximize classification performance. In this study we go one step further and address this integration concept along with the motive of embedding intrinsic data characteristics into classical pattern recognition tools, thus implanting filtering criteria into a wrapper operation, and eventually improving performance over existing wrapper methods. Our motive can also be seen as an attempt to bridge the gap between the two 'philosophies'. One proposed method is based on a Linear Neuron (LN) enriched with a variation of Fisher's metric and is referred to as Recursive Feature Elimination based on Linear Neuron Weights (RFE-LNW). A second approach is based on SVMs and is referred to as RFE based on the Fisher's metric and Support Vectors (RFE-FSVs). It makes

use of a fundamental property of support vector machines, which enables us to exploit various kernels in combination with the Fisher's ratio. More details on these two methodologies are provided in sections that follow. The proposed algorithms are tested under the same conditions and using various types of experiments in order to reveal the consistency of their performance.

A final point we address in this study is that of linearity. Due to the high dimensionality of the problem, most studies approach marker selection as a linear one using linear approaches. However, the field of interest is very complex with thousands of features and probably innumerable interconnections and dependencies among those features. A question of interest is immediately posed as: could non linear kernels help towards a better solution? The experimental section demonstrates that nonlinear approaches could indeed be useful to the process of marker selection, especially towards the end of the selection process where only a few markers have survived.

2.3 Methods

2.3.1 Background Knowledge on SVMs and GEMS

SVM [11] attempts to find the best separating hyperplane to distinguish between the two classes of interest, positive (+1) and negative (-1). This is done by maximizing the distance $\frac{2}{\|\boldsymbol{w}\|}$ between the two parallel lines $(\boldsymbol{w}\cdot\boldsymbol{x})+b=1$ and $(\boldsymbol{w}\cdot\boldsymbol{x})+b=-1$, which form the margin of separation of the two classes as shown in Figure 14. The final separating hyper-plane passes through the middle of this margin with equation $(\boldsymbol{w}\cdot\boldsymbol{x})+b=0$. The decision function then, is a function of the form:

$$f\left(x\right)=\operatorname{sgn}\left(\left(\boldsymbol{w}\cdot\boldsymbol{x}\right)+b\right) \tag{2.1}$$

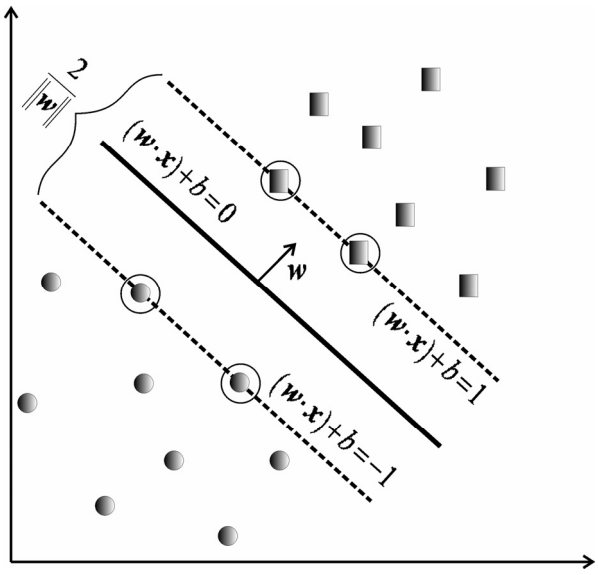

Figure 14: Illustration of the binary classification problem, showing the margin of separation between the two classes; circled points represent the support vectors.

where $\boldsymbol{w}$ represents the direction vector of the hyper-plane. The sign of the value returned by equation (2.1) indicates the predicted class associated with example $\boldsymbol{x}$, while $|f(x)|$ indicates the confidence level of the resulting decision. The SVM problem can be equivalently formulated as follows:

$$\begin{aligned} &minimize \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{j=1}^{n}\xi_j^2 \\ &subject\ to\ y_j\left(\left(\boldsymbol{w}\cdot\boldsymbol{x}_j\right)+b\right) \geq 1-\xi_j,\ \xi_j \geq 0,\ j=1,\ldots n \end{aligned} \tag{2.2}$$

By the duality theory, a tutorial of which can be found in [12], the problem can be transformed to the following maximization problem, where λ represents the vector of Lagrange multipliers and y_i represents the label (either +1 or -1) of the ith sample:

$$\begin{aligned} &\underset{\boldsymbol{\lambda}\in\mathbb{R}^n}{maximize} \sum_{j=1}^{n}\lambda_j - \frac{1}{2}\sum_{i,j=1}^{n}\lambda_i\lambda_j y_i y_j\left(\boldsymbol{x}_i\cdot\boldsymbol{x}_j\right) \\ &Subject\ to \begin{cases} \sum_{j=1}^{n}\lambda_j y_i = 0 \\ 0 \leq \lambda_j \leq C,\ j=1,\ldots,n \end{cases} \end{aligned} \tag{2.3}$$

Towards the solution of this problem, we obtain the following expression for the direction vector $\boldsymbol{w}$:

$$\boldsymbol{w} = \sum_{j=1}^{n} \lambda_j y_j \boldsymbol{x}_j \tag{2.4}$$

which is actually an expansion of those training samples with non-zero λ_j, i.e. the support vectors. It can be proved that support vectors lay on the borders of the class regions (as Figure 14 illustrates) and can be used to find b by substituting one of the support vectors to the following equation:

$$y_j \left(\left(\boldsymbol{w} \cdot \boldsymbol{x}_j \right) + b \right) = 1 \tag{2.5}$$

An important issue making SVMs very attractive is that they allow the use of kernels, so that the dot product in equation (2.3) can be replaced by a kernel function in the following form:

$$\begin{aligned} &\underset{\boldsymbol{\lambda} \in \mathbb{R}^n}{maximize} \sum_{j=1}^{n} \lambda_j - \frac{1}{2} \sum_{i,j=1}^{n} \lambda_i \lambda_j y_i y_j k\left(\boldsymbol{x}_i, \boldsymbol{x}_j \right) \\ &Subject\ to \begin{cases} \sum_{j=1}^{n} \lambda_j y_j = 0 \\ 0 \le \lambda_j \le C, i = 1, \ldots, n \end{cases} \end{aligned} \tag{2.6}$$

Besides the linear kernel in equation (2.3), other types of kernels such as polynomials of any degree, as well as Radial Basis Functions (RBF) can be used in the forms of:

$$\begin{aligned} k\left(\boldsymbol{x}, \boldsymbol{y} \right) &= \left(1 + \left(\boldsymbol{x} \cdot \boldsymbol{y} \right) \right)^d \\ k(\boldsymbol{x}, \boldsymbol{y}) &= \exp\left(-\gamma \left\| \boldsymbol{x} - \boldsymbol{y} \right\|^2 \right) \end{aligned} \tag{2.7}$$

For a detailed essay on SVM the interested reader may refer to [11].

GEMS (Gene Expression Model Selector) [9], [10] is a wrapper gene selection method that employees an SVM classifier to assist its task. The heart of GEMS

consists of a nested cross validation procedure presented in table 1. Gene selection is based on Fisher's ranking criterion as in [3], which will be presented in more detail in the next section. At each cross validation cycle, the highest rank gene that also achieves the best classification performance across the nested cross validation procedure is selected as a significant one and it is added to the list of the previously selected genes, while the performance of the list is being recorded. The procedure continues iteratively until the pre-specified number of genes is selected. Finally, the list of genes achieving the highest classification accuracy is reported as the final set of markers.

1 Repeat n times
- Training set $\leftarrow n-1$ subsets;
- Testing set $\leftarrow$ remaining subset;

1.1 Repeat for $i = 1 \ldots k$ ($number\ of\ posible\ C\ values$):
   a. Repeat $n-1$ times (for samples only in the training set)
      - Training_validation set $\leftarrow n-2$ subsets;
      - Testing_validation set $\leftarrow$ remaining subset;
      - Train an SVM classifier using parameter C_i;
      - Test it on the Testing-validation set.
   b. Record $P(i)$, the performance of the SVM classifier over $n-1$ Testing_validation sets.

1.2 Determine C_j where $j = \arg\max P(i)\ for\ i = 1 \ldots k$;

1.3 Train the SVM classifier on the training set using parameter C_j.

1.4 Test the classifier obtained in step 1.3 on the testing set.

2. Return p, the best performance of the classifier over n testing sets.

Table 1: Nested cross validation process of the Gene Expression Model Selection (GEMS) procedure.

2.3.2 The RFE-SVM Method

The RFE-SVM method [1] is based on SVMs [11] and the idea of ranking features according to the absolute value of the components of the direction vector $\boldsymbol{w}$. As expressed in equation (2.4), each individual component of $\boldsymbol{w}$ is associated with an individual component of vector $\boldsymbol{x}$, which is the expression level of an individual feature. Thus, every feature (gene) is multiplied by a weight; the larger the absolute

value of its weight, the more important that feature is according to RFE-SVM, in the sense that it contributes more to the decision function of equation (2.1). As a consequence, genes can be ranked according to the absolute value of the individual components of $\boldsymbol{w}$. A general overview of the method is given in Table 2.

1. Let m be the initial number of features.
2. While $(m \geq 0)$
3. Estimate the direction vector $\boldsymbol{w}$ of the separating hyperplane using linear SVM.
4. Rank features according to the components of $|\boldsymbol{w}|$.
5. Remove the feature with the smallest weight in absolute value $(m \leftarrow m-1)$. More than one features can be removed in each iteration.
6. Estimate classification accuracy of the m surviving features using a linear SVM classifier.
7. End While
8. Output as marker genes the set of surviving features achieving maximum accuracy performance.

Table 2: The Recursive Feature Elimination based on SVM (RFE-SVM) algorithm.

2.3.3 Differentially Expressed Genes

The basic idea behind the development of the proposed methodologies is the identification and eventually the selection of differentially expressed genes. This idea is not new in marker selection; it has been stated in various studies a number of which were cited in section 2.2. In all these studies domain experts are using variations of Fisher's coefficient which is given by the following equation:

$$f_1(g_i) = \frac{\left(\mu_+(g_i) - \mu_-(g_i)\right)^2}{\sigma_+(g_i)^2 + \sigma_-(g_i)^2} \tag{2.8}$$

A variation of Fisher's coefficient could be expressed as:

$$f_2(g_i) = \frac{\sum_{j=1}^{n} \left| g_{ij} - c(g_i) \right|}{\sigma_+(g_i) + \sigma_-(g_i)} \tag{2.9}$$

where, $\mu_+(g_i)$, $\mu_-(g_i)$, $\sigma_+(g_i)$ and $\sigma_-(g_i)$ are the means and standard deviations of the expressions of gene g_i in positive and negative class respectively, n is the number of samples and

$$c(g_i) = \frac{(\mu_+(g_i) + \mu_-(g_i))}{2} \tag{2.10}$$

Another variation of low computational cost is given as:

$$f_3(g_i) = \frac{|\mu_+(g_i) - \mu_-(g_i)|}{\sigma_+(g_i) + \sigma_-(g_i)} \tag{2.11}$$

Elaborating more, one could easily verify that equation (2.8) as well as equations (2.9) and (2.11) essentially express the same concept. When using these equations to assign weights to a set of given genes, it is obvious that genes which differentiate more their expression in the two situations (say -3 in the pathological cases and +3 in the normal) are assigned higher weights than those which differentiate less between the two classes. Genes that express themselves in exactly the same way between the two situations (they take the same expression in both pathological and normal states) are assigned the minimum weight of zero. We propose to use such a metric in a wrapper fashion, embedded properly within the learning procedure of linear neurons and SVMs in order to assist the task of marker selection.

2.3.4 The RFE-LNW Approach

Most marker selection approaches applied to the field of DNA microarray, due to the high dimensionality of the data, use linear tools to assess the problem. RFE-SVM is such a method where a linear kernel is used to estimate the weight vector of the separating hyperplane, the absolute value of which is then used as the ranking criterion of genes. On the other hand, due to its design (a linear combination of

inputs) a Linear Neuron (LN) can also approximate any linear function. Thus, we propose to use such a Linear Neuron to approximate the separating hyper-plane between positive and negative classes. Taking advantage of such an open architecture, we could choose among a variety of learning schemes, or easily embed a new learning procedure properly adapted to the underlined problem while we could further expand to multilayer and multi-class formulations. This is applied as a single

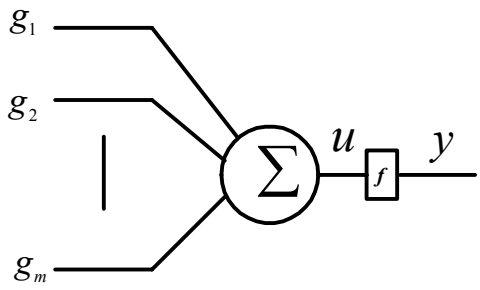

Figure 15: A single neuron adapted to the marker selection problem.

neuron network of *m* inputs (Figure 15), where *m* corresponds to the number of genes. Considering two possible outcomes at the output layer, namely *output 0 for the negative class and 1 for the positive class,* we can use such a Linear Neuron (LN) to approximate the separating hyper-plane that distinguishes the two classes of interest (negative, positive). More specifically, using the sigmoid function $f(u)$ we obtain:

$$y = \frac{1}{1+e^{u}} = f(u) \tag{2.12}$$

$$u = \sum_{i=1}^{m} w_i g_i \tag{2.13}$$

$$f'(u) = y(1-y) \tag{2.14}$$

Note that $f'(u) \geq 0$ since y ranges 0 to 1.

2.3.5 Training the RFE-LNW

In this section we provide the basic mathematical background of the training procedure used to iterate the weights of the Linear Neuron. The error function of a single neuron that is to be minimized is given by:

$$E = \frac{1}{2}\sum_{j=1}^{n}\left(d_j - y_j\right)^2 \tag{2.15}$$

where n corresponds to the number of samples, d_j represents the desirable neuron output associated with sample j and y_j is the actual output produced by this neuron for the given sample. Through the gradient descent method for the minimization of equation (2.15), we update the weight w_i associated to gene g_i as follows:

$$w_i(t+1) = w_i(t) - \left(\mu\frac{\partial \mathrm{E}}{\partial w_i}\right) = w_i(t) - \mu\sum_{j=1}^{n}\left(\frac{\partial \mathrm{E}}{\partial y_j}\frac{\partial y_j}{\partial u_j}\frac{\partial u_j}{\partial w_i}\right)$$

We propose to influence the update through the Fisher's metric in equation (2.9) as follows:

$$w_i(t+1) = w_i(t) - \frac{\mu}{2}\sum_{j=1}^{n}\left[\left(\frac{\partial \mathrm{E}}{\partial y_j}\frac{\partial y_j}{\partial u_j}\frac{\partial u_j}{\partial w_i}\right)\frac{\left|g_{ij} - c(g_i)\right|}{\sigma_+(g_i) + \sigma_-(g_i)}\right] =$$

$$w_i(t) - \frac{\mu}{2}\sum_{j=1}^{n}\left(-2\left(d_j - y_j\right)y_j\left(1-y_j\right)g_{ij}\right)f_2(g_i) =$$

$$w_i(t) + \mu\sum_{j=1}^{n}\left(d_j - y_j\right)y_j\left(1-y_j\right)g_{ij}f_2(g_i) =$$

$$w_i(t) + \mu\sum_{j=1}^{n}\left(d_j - y_j\right)f'\left(u_j\right)g_{ij}f_2(g_i).$$

Finally

$$w_i(t+1) = w_i(t) + \mu \sum_{j=1}^{n} e_j f'(u_j) g_{ij} \frac{\left| g_{ij} - c(g_i) \right|}{\sigma_+(g_i) + \sigma_-(g_i)} \tag{2.16}$$

where t represents current iteration, μ is the learning rate and

$$e_j = \left(d_j - y_j \right). \tag{2.17}$$

working with signs, which is an idea introduced in resilient back propagation learning [13], we express (2.16) as follows:

- In case that $f_2(g_i) = 1$, which is similar to the standard back-propagation procedure we get:

$$w_i(t+1) = w_i(t) + \mu \sum_{j=1}^{n} e_j f'(u_j) g_{ij} \tag{2.18}$$

$$w_i(t+1) = w_i(t) + \mu \sum_{j=1}^{n} sign\left(e_j f'(u_j) \right) sign\left(g_{ij} \right) \tag{2.19}$$

- or in general:

$$w_i(t+1) = w_i(t) + \mu \sum_{j=1}^{n} sign\left(e_j f'(u_j) \right) sign\left(g_{ij} \right) f_2(g_i) \tag{2.20}$$

$$w_i(t+1) = w_i(t) + \sum_{j=1}^{n} \left| d_j - y_j \right| sign\left(e_j f'(u_j) \right) sign\left(g_{ij} \right) f_2(g_i) \tag{2.21}$$

Equation (2.18) is the basic gradient descent learning algorithm proven to drive the error function (2.15) to a minimum. Equation (2.19) also converges since by keeping the sign of the gradient we are heading towards the direction of the minimum, which eventually will be reached (except in cases when trapped to a local minimum) by

using the appropriate learning rate. In fact, we expect (2.19) to converge faster than (2.18) since e_j and $f'\left(u_j\right)$ in (2.18) can take very small values resulting in low modifications of w, implying low modification of the error function and slowing down convergence [13]. Taking only the sign of the gradient in (2.19) we are heading towards the direction of the minimum taking 'larger stable steps' and speeding up convergence at least when the process is far from a minimum. This derivation is very helpful in its application of the marker selection process, since it forces the algorithm to converge much faster, especially at the first steps of the process where the number of attributes (genes) is extremely large. Equation (2.20) differs from (2.19) only in the coefficient $f_2\left(.\right)$. Following the same reasoning, (2.20) can be proved to converge to a minimum, but in addition it takes into consideration and eventually measures through the summation term an approximation of the Fisher's metric. Low dimensionality, however, may slow down convergence. As the process elevates and the problem dimensionality is significantly reduced, the samples over attributes ratio increases and the problem of estimating the separating hyperplane becomes harder, slowing down convergence. This necessitates the increase of either the number of epochs or the learning rate. At these last steps, equation (2.21) could be used to speed up convergence by using a variant learning rate $\left|d_j - y_j\right|$. It is straight forward to show that while we are far from the target $\left|d_j - y_j\right|$ will take a 'large' value speeding up convergence, but, while we are approaching the goal $\left|d_j - y_j\right|$ will start taking lower values, thus, slowing down convergence. In other words, at the last steps of the feature selection process we are heading towards the goal fast when we are far away from it, but we slow down when we are approaching the target for better fine tuning of the separating hyperplane. In our proposed iteration scheme, equation (2.20) in

combination with equation (2.21) are the final weight update rules used to asses the weights, while equations (2.18) and (2.19) are mostly used for explanation and justification purposes.

As a concluding remark of this section, we point out that by training a *single neuron* with an appropriate learning procedure, we can eventually apply a filter criterion such as Fisher's ratio in a wrapper fashion.

2.3.6 Emphasizing Differentially Expressed Genes

The selection of differentially expressed genes is a desirable goal of any marker selection approach [14]. Thus, we need to prove that among the extremely large number of genes our method is fed with, it will finally select markers that are expressed in a different way between the (two) situations of interest. Figure 16 shows the expression level of a hypothetical gene g_i in negative $(C=0)$ and positive $(C=1)$ class respectively. In cases (a) and (b) the hypothetical gene is differentially expressed in the two classes of interest, green (negative values) in negative class and red (positive values) in positive class or visa versa. On the other hand, cases (c) and (d) show no differentiation in the expression level of the specific gene in the two situations of interest. Considering case (a) in combination with equation (2.19) and focusing on the negative class (green part), we notice that the term $sign\left(e_j \cdot f'\left(u_j\right)\right) \cdot sign\left(g_{ij}\right) \geq 0$ holds; indeed $e_j \leq 0$ (since $d_j = 0$ (2.17) and $y_j \in [0 \cdots 1]$), $f'\left(u_j\right)$ from equation (2.14) is positive and $g_{ij} = -3$. Now focussing on the positive class (red part) of Figure 16 (a) and using the same reasoning we notice again that $sign\left(e_j \cdot f'\left(u_j\right)\right) \cdot sign\left(g_{ij}\right) \geq 0$. Since the term e_j in a realistic scenario is most often non zero, the summation term of equation (2.19) produces a positive result. Following about the same reasoning in case (b) of Figure 16 one may show that

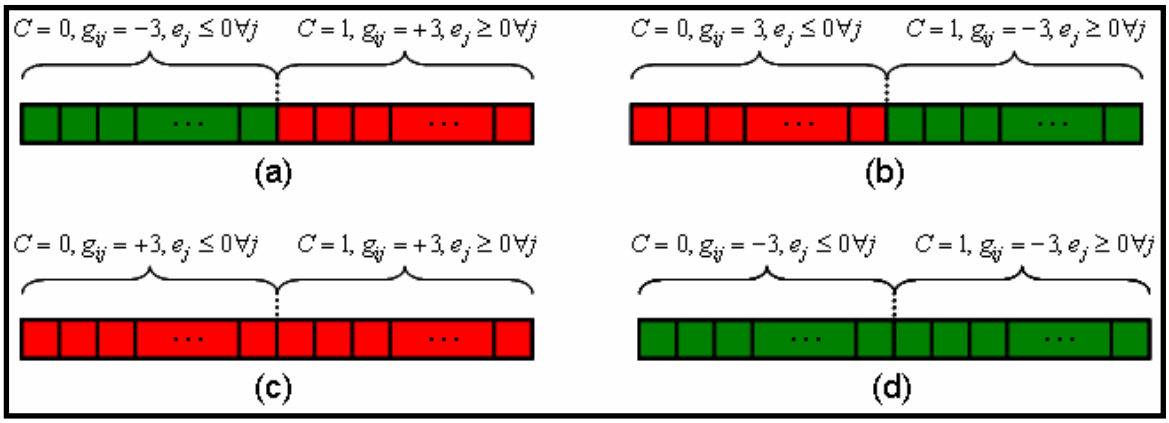

Figure 16: Differentially expressed genes versu non differentially expressed

equation (2.19) produces a negative result, while the summation term for this equation in cases (c) and (d) produces results close to zero since the term values in the two classes negate each other. If we now consider the absolute values of the assigned weights, we see that differentially expressed genes (cases (a) and (b)) take larger values than genes that do not differentiate their expression (cases (c) and (d)) in the two situations of interest. Notice on the contrary that equation (2.18) can not produce the same effect, since it depends on the value and not on the sign of term e_j , which eventually could be very low and diminish the expected result.

On the other hand, Equation (2.19) is unfair to differentially expressed genes since by taking only the sign value, genes that are more differentially expressed will be assigned the same weight with genes that are less differentially expressed. A more fair solution would be to assign higher weights to more differentially expressed genes, which is achieved by equation (2.20) where the magnitude term $f_2(g_i)$ is also introduced.

2.3.7 Incremental Versus Batch Learning

Equations (2.18)-(2.21) update the weight term $w(t+1)$ after all the examples are presented to the network, that is after the summation terms have been evaluated. This is referred to as batch training in neural network theory. Alternatively, one may

update the weights incrementally considering one sample at a time, in which case the summation terms can be dropped. To emphasize this update strategy we present the above set of equations again as follows:

$$w_i(t+1) = w(t) + \mu e f'(u) g_i \tag{2.22}$$

$$w_i(t+1) = w(t) + \mu \cdot sign\left(e \cdot f'(u)\right) \cdot sign(g_i) \tag{2.23}$$

$$w_i(t+1) = w(t) + \mu \cdot sign\left(e \cdot f'(u)\right) \cdot sign(g_i) \cdot f_2(g_i) \tag{2.24}$$

$$w_i(t+1) = w(t) + |d-y| \cdot sign\left(e \cdot f'(u)\right) \cdot sign(g_i) \cdot f_2(g_i) \tag{2.25}$$

For the experiments conducted in this work, weights are updated in an incremental fashion, since for the specific domains it was proved to produce better results than the batch mode weight-update method. However, we think that this is a domain specific decision and we can not draw a safe general conclusion in favor of one or the other weight update mode. Note that for the experiments conducted in this work, equation (2.25) was applied from the point of 100 surviving genes up to the end for the process.

2.3.8 Algorithmic Presentation of RFE-LNW

Table 2 provides an algorithmic overview of RFE-LNW method as described in the previous sections. Notice that besides the advantage of using a single neuron as our only learning component, we have also managed to apply filter criteria in a true wrapper fashion, where weights are re-evaluated and potentially adapted from iteration to iteration. Indeed, by eliminating genes we are actually reducing the dimensionality of the problem and thus, the new estimated hyperplane is re-evaluated in a new reduced space with a new direction vector $\boldsymbol{w}$. In real-world applications it

seems appropriate that gene weights are changing from iteration, to iteration since a large feature space with many insignificant genes can obscure the influence of truly

1. Let m be the initial number of features
2. While $(m \geq 0)$
3. Update the weight vector $\boldsymbol{w}$ using equations (2.24) and (2.25).
   (For the experiments conducted in this study, equation (2.24) was used as long as the number of surviving features was greater than 100, whereas equation (2.25) was used otherwise).
4. Rank the genes according to the absolute values of vector $\boldsymbol{w}$.
5. Remove the feature with the smallest weight in absolute value, $(m \leftarrow m-1)$. More than one features can be removed in each iteration.
6. Estimate the classification accuracy of the m surviving features using a linear SVM classifier.
7. End While
8. Output as marker genes the set of surviving features achieving the best classification accuracy.

Table 3: The RFE-LNW method, based on the weight assignment of a properly trained linear neuron.

important ones, which become more relevant as the dimensionality of the problem reduces. Notice the difference with the filter method where the fisher metric of surviving genes remains stable along the entire phase of the feature selection process and so does the separating boundary it defines.

2.3.9 RFE-SVM and RFE-LNW

In this section we use a simple example to compare the proposed methodology (RFE-LNW) with the well established RFE-SVM. It is well known that SVMs find the direction vector $\boldsymbol{w}$ and the shifting parameter b of a line that maximizes the distance between positive and negative classes of the given samples. Alternatively, through equations (2.20) or (2.24) RFE-LNW searches for approaching the direction vector $\boldsymbol{w}$ of the line that maximizes the 'difference' on the expression level of the selected

markers. The result of these two quiet different philosophies is demonstrated through the following simple example. Consider that we are given 4 samples, 2 negative and 2 positive, which are depicted in Figure 17; negative samples are represented by asterisks while positive are represented by circles. These samples are described by 2 genes each, defining a two dimensional space. Let g1 be the single descriptive marker on this toy domain, *where a negative values of g1 designate the negative class while positive values designate the positive class*.

Letting the two systems operate on the examples given, we notice a quiet different behavior on their learning philosophies depicted in Figure 17. Since RFE-SVM tries to maximize the distance between the two given classes, it derives the dashed line. RFE-LNW on the other hand, using a learning rate of 0.1 and 200 epochs in combination with equation (2.24) discovered the underlined rule, expressed by the solid line in Figure 4. Through this simple example we aim at demonstrating that maximization of the distance between the two classes of interest may not always be the best choice for marker selection and that additional prior constraints reflecting the intrinsic domain properties might be necessary in order to derive not only stochastically but also biologically sound results. This is exactly a targeted advantage of our proposed methodological scheme which is demonstrated through the neuron but can also be extended to the SVM formulation as in the next section. An add-on advantage of the proposed approach is that it offers a general solution for using filter criteria in a wrapper fashion, allowing other filter criteria besides Fisher's metric to be used as well. A drawback of the LNW compared to SVM formulation is that it suffers from the local minimum problem and requires the fine tuning of a number of parameters, such as the learning rate, number of epochs and the point in the feature selection process that switches to a variant learning rate through equation (2.25).

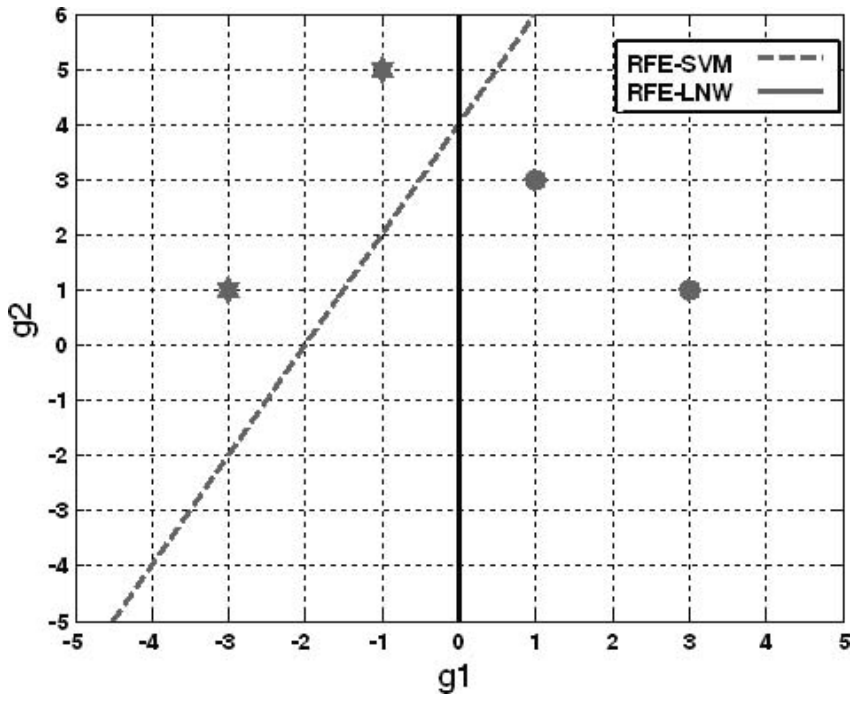

Figure 17: Illustration of learning differencies between the RFE-LNW and RFE-SVM methodologies.

An alternative approach which makes use of SVMs and Fisher's ratio is presented in the next section, referred to as RFE-FSVs method. This approach is based on SVMs and hence the local minimum problem is resolved. Besides, it has a smaller number of parameters to refine as is presented in subsequent sections.

2.4 The RFE-FSVs Approach

According to SVM theory, the direction vector $\boldsymbol{w}$ of the separating hyperplane is given by equation (2.4), as an expansion of the samples whose λ_j's are non zero, i.e. the support vectors. Based on this equation the individual components of the direction vector $\boldsymbol{w}$ could be found by:

$$w_i = \sum_{j=1}^{n} \lambda_j y_j x_{ij} \tag{2.26}$$

Let *SVs* be the set of support vectors and *S* be the set of indices defined as:

$$S = \{k : \boldsymbol{x}_k \in SVs\} \tag{2.27}$$

Focusing only on support vectors (2.26) can be written as:

$$w_i = \sum_{j \in S} \lambda_j y_j x_{ij} \tag{2.28}$$

We propose to introduce the metric (2.11) into the above formulation as:

$$w_i' = \sum_{j \in S} sign(\lambda_j) \cdot y_j \cdot sign(x_{ij}) \cdot \frac{\left| \mu_{s+}(g_i) - \mu_{s-}(g_i) \right|}{\sigma_{s+}(g_i) + \sigma_{s-}(g_i)} \tag{2.29}$$

where $\mu_{s+}, \mu_{s-}, \sigma_{s+}, \sigma_{s-}$, are the means and standard deviations of the support vectors for positive and negative classes. Note that in equation (2.29) the component w_i' is computed based only on the support vectors, since λ_j is zero for non support vectors and so does the Fisher metric utilized. Hence, the direction vector $\boldsymbol{w}'$ defined in equation (2.29) expresses a Fisher's line that passes through the origin and retains the same direction sign with the line defined by the conventional SVMs approach. This new line can be used for defining the ranking criterion of surviving genes. Also note that by using different kinds of kernels we are supplied with different sets of support vectors and thus, different Fisher lines. Overall in the proposed RFE-FSVs methodology we propose to use the absolute values of the components of the direction vector $\boldsymbol{w}'$ defined in (2.29) as a new ranking criterion, while preserving the general structure and the iteration scheme of the conventional RFE-SVM approach.

Besides the mathematical formulation of the proposed methodology, we will also attempt to reveal an intuitive reasoning hidden behind the RFE-FSVs method. As it was explained in section 2.3.2, support vectors lie on the margin of the separating hyperplane (Figure 14). A significant property of SVMs is that the learning process never changes as long as all support vectors remain the same. More specifically, the system learns the exact same separating hyperplane by keeping only the support vector samples (which might be only a small proportion of the entire data set) and

ignoring all remaining samples. The converse however is not true; by ignoring some or all of the support vectors, the system is forced to learn a different boundary of separation. This means that by keeping only a *small proportion* of the entire training set, namely the support vectors, the system learns the same rule as if the entire set was presented to it. Thus, SVMs are mostly based on quality rather than on quantity of the supplied samples, in the sense that a few *but representative* cases could be enough to derive the underling rule. Not representative samples, on the other hand, used as support vectors could lead to a peculiar solution such as the one presented in section 2.3.9. In order to minimize such a side effect, kernels could play the role for locating appropriate and representative samples, as well as for better refining the separating hyperplane, since as the process progresses and dimensionality is reduced non-linear kernels could become better approximators. Furthermore, support vectors lie around the margin of separation between the two classes, which is the critical region to distinguish between the two situations of interest (Figure 14). Based on this property, one issue of particular algorithmic but also biological interest concerns the gene topology that forces a specific patient to cross the border of separation from one class to the other. The idea then behind the proposed methodology is to focus on the class borders and examine the factors that cause this misallocation. The borders are determined by the support vectors and by selecting different kinds of kernels we can obtain a variety of support vectors, which can also be viewed as different sets of domain representatives. Kernels then play the dual role of search engines that help us locate a representative subset of crucial samples, as well as for better fine tuning of the separating hyperplane, as dimensionality is decreased and linearity becomes questionable. The use of kernels could lead to an increase of the generalization performance of the selected markers by focusing on characteristic properties of the

problem domain. Notice that both proposed methodologies take into account intrinsic domain characteristics expressed through variations of the Fisher's ratio. Also note that RFE-FSVs by using the support vectors focuses at the class borders, while RFE-LNW on the other side aims on the entire class topology. In any case however, no safe conclusion can be drawn in favor of one or the other method as it seems to be a trade off situation and a domain specific dilemma.

By enriching the proposed methodology (RFE-FSVs) with intrinsic characteristics towards the appropriate application of Fisher's criterion in a wrapper fashion, genes are ranked according to the absolute values of the weight vector given by equation (2.29). The gene weights are changing and adjusted along the process, since the set of support vectors is not fixed in the entire evaluation phase. One argument that could be posed as a criticism for this approach is that since we work on a very large dimensional space there is an over fitting problem whenever a large proportion of the samples become support vectors. However, over fitting is not in fact a problem since we use the support vectors only for estimating the ranking criterion and classifying the training set. For classification purposes of new unseen samples a *linear kernel* is used as in the conventional RFE-SVM approach. Another criticism could be raised that the proposed method may degrade to the original filter one. Nevertheless, this is only true in the very seldom case where all samples are support vectors across the entire feature elimination process. A specific study addressing the above issues is conducted later in section 2.6.

2.4.1 Algorithmic Presentation of RFE-FSVs

We are now introducing the proposed feature selection method referred to as RFE based on Fisher's metric and Support Vectors (RFE-FSVs) which is described in the Table 4

1. Let n be the initial number of features
2. While $(m \geq 0)$
3. Create and train the SVM classifier using any type of kernel.
4. Locate the Support Vectors (SVs).
5. Based on the Support Vectors only rank the genes according to the value returned by equation (2.29).
6. Remove the feature with the smallest weight in absolute value, $(m \leftarrow m-1)$. More than one features can be removed in each iteration.
7. Estimate the classification accuracy of the m surviving features using a linear SVM classifier.
8. End While
9. Output as marker genes the set of surviving features achieving the best classification accuracy.

Table 4: The Recursive Feature Elimination based on the Fisher's metric and Support Vectors (RFE-FSVs) algorithm**.**

Noteworthy aspects of the above implementation are the following:

1. It provides an alternative way to embody a filtering criterion within a wrapper methodology, taking into consideration intrinsic characteristics of the data.

2. By focusing only on the support vectors for ranking features, we can actually use any type of kernel (besides the linear one) to approach the problem of marker selection.

3. Any type of correlation coefficient could be used as a criterion of feature ranking.

4. By using various types of kernels, the proposed methodology can be directly applied to non-linearly separable problems.

The performance of the proposed methodologies and the validation of their properties are elaborated in the next sections through direct application on two data sets.

2.5 Applied Data Sets

The two data sets tested in this study are from the Leukemia and Breast Cancer (BC) application domains published in [3] and [4], respectively. Both data sets consist of a training set and an independent test set. The Leukemia domain contains 7129 genes, where the training set consists of 38 samples (27 ALL and 11 AML) and the test set of 34 samples (20 ALL and 14 AML), all normalized to a zero mean and standard deviation one, as suggested in the original publication [3]. The BC data set contains 24481 genes and 78 samples on the training set, 44 of which is characterized negative and correspond to patients that remain disease free for a period of at least five years, whereas the remaining 34 are characterized positive and correspond to patients that developed a relapse within a period of five years. 293 genes expressing missing information for all 78 patients were removed and the remaining 13604 missing values were substituted using Expectation Maximization (EM) imputation [15]. The independent test set consists of 19 samples, 7 negative and 12 positive.

In a first attempt to compare the two domains, we performed an unlabelled cluster evaluation procedure. Using all genes of the leukaemia domain, this process resulted in two main clusters. By assigning then the known labels to the clustered samples, the two classes (ALL and AML) were almost completely discovered with only one sample being misallocated. A hierarchical clustering in combination with a Pearson correlation and average linkage was used. On the contrary, the BC domain can not be appropriately clustered, leading to the conclusion that there is a lot of overlap between the two classes.

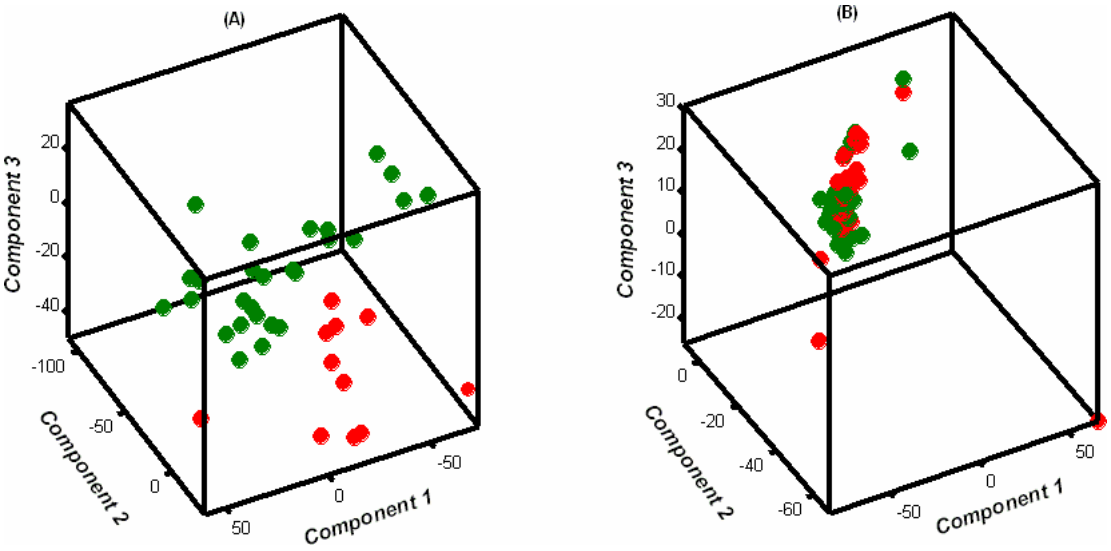

Figure 18: PCA analysis on two domains, where we observe that the Leukemia domain (on the left) presents almost no overlap which, is not the case in the BC domain (on the right).

This result was also verified by a PCA evaluation performed on the two domains, the results of which are visualized in Figure 18 for the three principal components of each data set. We observe that in the Leukemia domain, (Figure 18 (A)) the two classes are well separated between each other, while in the BC domain (Figure 18 (B)) there is a lot of overlap between the two classes and in fact they are very tightly bound within each other. Taking also into account the overall variance of the two domains, we notice that in the case of BC the variance is much smaller (0.06) compared to the 0.97 value of the leukemia domain. From the above analysis on the two domains, we can conclude that the BC data set is much less separable than the Leukemia one, making gene selection a much more difficult and intriguing problem. It is worth stressing the fact that no preprocessing step is conducted on any of the two domains in the subsequent experiments. In the original study on BC domain [4], two preprocessing steps were taken that lead to 231 final set of genes, on which the actual gene selection process was applied.

2.5.1 Experimental Scenarios - Results

Three series of experimental scenarios are conducted on the tested domains and the accuracy evaluation criteria (reported in Appendix I) are appropriately assessed in the conducted experiments.

ILOO Accuracy Performance

In this first experimental scenario, the Internal Leave One Out Success Rate (ILOO-SR) is assessed for both tested data sets. Even though it is not an unbiased estimator and it introduces a significant amount of bias [16], ILOO is used here only as a mean of demonstrating the learning ability of the tested methodologies on the training set and not as an actual measure of success rate estimator. For a more unbiased and more realistic results, we use independent test set and 10-fold cross validation strategies, which correct for any introduced bias [16].

Independent Test Set Evaluation and 10-Fold Cross Validation

In the second series of experiments, independent test set evaluation is used to derive the final set of marker genes by feeding the training samples to the gene selection procedure and proceed as follows: For each iteration of the feature selection process a pre-specified number of genes is eliminated according to the scenario presented in column 2 of Table 1 (leukaemia data set) and Table 3 (BC data set) in Appendix I. A classifier is build from the training set based on the survived features while it is tested on the independent test set provided for the two data sets. The minimum set of genes achieving the highest classification accuracy is then reported as the final set of markers.

In the third series of experiments, a 10-fold cross validation scheme is assessed 10 times as follows: The training samples are randomly divided $10(folds)\times 10$ times into training groups consisting of 90% of the training set and test groups consisting of the

remaining 10%; such a validation scheme has been proposed by Kohavi in [17]. For each of the one hundred runs, a gene selection procedure similar to that used for the independent test set evaluation (previous paragraph) is applied, yielding both a set of marker genes and a classification accuracy measured on the corresponding test group by the selected markers. The overall average of the number of markers selected, as well as the accuracy performance achieved in each run by those markers on the corresponding test group is reported in order to assess the final cross validation performance of the methodology under consideration. Markers at the end of each one of the one hundred runs are selected according to the following rule: The set of marker genes of size less than or equal to 100 providing the highest classification accuracy on the corresponding test group but also classifying perfectly the training set are identified as the selected set of marker genes.

In all previously conducted experiments, unless otherwise stated, RFE-LNW was applied with 0.01 learning rate and 3000 epochs up to 100 genes. Beyond 100 genes, a variant learning rate through equation (2.25) was applied with 200 epochs up to the end of the process. A linear SVM classifier was used for accuracy assessment in all experiments. TIGR-MEV version 3.1 was used as the expression profile viewer. Note that for RFE-FSVs method, even though a kernel was used as a tool to search for a representative set of support vectors and for classifying the training set, the accuracy of the method on the independent test set was estimated using a linear SVM classifier.

GEMS Experiment

We also conducted a series of experiments to compare the performance of the proposed methodologies to that of the GEMS approach. GEMS software [18] has been designed in such a way that it can not be applied directly in the previously described experimental procedures. It uses a 10-fold cross validation strategy similar

to that presented before, but it provides as a final set of marker genes those that succeed the best classification performance across the 10 runs of each fold. Such a 10-fold cross validation strategy was applied 10 times, and hence 10 set of marker genes were given as possible candidates. Outside GEMS software then, we tested the prediction ability of those marker sets to the corresponding independent test sets (which have not been used in the 10-fold evaluation process) and the marker set achieving the best performance was the final derived marker set for this experimental scenario. The results of the proposed methodologies are tested in a similar manner, so that a direct comparison to GEMS can be performed. In this experimental set an additional test (Q-Statistic) [19] measuring the statistical significance of the derived results is applied, revealing some interesting characteristics of the underlined methodologies. Note that, for this last experimental step, a linear SVM classifier was used according to the parameter values returned by GEMS that were estimated appropriately during the 10-fold cross validation procedure.

2.5.2 Experimental Results on Leukemia

ILOO Accuracy Performance

The ILOO (Internal Leave One Out) accuracy performances of RFE-LNW and RFE-FSVs are summarized in Table 1 of Appendix I. We point out that we took advantage of the ability of the RFE-FSVs method to employ various types of kernels and we present the results for a 4^{th} degree polynomial kernel (RFE-FSVs-4DK) and a combination of an RBF and a 7^{th} degree polynomial kernels (RFE-FSVs-RBF7DK). In the case of kernel combination, we use an RBF kernel with $\gamma = 0.01$ as long as the number of surviving genes is larger than 100, whereas a 7^{th} degree polynomial kernel was used otherwise. With this learning scheme, we emphasize the fact that while dimensionality is still large, an RBF kernel which defines different class regions is

used as a means of locating support vectors and estimating feature weights, whereas as the process iterates and dimensionality decreases (making the class distinction problem even harder) a 7^{th} degree polynomial kernel is used in place of the RBF.

We observe from Table 1 of Appendix I that all methods perform well on the ILOO evaluation criterion with an average ILOO-SR of over 99%, implying that they can learn and generalize fairly well on the training set.

Independent Test Set Evaluation and 10-Fold Cross Validation

The performance of the two methods is also evaluated on an independent test set, the detailed results of which are presented in Table 2 of Appendix I. An overview of the result on the independent test set is also presented in the right part of Table 5 below. We point out that the 97.06% accuracy achieved by the RFE-FSVs-RBF7DK, which corresponds to one missed sample, is very close to the best results ever reported on this domain, i.e. 100% accuracy with three selected genes respectively in [20]. Another outstanding result was reported in 0 with the RFE-SVM method, where 8 genes achieved 100% accuracy on the independent test set. Unfortunately, we are not given the gene names that achieved this remarkable performance on these two studies and we could not reproduce these results. Concerning the RFE-LNW and the RFE-FSVs-4DK, an Independent Test Set Success Rate (ITS-SR) of 94.12% was achieved, which corresponds to two missed examples with only two genes.

As a concluding remark on this set of experiments we point out that the proposed methodologies produce comparable results, very close to the highest accuracy reported in the international bibliography, establishing that application of filtering criteria in a wrapper manner can be effectively applied. Their main advantage however, will become evident in the harder domain of breast cancer examined next, demonstrating a consistently high performance on a variety of data sets.

	Training Set 10 Fold Cross Validation		Independent Test Set			
Method	**Success Rate**	**Genes**	**Success Rate**	**Genes**	**Sensitivity**	**Specificity**
RFE-LNW	99.00%	3	94.12%	2	0.93	0.95
RFE-FSVs-4DK	99.00%	2	94.12%	2	1.00	0.93
RFE-FSVs-RBF7DK	99.75%	2	97.06%	3	0.93	1.00

Table 5: Leukemia Domain: performance of the tested methodologies on the 10-fold cross validation scheme as well as on the independent test set.

The selected markers for the proposed methodologies are reported in Table 6 where we notice that X95735 (Zyxin) is common to all marker selection schemes. We point out the fact that all reported genes except the M19507 were also included in the set of 50 markers published by Golub et al. in [3].

The performance of the 10-Fold cross validation scheme is presented on the left part of Table 5, where we observe that the RFE-FSVs-RBF7DK is slightly the best performer with a 99.75% average accuracy over 100 runs. RFE-LNW and RFE-FSVs-4DK achieved also very good results of 99% each. A very interesting fact that was revealed in this series of experiments is that the most frequently selected gene by all tested methodologies is X95735 which is responsible for the protein of zixyn. This has also been reported in various studies [21], [22], [23] to be a highly informative gene. X95735 was always selected by RFE-LNW methodology (100 times, perfect frequency rate), 75 times by the RFE-FSVs-4DK and 81 times by the RFE-FSVs-RBF7DK, implying that the proposed methodologies comply with already established knowledge on the specific domain.

RFE-FSVs-4DK		
Accession	**Symbol**	**Description**
X95735	*ZYX*	*zyxin*
U22376	MYB	v-myb myeloblastosis viral oncogene homolog (avian)
RFE-FSVs-RBF7DK		
M19507	MPO	myeloperoxidase
M23197	CD33	CD33 molecule
X95735	*ZYX*	*zyxin*
RFE-LNW		
M27891	CST3	cystatin C (amyloid angiopathy and cerebral hemorrhage)
X95735	*ZYX*	*zyxin*

Table 6: Leukemia Domain: Markers selected by tested methodologies; zyxin is common to all marker sets.

GEMS Experiment

GEMS was applied 10 times using a 10-fold cross validation strategy as described in section 2.5.1 (10x10 runs). The 10 marker sets derived through the 10-fold cross validation process are tested on the independent test and the one achieving the highest performance is recorded as the final result. We apply a similar methodology to the proposed gene selection procedures, but on the same average number of genes as those selected by GEMS. Ten marker sets are derived as possible outcomes as well and the one achieving the highest performance on the independent test set is recorded as the final score, Figure 19 provides a general overview of the derived results. GEMS selected 4 genes on the average with 61.76% maximum best performance. We measure the performance of the tested methodologies in a similar manner, i.e. for each one of the 10 folds, we record the set of 4-genes among the 10 runs that achieved the best accuracy performance. Then those 10 possible 4-gene candidates were tested on

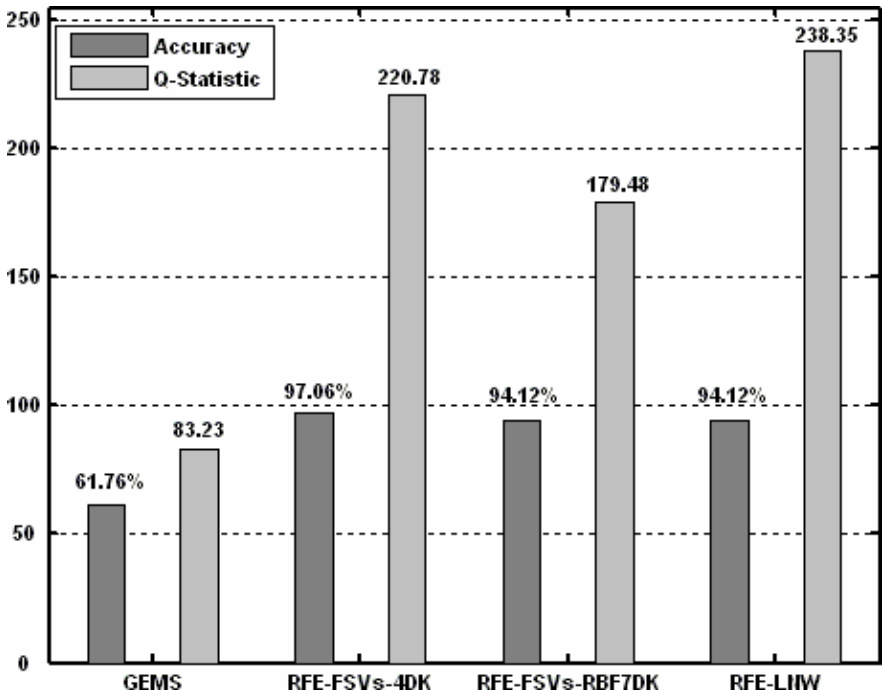

Figure 19: Performance evaluation of the tested methodologies. Results are recorded at the level of 4 genes for all methods, since this is the average number of genes selected by GEMS along the 10-fold cross validation process. Significant difference on both classification accuracy and statistical significance is observed in favor of the proposed methodologies.

the independent test set and the one achieving the highest classification accuracy was recorded. RFE-FSVs-4DK (Figure 6) is the best accuracy performer with 97.06% success rate (missing only one sample), while RFE-FSVs-RBF7DK and RFE-LNW succeeds also a remarkable performance of 94.12% (missing two samples).

A possible explanation on the significant advantage of the proposed methodologies could be given by conducting an additional experiment measuring the statistical significance of the derived classification result on the final set of marker genes. For each one of the previously four marker sets, the statistical significance of the classification result on the training set was measured using the Q-Statistic score introduced in [19]; the highest the value of the score the more significant the derived result is. Using such a score we can visualize in Figure 19 a significant statistical difference in favour of the proposed methodologies. GEMS achieved a score of 83.23 units while the minimum performance reached by the proposed methodologies is

179.48 achieved by RFE-FSVs-RBF7DK. Notice the fact that RFE-LNW is the best performer on this test with a score of 238.35 units.

2.5.3 Experimental Results on Breast Cancer

In this domain, besides testing the two proposed methodologies we are also evaluating the results of the well known method referred to as RFE-SVM introduced in section 2.3.2, for two reasons. First, to study the results of a representative wrapper method (such as RFE-SVM) and second, to study consistently the behavior of RFE-SVM on a 'harder' domain such as BC. Notice that RFE-SVM has been extensively studied mainly on the Leukemia domain in reference 0.

ILOO Accuracy Performance

The ILOO detailed accuracy results of the tested methodologies are presented in Table 3 of Appendix I, where RFE-FSVs is applied with a 7^{th} degree polynomial kernel (experimentally found to produce best results). Concerning the overall performance of the methodologies on the ILOO criterion, we observe that RFE-LNW (average performance 98.47%) is slightly better on the average than RFE-SVM (97.83%) and better than RFE-FSVs, which is still performing well with an average ILOO-SR of 86.42%. The RFE-SVM selection process was applied with various C values as reported in [24]; we present here the result for C = 1000; values of 10, 100, 1000, 10000 produced almost identical results, while C values of less than 1 did not produce any better results.

Independent Test Set Evaluation and 10-Fold Cross Validation

A detailed report concerning the entire process of the test set evaluation is reported in Table 4 of Appendix I, while an overview of the performance of the methods is reported on the in the right part of Table 7. We observe that RFE-FSVs-7DK is superior to other methods, achieving a success rate of 94.74% (only one sample missed) by selecting 73 genes. We point out that Van't Veer et al., [1] achieved a

success rate of 89.47% (two missed samples) by selecting 70 genes. The result produced by RFE-FSVs-7DK is also better than that reported in [20], where 8 genes give an accuracy of 89.47%, as well as the result reported in [26] where 44 genes gave an accuracy of 89.47%.

RFE-LNW produces noticeable results as well. Note that by selecting 44 genes it gives a classification accuracy of 89.47%; a result also comparable to that reported in [1], [20] and [26], where 70, 8 and 44 markers respectively give the same classification result. RFE-SVM was tested with various C values as suggested in [24]. More specifically, C values of 1, 10, 100, 1000 and 10000 as in ILOOCV produced approximately the same results and C values less than 1 did not produce any better

	Training Set 10 Fold Cross Validation		**Independent Test Set**			
Method	**Genes**	**Success Rate**	**Genes**	**Success Rate**	**Sensitivity**	**Specificity**
RFE-SVM	33	76.00%	32	78.95%	0.75	0.86
RFE-LNW	17	82.04%	44	89.47%	0.92	0.86
RFE-FSVs-7DK	21	84.71%	73	94.74%	0.92	1.00

Table 7: Breast Cancer Domain: Performance of the tested methodologies on the 10-Fold cross validation scheme as well as on the independent test set.

results. Concerning the performance of RFE-SVM in general, we observe that it achieves a success rate of 78.95% (4 samples missed) by selecting 32 genes, which is a respectable result as well. Concerning the accuracy achieved by these methodologies in the tested domains, one may notice that the performance of RFE-FSVs and/or RFE-LNW is quite remarkable and it is comparable or better to the best

published on BC domain, while both methodologies derive also high level results in the leukemia domain.

Performance of the 10-Fold cross validation scheme is presented at the left part of Table 7. RFE-FSVs-7DK and RFE-LNW are the best performers on the cross validation scheme with about 85% mean accuracy and 21 markers for the former and 82% mean accuracy and 17 markers for the latter ,over all 100 trials. This experiment reveals a substantial advantage of the proposed methodologies over the RFE-SVM, which achieved 76% mean accuracy and 33 markers. Thus, for 'hard' domains, where a lot of overlap exists between the two classes of interest, the application of filtering criteria in a wrapper fashion may be one of the keys for improving performance. Focusing on qualitative aspects of the derived results, we study the expression profiles of the markers selected by each method (i.e. the set of genes that gave the highest classification accuracy on the independent test set on the BC domain) by visualizing their behavior in Figure 20 (A, B, and C), where genes are ranked in increasing order according to Fisher's metric given by $\frac{\mu_+(g_i)-\mu_-(g_i)}{\sigma_+(g_i)+\sigma_-(g_i)}$. We can observe that the two proposed methodologies show a clear advantage over the RFE-SVM approach by clearly discriminating the classes of interest.

More specifically, the RFE-FSVs (Figure 20-A) and RFE-LNW (Figure 20-B) methods select genes that share an intrinsic characteristic related to a significant difference in the expression of the selected markers from one (negative) to the other (positive) class. Markers selected by RFE-SVM Figure 20-C), on the other hand, are not as significantly differentiated between the two classes and thus, they hardly reflect differentially expressed regions; this is a drawback for domain experts, who are

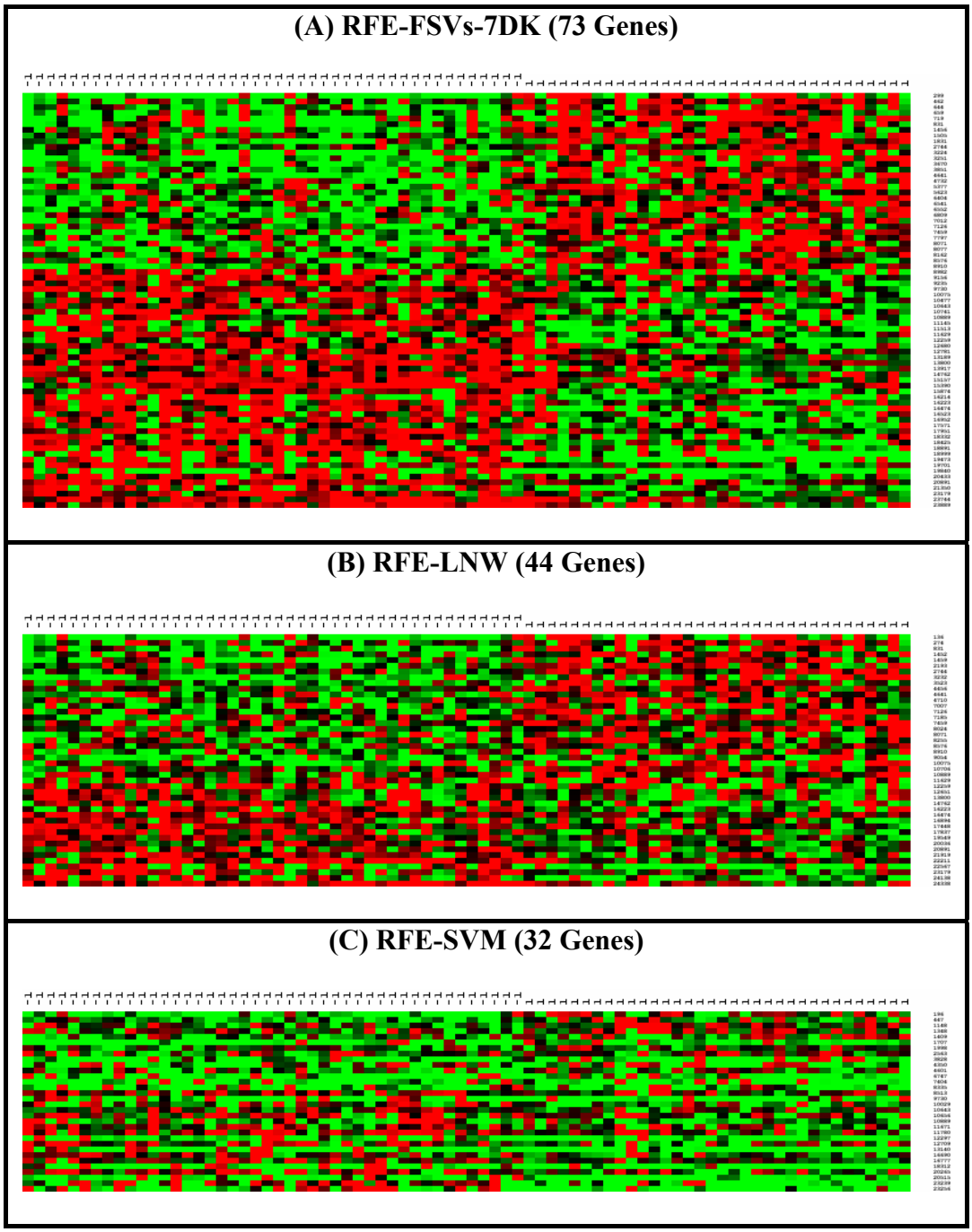

Figure 20: Expression profile analysis of the selected markers derived by the tested methodologies in breast cancer; genes selected by the proposed methodologies (RFE-FSVs-7DK and RFE-LNW) demonstrate significantly higher expression variation between the two classes of interest than those selected by RFE-SVM.

searching for genes that differentiate their expression significantly in the two classes of interest, as was pointed out by studies reported in the introduction section.

On the leukemia domain, RFE-SVM methodology achieves outstanding results both in terms of classification accuracy as well as in terms of the expression profiles of the selected markers as reported in [1]. Nevertheless, it does not achieve the same level of result on the BC domain. This is due to the fact that BC is a 'harder' data set with a lot of overlap between the two classes as was presented in section 2.5, and thus the need for other constraint characteristics besides accuracy (such as differentiation on the expression level of the selected genes) becomes crucial. The two proposed methodologies show a more stable performance both in terms of accuracy as well as of the 'quality' of the expression profiles on the tested domains as illustrated in Figure 20 (A) and Figure 20 (B).

The genes selected by the tested methodologies are reported in Table 1, Table 2 and Table 3 in Appendix II. A worth mentioning point is the fact that there are 18 common genes found between the markers selected by RFE-LNW and RFE-FSVs while there are only 3 common genes between RFE-SVM and RFE-FSVs and 1 common gene between RFE-SVM and RFE-LNW.

Biological Relevance

Proceeding one step further, we assess the biological significance of the derived result using the biological knowledge that might be hidden behind the underlined gene signatures discovered by the proposed methodologies in the previous paragraph. To assess such an approach we searched for the Gene Ontologies Biological Processes (GOBPs) [27] hidden behind the underlined gene signatures and measured their overlap with those GOBPs underlined by two well known and broadly accepted breast cancer signatures namely, Van't Veer's [4] and Wang's [25], respectively. We locate

46 different GOBPs covered behind the 70-gene Van't Veer's signature and 71 different GOBPs hidden behind the 76-gene Wang's signature, while they share six GOBPs in common. Results are outlined in Table 8 where we notice that RFE-FSVs-7DK demonstrates the highest degree of overlap with 18 GOBPs in common with Van't Veer's signature and 16 GOBPs common with Wang's signature. RFE-LNW comes next with 8 and 16 GOBPs, respectively, in common with the underlined gene signatures. RFE-SVM shows the minimum overlap with 3 and 6 GOBPs, respectively.

Method	GO Overlap - Van't Veer	GO Overlap - Wang
RFE-FSVs-7DK	18	16
RFE-LNW	8	16
RFE-SVM	3	6

Table 8: Breast Cancer Domain: GO overlap of the underlined methodologies with Van't Veer's and Wang's signatures. RFE-FSVs-7DK and RFE-LNW demonstrate a significant overlap with both signatures. The overlap between Van't Veer's and Wang's signatures in terms of GOs is 6.

Regarding the absolute gene overlap, we report an impressive overlap of 17 genes (25% overlap) found between the 73-gene signature of RFE-FSVs and the 70-gene Van't Veer's signature. RFE-LNW and RFE-SVM on the other hand demonstrated an overlap of 5 and 3 genes respectively.

Taking into consideration the above results, we observe a strong evidence that the proposed methodologies search towards a biologically meaningful path. This evidence is obviously stronger for RFE-FSVs, which demonstrates a higher level of overlap both in terms of GOs and absolute genes, but evidence is also strong for RFE-LNW since it also demonstrates a significant degree of GO overlap.

GEMS Experiment

In this experimental procedure GEMS, and other approaches were applied in a similar manner as in leukemia data set in section 2.5.2. Results are visualized in Figure 21,

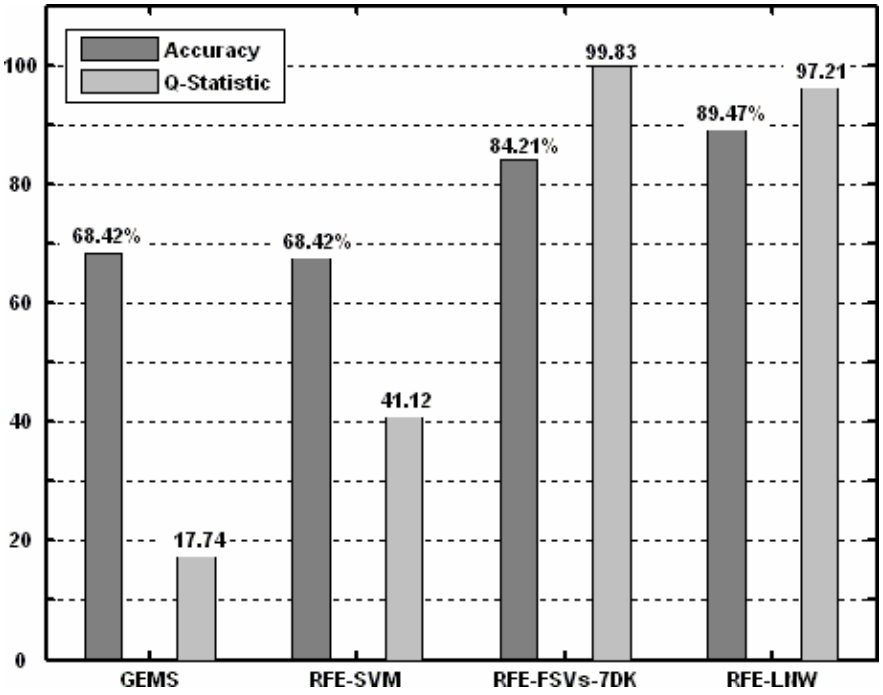

Figure 21: Performance evaluation of the tested methodologies along with GEMS. Results are recorded at the level of 55 genes, i.e., the average number of genes selected by GEMS. Significant difference in favor of the proposed methodologies is observed, both in terms of classification accuracy and statistical significance.

where we notice a significant advantage of the proposed methodologies (RFE-LNW, RFE-FSVs-7DK) both in terms of accuracy performance and statistical significance.

RFE-LNW achieved an accuracy performance of 89.47% on the independent test set (missing only 2 samples) while RFE-FSVs-7DK reached a success rate of 84.21% (missing 3 samples). Both GEMS and RFE-SVM reached an accuracy performance of 68.43% (missing 6 samples). A significant advantage on the proposed methodologies is also noticed on the statistical significance of the derived result, as demonstrated in Figure 21 (Q-Statistic score), further supporting the over all advantage of the proposed selection approaches.

2.6 On the Utilization of Kernels and Support Vectors

In this section a final experiment was conducted to show the distribution of support vectors across the entire feature elimination process and addresses the over-fitting problem. Over-fitting occurs when a large proportion of samples become support

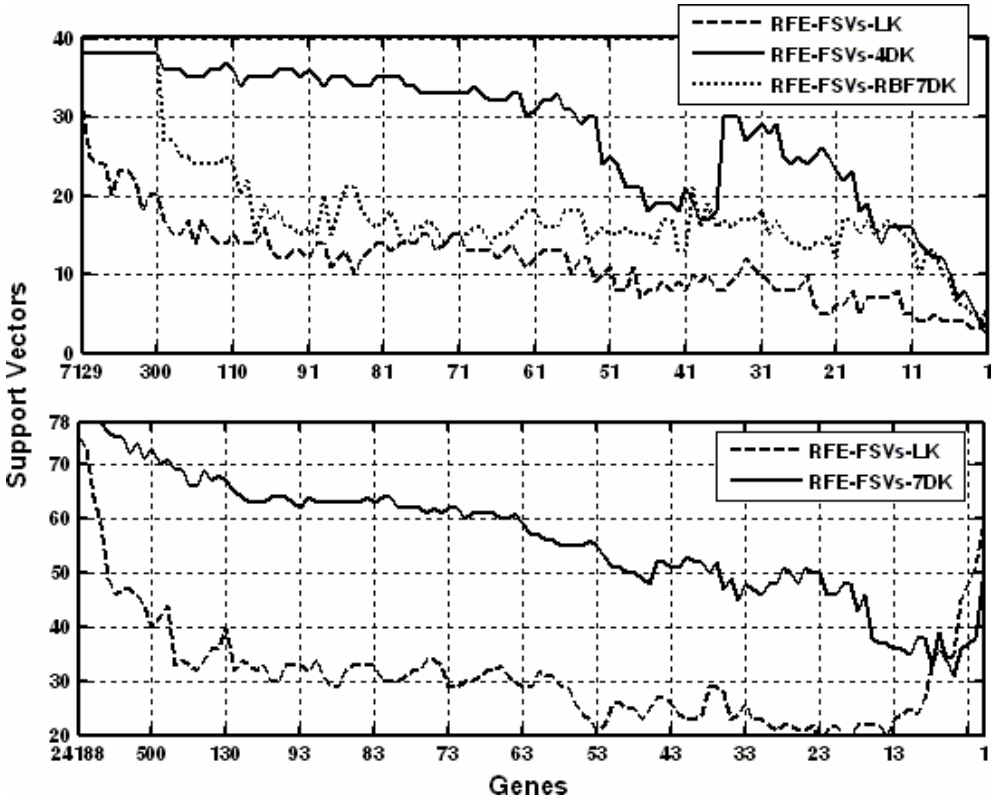

Figure 22: Support vector allocation by the kernels used on the Leukemia domain (upper graph) and on BC domain (lower graph).

vectors and hence, the classifier can not generalize well on new unseen data. An extreme case of over- fitting exists when all training samples are support vectors.

Results of the linear as well as of the polynomial kernels used in both tested domains are depicted in Figure 22. Notice that the scale in the horizontal axis of Figure 22 follows the arrangement of column 2 in Tables 1 and 3 of Appendix I respectively. We observe that by using polynomial kernels the average number of support vectors increases, as it is expected. This verifies the fact that polynomial kernels are more sensitive to over fitting, since a large proportion of the samples become support vectors. Even though this is true especially when dimensionality is still very large (initial steps of the feature elimination process), it is not a drawback of the proposed methodology, since support vectors are only used to assess the ranking criterion and not for classification of new unseen data. More specifically, even though the training set is classified using a polynomial kernel, new unseen samples are tested

linearly where over-fitting is potentially much less of a problem. As depicted in Figure 22, the number of support vectors used by a linear kernel in both domains reduces drastically. A final comment regarding the RFE-FSVs method pertains to the possibility that it could degrade to the initial filter method only in the very seldom case where all samples are always kept as support vectors. In the case of kernel combination, as in leukaemia data set where an RBF kernel was combined with a polynomial one, we observe that (Figure 22 upper graph) up to the 300 surviving genes all samples are retained as support vectors but after this cut off point a significant decrease is observed, rendering the performance of our method quite different from that of a filter approach.

2.7 Discussion and Conclusion

Within the field of marker selection, wrapper approaches are very much dependent on the classifier or the pattern recognition approach used to assign the weights for ranking the features (genes). On the other hand, filter methods take into account only intrinsic characteristics of the data, such as the differentiation ability of each gene. Our aim in this study is to bring these two methodologies together by embedding filtering criteria in the wrapper 'philosophy' and integrate the two approaches. Special attention is given to Fisher's ratio which has been extensively used; but other filter metrics can be considered as well.

The methodologies proposed produce comparable results and at cases even better than the best reported, demonstrating that by wrapper filtering criteria we fully utilize the advantages of wrapper methods while on the other hand we are not neglecting the very important property of differential expression, which is mostly addressed by filter methods. This merging results to stable performance along different problem domains

demonstrated on two data sets, one considered 'easier' while the other considered a 'harder' domain.

Alternatively, the proposed methodologies could be seen as a means of selecting markers that significantly differentiate their expression between positive and negative classes, but also lead to high classification performance. This aspect has not been addressed thoroughly within bioinformatics research, even though domain experts and biostatisticians either implicitly or explicitly are searching for it.

Inspired by the results of this study, we proceed on to the next section, where the evaluation platform is further extended, completing the evaluation measures introduced here. We focus in RFE-LNW and compare its performance to that of RFE-SVM using additional evaluation criteria, but with aim in deriving an effective and reliable gene signature through such an evaluation measure. Even though additional data sets are used for testing, special focus is paid in breast cancer for a final gene signature derivation.

References

[1] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, Gene selection for cancer classification using Support vector machines, *machine learning*, 36 (2002) 389-422.

[2] I. Inza, P. Larranaga, R. Blanco and A. J. Cerrolaza, Filter versus wrapper gene selection approaches in DNA microarray domains, Artificial Intelligence in Medicine 31 (2004), 91-103.

[3] R. T. Golub, K. D. Slonim, P. Tamayo, C. Huard, et al., Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, *Science*, 286 (1999) 531-536.

[4] L. J. Van't Veer, H. Dai, M. J. Van de Vijver, Y. D. He, et al., Gene expression profiling predicts clinical outcome of breast cancer. *Letters to Nature* 415 (2002) 530-536.

[5] C. Nutt, D. Mani, R. Betensky, P. Tamayo et al., (2003), Gene Expression-based Classification of Malignant Gliomas Correlates Better with Survival than Histological Classification, *Cancer Research*, 63 (2003) 1602-1607.

[6] S. Armstrong, J. Staunton, L. Silverman, R. Pieters, et al., MLL translocations specify a distinct gene expression profile that distinguishes a unique leukaemia, *nature genetics*, 30 (2002), 41-47.

[7] X. Zhou and K. Z. Mao, LS Bound based gene selection for DNA microarray data, *Bioinformatics*, 21:8 (2004) 1559-1564.

[8] E. Ke Tank, PN Suganthan and X. Yao, Gene selection algorithms for microarray data based on least square support vector machine, *BMC Bionformatics* 7:95 (2006) doi:10.1186/1471-2105-7-95.

[9] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin and S. Levy, A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis, Bionformatics, 21:5 (2005) 631-643.

[10] A. Statnikov, I. Tsamardinos, Y. Dosbayev and C. F. Aliferis, GEMS: A system for automated cancer diagnosis and biomarker discovery from microarray gene expression data, International Journal of Medical Informatics 74 (2005) 491-503.

[11] N. V. Vapnik, *The Nature of Statistical Learning Theory* (Springer–Verlag New York, 1999).

[12] S. Boyd. And Vandenberghe L, *Convex Optimization* (Oxford University Press 2004).

[13] Riedmiller M. and Braun H. (1993), A direct adoptive method for faster backpropagation learning: The RPROP algorithm. *Proceedings of the IEEE International Conference on Neural Networks (ICNN)*, 586-591.

[14] R. Simon, M. D. Radmacher, K. Dobbin, L. M. McShane, Pitfalls in the use of DNA Microarray Data for Diagnostic and Prognostic Classification, *Journal of the National Cancer Institute* 95 **(**2003) 14–18.

[15] A. Little and D. Rubin, *Statistical Analysis with Missing Data* (Wiley Series in Probability and Mathematical Statistics, (1987).

[16] C. Ambroise and G. J. McLachlan, Selection bias in gene extraction on the basis of microarray gene-expression data, *PNAS* 99 (2002) 6562-6566.

[17] Kohavi R., A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceeding of Fourteenth International Joint Conference on Artificial Intelligence (IJCAI)* (1995) San Mateo CA, pp. 1137 1143.

[18] http: //www.gems-system.org

[19] Goeman J., van de Geer S., de Koort F and van Houwelingen H., A global test for groups of genes: testing association with clinical outcome, *Bioinformtics* 20 (2003) 93-99.

[20] Li F. and Y. Yang, Analysis of recursive gene selection approaches from microarray data, *Bionformatics* 21 (2005) 3741-3747.

[21] H. Takahashi., T. Kobayashi and H. Honda, Construction of robust prognostic predictors by using using projective adaptive reasonance theory as gene filtering method, *Bioinformatics*, 21 (2004) 179-186.

[22] S. G. Baker and B. S. Kramer, Identifying genes that contribute most to good classification in microarrays, *BMC Bioinformatics*, 7:407 (2006), doi: 10.1186/1471-2105-7/407.

[23] Y Wang, IV Tetko, MA Hall, E. Frank, A Facius, K.F.X. Mayer, H.W Mewes, Gene Selection for microarray data for cancer classification – a machine learning approach, Computational Biology and Chemisry, 29 (2005), 37-46.

[24] T. M. Huang and V. Kecman, Gene extraction for cancer diagnosis by support vector machines – An Improvement, *Artificial Intelligence In Medicine* 35 (2005) 185-194.

[25] Y. Wang, J. Klijn, Y. Zhang, A. Sieuwerts, M. Look, F. Yang, D. Talantov, M. Timmermans, M. Meijer-van Gelder, J. Yu, T. Jatkoe, E. Berns, D. Atkins and J. Foekens, Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer, *Lancet* 365 (2005) 671-679.

[26] R. Shen, D. Ghosh, A. Chinnaiyan and Z. Meng, Eigengene-based linear discriminant model for tumor classification using gene expression microarray data, *Bioinformatics* 22 (2006) 2635-2642.

[27] http://www.geneontology.org/

CHAPTER 3

The Linear Neuron as Marker Selector and Clinical Predictor in Cancer Gene Analysis

3.1 Abstract

Objective: Having established the integration of filter and wrapper methods by means of network structures, in this chapter we study the behavior of the linear neuron using more stringent evaluation criteria along with additional application domains. Special attention is paid in breast cancer, our aim is to show that the proposed methodology when used along with appropriately designed evaluation criteria leads to a statistically significant gene signature comparable to bench mark results in breast cancer.

Methods and Materials: We explore the proposed approach in terms of accuracy evaluation criteria, which are used to assess the performance of the proposed methodology, but we also evaluate the produced results in terms of cluster quality and survival prediction. Cluster quality reflects the ability of the method to select differentially expressed genes, which in turn leads to better clustering and survival prediction.

Results: We directly compare the proposed methodology with RFE-SVM, a well known and broadly accepted method demonstrating remarkable performance on various data sets of clinical interest.

3.2 Introduction

Among the various feature selection methods proposed, RFE-SVM [1] (see also section 2.3.2) is an approach that has shown remarkable results in leukemia [2] and colon cancer [3] datasets. However, depending on the data distribution and the complexity of the classification problem, the algorithmic design based on the philosophy of RFE-SVM may lead to ill-defined and ill-distinctive clusters of selected

gene signatures, as it is demonstrated in the results section and to some extend in chapter 2. This issue has been implicitly addressed in [4], showing that wrapper methods do not provide sufficient focus for further investigation (of their result) because many genes may be included by chance. In this study we go one step further and demonstrate that such a lack of focus could lead to the production of ill-distinctive clusters, by-passed however, by the proposed RFE-LNW approach. The proposed network acts as a linear filter for classification producing more compact and distinct clusters of markers.

The idea of applying a linear neuron to the problem of gene selection is a novel approach, introduced in chapter 2 and in study [5]. It is based on the ability of a single neuron to approximate any linear function. This idea absolutely complies with the philosophy of linear methods such as the RFE-SVM, which is based on linear support vector machines. In this study, we apply RFE-LNW in combination with an appropriately designed evaluation platform so that biologically relevant results are eventually derived.

Besides methodology, another important issue concerning gene selection relates to the measures used for validation of the classification performance (prediction rule) of the selected markers. It is a common practice to assess the performance of a method by its Leave One Out Cross Validation (LOOCV) error. Two types of LOOCV schemes are generally considered and both are assessed in this study. The first one addresses the removal of the left-out sample before the selection of differentially expressed genes and the application of the prediction rule, while the second approach handles the removal of the left-out sample after the selection process but before the application of the prediction rule. The first is usually referred to as the External

LOOCV (ELOOCV) while the second is referred to as the Internal LOOCV (ILOOCV) [6].

It is obvious that ELOOCV is a more unbiased estimator of the error rate since it is totally independent of the selection process. However, ILOOCV provides a measure that can not be neglected, as it expresses the training ability of a selection rule within the training set. In other words, ILOOCV indicates the selection rule(s) that can learn or generalize better on the training set. The ILOOCV scheme in combination with an independent test set was used to assess the performance of RFE-SVM in [1]. Towards a fair and statistically sound comparison of the methods in this study, we assess both of these measures in combination with a 10-fold cross validation process in all evaluation steps. Proceeding one step further than classification accuracy and cluster quality, we test the result derived through the ELOOCV procedure as a set of selected genes that can provide high classification accuracy on the independent test published by Van't Veer et al. in [7]. In addition, this set of markers is proved to be an efficient survival predictor for the 234 new cases published by Van De Vijver et al. in [8].

To reveal the basic differences of the two underlined methodologies (RFE-SVM and RFE-LNW) we examine their learning ability on three data sets: a) the data set of Diffuse Large B-Cell lymphoma published in [9], b) the colon cancer data set in [3] and c) the breast cancer data set published by Van't Veer et al. in [7].

3.3 The RFE-LNW Ranking Criterion

Within RFE-SVM (see section 2.3.2) a linear kernel with equation (2.1) is used to estimate the weight vector of the separating hyperplane. On the other hand, a Linear Neuron (LN) can approximate any linear function. The architecture of such a network is depicted in Figure 15 (chapetr 0), where a single neuron fed by all inputs (genes) is actually the only learning component. We use such a linear neuron to approximate the

separating hyper-plane between positive and negative classes in the place of the linear SVM used in RFE-SVM. This is applied as a linear filter of *m* inputs, where *m* corresponds to the number of genes. For the appropriate operation of the proposed methodology we consider two possible outcomes at the output layer, namely *output 0 for the negative class and 1 for the positive class.* We can then use such a network to approximate the separating hyper-plane, and hence the ranking criterion of genes. The rule that is used to iterate the linear neuron weights is given below and was further analyzed in section 2.3.4

$$w_i(t+1) = w_i(t) + \mu \cdot sign\left(e \cdot f'(u)\right) \cdot sign(g_i) \cdot K_i \quad (2.30)$$

$$K_i = \frac{|g_i - c|}{\sigma_+(g_i) + \sigma_-(g_i)} \quad (2.31)$$

$$c = \frac{\mu_+(g_i) + \mu_-(g_i)}{2} \quad (2.32)$$

where $\sigma_+(g_i)$, $\sigma_-(g_i)$, $\mu_+(g_i)$ and $\mu_-(g_i)$ are the standard deviations and mean values of gene g_i in the positive and negative classes, respectively. A close alternative for the weight factor K_i is defined as:

$$K_i = \frac{|\mu_+(g_i) - \mu_-(g_i)|}{\sigma_+(g_i) + \sigma_-(g_i)} \quad (2.33)$$

Elaborating in equation (2.30) with either (2.31) and (2.32) or (2.33) we notice that by taking sign values the method proceeds towards the direction of the minimum, which eventually will be reached by selecting the appropriate learning rate and an adequate number of iterations (epochs). The idea of using the sign of the gradient instead of its actual value was instigated by resilient backpropagation method, which is used to speed up convergence [10].

The term K_i introduced in equation (2.31) or alternatively in (2.33) is a variation of Fisher's coefficient and plays an important role in the gene selection problem. Elaborating on equations (2.30) and (2.31) or (2.32) one can easily verify that genes which differentiate their expression across samples in the two classes of interest i.e., green in negative class (-3 expression value) and red in positive class (+3 expression value) will take higher K values than genes showing the same expression across samples within the two classes. By introducing term K_i in equation (2.30), we aim at assigning higher weight values to those genes that significantly differentiate their expression from negative to positive class (for more details refer to section 2.3.3).

Employing the proposed learning procedure we are not only interested in a hyperplane that distinguishes the two classes of interest but we are also searching, through the coefficient K_i, for a hyperplane that assigns higher weights to those genes that differentiate their expression more. Since the method is based on gradient descent, convergence is guaranteed but the algorithm may be trapped to a local minimum, which is a known problem of neural network methods. In our methodology, however, the neural kernel is only used for ranking genes, rendering the requirement of a global minimum of less importance than in other optimization frameworks. Furthermore, the proposed methodology could be extended to a multilayer network and applied to non linear problems.

3.4 Cluster Quality Measure

The validity or quality of features that survive at each step of the elimination process is an issue of particular interest. Based on the desirable properties of markers, the surviving genes should form well defined clusters related to the pathology states of interest. In other words, the clusters of selected genes should express small intra-class but large inter-class distance [11]. One measure to asses cluster quality is known as

Davies-Bouldin (DB) index [12] that has been extensively used to assess cluster quality in various fields besides DNA microarray analysis [13], [14], [15]. The DB index for a partition U that is composed of two clusters, namely X_P corresponding to the positive class and X_N to the negative class, is given by:

$$DB(U) = \frac{\Delta(X_P) + \Delta(X_N)}{\delta(X_P, X_N)} \qquad (2.34)$$

where $\delta(X_P, X_N)$ corresponds to inter-class distance given by:

$$\delta(X_P, X_N) = \frac{1}{|X_P||X_N|} \sum_{\substack{x \in X_P \\ y \in X_N}} d(x, y) \qquad (2.35)$$

where $|.|$ denotes set cardinality, $d(x,y)$ is the Euclidean distance between two samples x and y and $\Delta(X_P)$ represents the intra-class distance given by:

$$\Delta(X_P) = \frac{1}{|X_P|(|X_P|-1)} \sum_{x,y \in X_P} d(x, y) \qquad (2.36)$$

with $\Delta(X_N)$ analogously defined. Optimization of the DB index minimizes intra-class distance while maximizing inter-class distance. Therefore, smaller values of DB reflect better clusters.

3.5 ELOOCV and 10-Fold Cross Validation

As pointed out in section 3.2, ELOOCV is a more unbiased estimator for evaluating the performance of a prediction rule on a totally independent test set ([6], [16]). It is repeated as many times (say *n*) as the number of examples the training set. One sample is left out from the training process at each iteration, and then that left out sample is tested on the classifier produced by the remaining $n-1$ samples. ELOO differs from ILOO in that the left out sample is not included in the training set during the gene selection phase. At each stage of the feature elimination process the training

set is used to construct the separating hyperplane on the space defined by the surviving genes (variables). Then the left out sample is tested using the generated hyperplane and either a success (1) or failure (0) is measured. When the process ends, we are left with an $m \times n$ matrix S composed of 1s and 0s as its elements S_{ij}, where *m* being the number of repetitions or number of cut off points according to the cut off strategy used in the ELOO iterations. The average vector *A* of *m* components is defined with elements:

$$A_i = \frac{1}{n}\sum_{j=1}^{n} S_{ij} \qquad (2.37)$$

each measuring the average success rate on the performance of the examined methodology at the ith stage of the feature elimination process. The component of *A* where maximum accuracy is observed deserves special attention, since it indicates the optimal cut off point and thus, the number of markers at which the examined methodology achieved its best performance. This value can also be used as a measure of stability regarding the number of selected markers. At the extreme case for instance where all components of *A* take consistently low values, the examined methodology is not expressing any stable performance on the number of markers it selects, since there is not a clear stage where the algorithm (in most of the cases) achieves its best acceptable performance.

3.6 Results

Three different data sets are used to test the two methods:

(1) The data set of Diffuse Large B-Cell Lymphoma (DLBCL) published in [9] consisting of 50 negative (non DLBCL) and 46 positive (DLBCL) patients described through a set of 4026 gene expression profiles. Missing values were substituted using (EM) imputation [17] and overall data set variance is 0.63,

i.e., variance of the distribution defined by the whole set of gene expression values across all samples.

(2) The colon cancer data set published in [5] consisting of 2000 genes and 22 normal as well as 40 tumor tissues. The data set was normalized to a zero mean and standard deviation of one as suggested in [5]; no missing values were encountered in this data set, overall variance is 0.98.

(3) The data set provided by [7] in Breast Cancer (BC), consists of 24481 gene expression profiles and 78 samples, 44 of which correspond to patients that remain disease free for a period of at least five years, whereas 34 correspond to patients that relapsed within a period of five years. 293 genes express missing information for all 78 patients and were removed, while other missing values were substituted using Expectation Maximization (EM) imputation [17], overall data set variance is 0.06. We are also provided with a set of 19 new samples [7] which is going to be used as an independent test set for evaluation purposes, as well as with a set of 234 new samples [8] used to independently test clinical prediction outcome.

Comparing these data sets we notice that BC consists of a significantly lager number of genes than the other two and it also expresses a significantly lower overall gene expression variance along the two classes. These two facts indicate that the BC data set could be proved the most complicated and the hardest one to classify. For this reason we employ it for elucidating additional properties of the proposed methodology.

As explained in section 3.2, three series of computational steps are conducted. In the first series the ILOOCV along with the quality of the selected markers using the DB index is assessed, whereas in the second series the ELOOCV in combination with

a repetitive 10-fold process is considered for deriving a promising gene set (gene signature). In a third step the performance of the derived gene signature is tested as a clinical predictor in BC data set.

3.6.1 ILOOCV – Cluster Quality Results

The ILOOCV is used as a measure of estimating the learning ability of the studied methodologies rather than as a measure of independent generalization performance. Besides ILOOCV, we also measure the quality of the selected features (genes) in terms of low intra but high inter class distance. The DB index is used as a measure to asses this feature. In addition, we also measure the area under the ROC curve (AUC) across the entire feature elimination process as well as the number of common genes in the last seven elimination steps of the process (from 64 to 1 surviving gene).

For the RFE-LNW approach, 300 epochs were used to train the network as long as the number of remaining features was greater than or equal to 1024 and after this point the number of epochs was fixed to 1000. A learning rate of 10^{-2} was used and equation (2.30) was applied in combination with (2.31) as our iterative learning process. The parameter selection may vary from problem to problem or from run to run and should be appropriately defined. In our application a learning rate within the range of approximately $\left[10^{-1}\cdots10^{-4}\right]$ most often provided satisfactory results. On the other hand, the optimal parameter selection is a problem which needs deeper investigation. For the steps conducted in this section the used parameters were experimentally tuned. We increase the number of epochs as we approach the end of the process for better approximation of the separating hyperplane.

DLBCL Results

In this data set RFE-LNW achieved an average accuracy of 97%, while the RFE-SVM reached an average accuracy of 95%. There exists a slight advantage of RFE-SVM up

to the point of 4 surviving genes. The overall performance of the two methods is depicted in Figure 23 (Panel 1) where we observe that the accuracy of RFE-LNW never drops below 90% (Figure 23, Panel 1.A), while the AUC measures follow about the same pattern (Figure 23-Panel 1.C). On the other hand, of particular interest are the last two steps of the process where we observe a significant difference in favour of RFE-LNW, which retains a success rate of over 90% while RFE-SVM drops to a rate of about 70%. This behaviour could be explained by studying the cluster quality measure (Figure 23-Panel 1.B). We observe that the difference of the DB index between the two tested methodologies takes its highest values in the same last steps, implying a connection between cluster quality and success rate. This fact could be further elaborated by looking at the number of common genes selected by the two tested methodologies within the range of 64 to 4 surviving genes (Figure 23, Panel 1.D). We observe that within this interval the percentage of common over the total surviving genes varies from approximately 25% to 50%. Having a rather high number of common genes, the tested methodologies express only a small difference on success rate and AUC measures between them. In contrast, during the last two steps the tested methodologies take quiet different directions, verified by the fact that there are no common genes selected. In addition, returning to Figure 23 (Panel 1.B) we observe that RFE-SVM selects genes that decrease both cluster quality (higher DB index) and classification accuracy. The RFE-LNW on the other hand, selects genes that increase cluster quality and hence it is rewarded by retaining a high success rate, similar to the previous steps of the process. Considering the overall cluster quality of the selected markers, we notice an overall significant advantage of RFE-LNW, as demonstrated in Figure 23 (Panel 1.B).

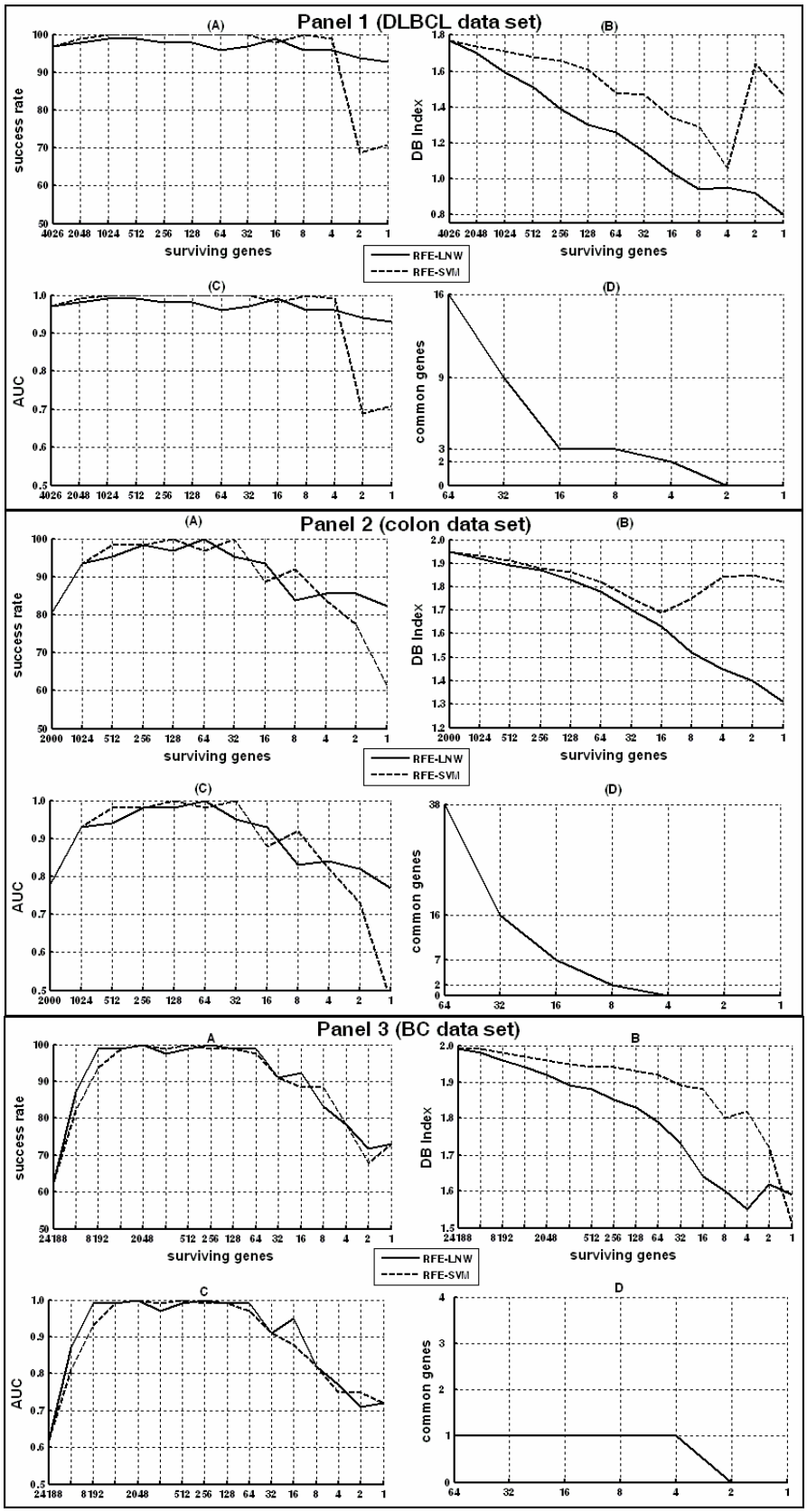

Figure 23: Performance of tested methodologies in DLBCL data set (Panel 1), in colon data set (Panel 2) and in breast cancer (Panel 3); accuracy evaluation (A), quality evaluation index (B), Area Under the ROC curve (C) and common genes between methods (D).

Colon Results

In Colon data [3], RFE-LNW succeeded an average success rate of 91%, while RFE-SVM reached an average rate of 89%. The learning performance of the underlying methods is presented in Figure 23 (Panel 2.A and 2.C), where we again notice that the performance of RFE-SVM falls significantly behind during the last two steps of the process, while in the previous steps both methods are almost equivalent. In the last steps of the process the performance of RFE-SVM falls at almost 60%, while RFE-LNW retains a relatively high performance of above 80%. As in the previous steps, we verify that as long as there is not significant difference in the quality performance of the two methods (Figure 23, Panel 2.B), there are also insignificant differences in their accuracy performance (Figure 23 Panel 2.A). This connection between cluster quality and accuracy is further related with the number of common genes between the two processes (Figure 23 Panel 2.D), where we observe that up to the point of 16 surviving features the percentage of common genes is rather high, varying between 59% (at 64 surviving genes) to 44% (at 16 surviving genes). However, from the point of 8 surviving genes we verify abrupt reduction in the number of common genes along with significant difference on the quality performance in favor of RFE-LNW.

BC Results

In BC data set [7], both methods achieve an average success rate of 89%. Their learning performance is visualized in Figure 23; Panels 3.A and 3.C. A measure that implies differences in the applied approaches is the cluster quality index along with the number of genes they share. As depicted in Figure 23, Panel 3.D there is only one common gene shared between the two methods within the range of 64 to 4 surviving genes and as in the previous two cases, there are no common genes in the last two steps of the process. Concerning the quality performance of the selected markers

(Figure 23 Panel 3.B) a significant advantage of RFE-LNW over RFE-SVM is verified. Even though we can not make a direct connection between cluster quality and performance evaluation, as in the previous two cases, an indirect effect is elucidated by studying the quality of the 64 markers selected by each method in Figure 24.A and Figure 24.B.

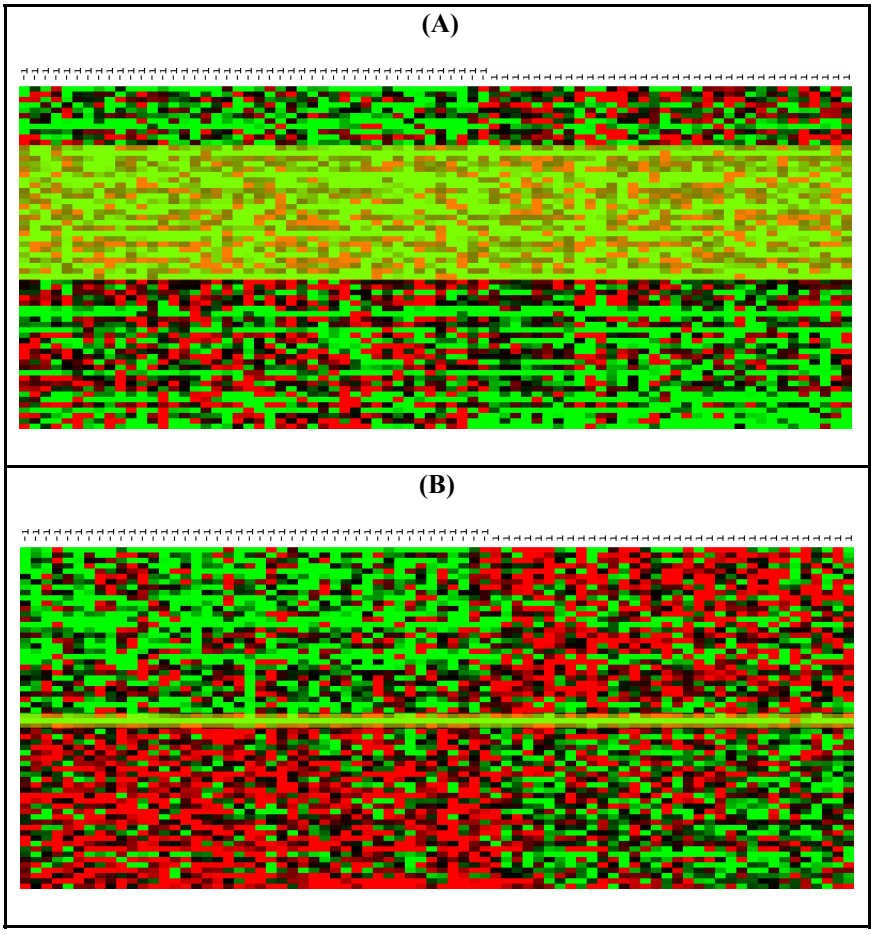

Figure 24: The 64 markers selected by RFE-SVM method (Panel A); genes are not as significantly differentiated as in the case of RFE-LNW, resulting to "ill defined" expression regions of selected markers. The 64 markers selected by RFE-LNW (Panel B) demonstrate significant differences on their expression levels revealing four distinct regions.

In those two figures genes are ranked in ascending order (from negative to positive values) according to Fisher's correlation. The markers selected by RFE-SVM (Figure 24.A), even though they show a local variation on their expression levels (unshaded regions of the graph), their expression difference between classes is not as significant

as in the case of RFE-LNW in Figure 24.B. Thus, RFE-SVM leads to ill-defined gene expression regions across classes, as mentioned earlier in the introduction section and

demonstrates also the lack of focus problem encountered in wrapper methods [4]. An additional aspect related to this RFE-SVM inadequacy pertains to genes that do not significantly differentiate their expression across classes (Fisher'r ratio close to 0, i.e. genes expressed in almost the same way across classes), corresponding to the shaded regions in the two figures. We observe that this region is much wider in the case of RFE-SVM, while it is much smaller in the case of RFE-LNW (consists of only one gene), supporting the case that the proposed methodology indeed selects more significantly differentiated genes. Thus, the DB-index advantage for RFE-LNW over RFE-SVM might not be translated into direct influence on the classification accuracy, but certainly reflects qualitative differences in the set of selected genes as confirmed by the figures of the two presented gene sets.

As a general concluding remark on these three results we state that the proposed methodology is able to generalize relatively well on the training set (as verified by the ILOOCV results), demonstrating also the convergence ability of the training method. Its accuracy performance on all tested data sets is comparable to that achieved by RFE-SVM. In 2 out of the 3 cases we observe an overall advantage of RFE-LNW during the last two steps of the selection process and a direct connection between cluster quality and accuracy performance. The third experimental step (on the more complex BC data set) reveals an indirect advantage of RFE-LNW related to the quality of the selected markers. As it is demonstrated by these results, RFE-LNW creates more compact and distinctive cluster of markers verified by the DBindex quality measure.

3.6.2 ELOOCV and 10-Fold Cross Validation

We measure and present graphically the ELOO scores of the two methods on the same data sets as in the previous section, where tick points of the graphs represent the components of vector A as described in equation (2.37). To train the RFE-LNW, 300 epochs were used as long as the number of remaining features was greater than or equal to 1024 and 500 epochs afterwards. A learning rate of 10^{-2} was used, except in the case of BC where a learning rate of 10^{-4} was employed. Notice that in contrast to the internal scheme, in external CV we are using smaller number of epochs in order to avoid close convergence and overtraining, so as to achieve better generalization performance. The learning rate for ELOOCV in the more complex BC data set was reduced to 10^{-4} for the same reason. These parameters, as in the case of ILOOCV, were assessed experimentally.

Even though ELOOCV permits the study of algorithmic performance with less bias than ILOOCV, its main disadvantage is that it induces high variance on the estimated accuracy. Thus, besides evaluating ELOO we also evaluate in a similar manner the method of 10-fold cross validation that was repeated 100 times in our attempt to produce more reliable and robust results. The given set of 78 samples was randomly partitioned 100 times, such that 90% of it was used for training purposes while the remaining 10% was used as an independent test set. In each run the success rate at each cut off point is measured and finally the overall average at each point of the feature elimination process is plotted.

DLBCL Results

The overall ELOOCV accuracy of the selection process is presented in Figure 25 (Pane 1.A), where it appears that the RFE-SVM performs better than the RFE-LNW up to 64 features. Nevertheless, after this step it starts reducing its performance until it

finally reaches around 70% in the last two steps of the process, as it was also reported in the ILOOCV result. The RFE-LNW on the other hand never falls below 90% in the entire phase of the selection process. In terms of absolute numbers, RFE-SVM reaches its best performance at 256 genes with 98% success rate. On the other hand RFE-LNW achieves its best performance at 128 with 94% success rate, but comparable performance is maintained at 4 genes with 93% success rate. As a conclusion, RFE-SVM achieves slightly better classification accuracy with a larger number of genes while RFE-LNW selects a smaller number of genes with slightly less accuracy but always preserved over 90%. The one gene that is most frequently selected by RFE-LNW (80/96 frequency rate) and achieves an accuracy of approximately 90% corresponds to GENE1637X (CCND2). This gene encodes the cyclin D2 protein involved in the phosphorylation of tumor suppressor protein Rb, which malfunctions in nearly all malignant gliomas as well as in many other solid tumors. Focusing on Rb, recent and very promising research in brain cancer [18] produces encouraging results even for therapeutic purposes.

Concerning 10-fold cross validation, (Figure 25, panel 1.B) we notice that the performance curves follow a smoother pattern due to the lower variance of the produced result. The performance difference of the two methods in the last two steps of the process is still verified, even though it is not as significant as in the case of the ELOOCV.

Colon Results

The overall ELOOCV performance of the two methods is visualized in Figure 25 (panel 2.A). We observe that RFE-LNW achieves its average best performance of 85% with only one gene, implying that in the 85% of the cases it classifies correctly the left out sample by using only one gene, which is almost always the same (61/62

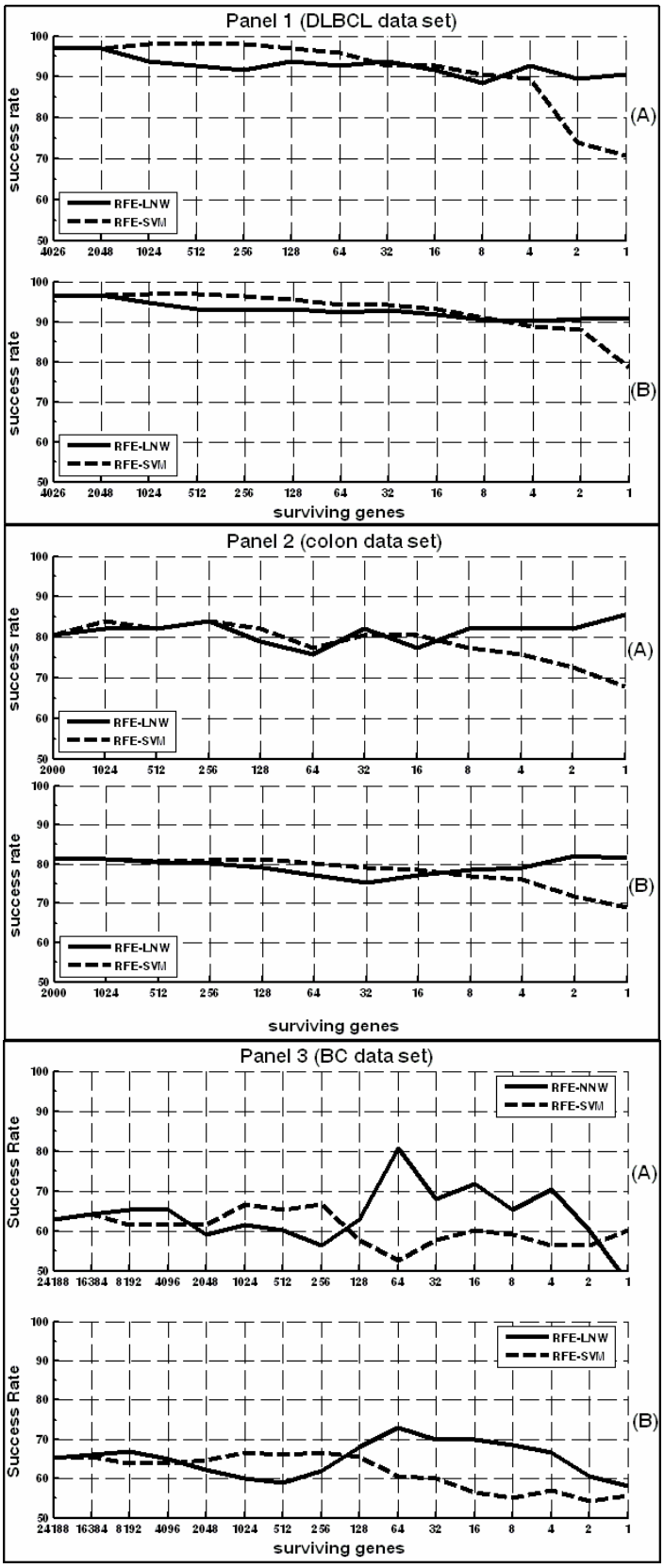

Figure 25: ELOOCV (A) and 10-fold (B) cross validation performance curves of the two methods in DLBCL (panel 1), colon cancer (panel 2) and BC (Panel 3) data sets. A significance difference is expressed in favour of RFE-LNW especially in breast cancer.

frequency rate) for all runs. Performance of over 80% is consistently maintained from the stage of 8 surviving genes. On the other hand RFE-SVM achieves its best performance of 84% at the 256 genes cut-off point. Then, from the point of 16 surviving genes its performance drops below 80%. In this application the performance of RFE-LNW is better, since by selecting a smaller number of genes on the average (only one) it achieves better classification accuracy. The performance of RFE-SVM up to 16 genes is close or slightly better, at places, than that of RFE-LNW. We again emphasize the drop in the performance of the RFE-SVM method in the last steps of the process, a fact that was also verified in the ILOOCV results and was directly linked to the produced cluster quality.

Concerning 10-fold cross validation (Figure 25, panel 2.B), maximum accuracy is still achieved by RFE-LNW with one gene, in complete agreement with the result derived by the ELOOCV procedure. Notice that the most frequently selected gene (80/100 frequency rate) is the one that was also selected by the ELOOCV process. This gene corresponds to M63391 (human desmin gene, complete cds), which belongs to the group of smooth muscle genes playing an important role in colon cancer [3]. The muscle index is a quantity used in [3] to reflect the muscle cell contents of a given sample. Most normal samples have higher muscle index than tumor samples. The specific gene has been reported in various studies as a very significant one in colon cancer, [20], [21], [24].

BC Results

In this data set we notice significant differences in the performance of the two schemes, especially for small numbers of markers as demonstrated in Figure 25 (Panel 3). Concerning the ELOO performance of RFE-LNW (Figure 25, panel 3.A), its maximum performance of 81% is achieved at 64 genes. We point out that those 64

genes may differ slightly from iteration to iteration (for different left out sample), so that for the specific data set we may end up with 78 different sets of 64 genes (since the training set consists of 78 samples). This ensemble of gene sets is further analyzed in the next section. Overall, the RFE-LNW achieves an average success rate of 81% with 64 genes, showing better generalization ability than that of RFE-SVM achieving 67% average accuracy with 256 genes. This improved performance of RFE-LNW is consistently preserved through external cross validation from 64 to 2 genes. The same qualitative result is achieved in 10-fold cross validation (Figure 25, panel 3.B), where we notice that maximum performance is still achieved by RFE-LNW on the 64 genes cut off point in agreement with the results of the ELOO procedure. We again notice that the 10-fold cross validation procedure produces smoother learning curves due to the lower variance of the produced result.

Wrapper methods have been accused for introducing a large amount of bias through their learning process [6]. Giving an appropriate computational set up, we demonstrate in APPENDIX I that the bias introduced through RFE-LNW, approaches that introduced by the filter method, while in APPENDIX II we provide a benchmark evolution of the gene selection problem in the applied BC data set.

3.6.3 Fusion of Selected Genes

Taking into account the promising results derived by the proposed methodology on the breast cancer data set, we proceed by examining the union of the different 78 and 100 sets of 64 genes derived through the ELOO and the 10-fold procedures, respectively. This union results into two different supersets of markers; the first one contains 200 different genes, while the second one consists of 507 genes. Notice that the superset derived through the ELOO procedures is much smaller than that derived through the 10-fold cross validation, indicating a great amount of gene overlap. In

fact, half of the 64 genes (50%) overlap with each other in all runs (a frequency of 78), while on the other hand only 6 genes overlap in all runs (a frequency of 100) of the 10-folld cross validation procedure. This result further emphasizes the role of the ELOO procedure as a stability measure, according to the discussion preceded in section 3.5. Another interesting but also logical result that was verified is that the 200 genes derived through the ELOO process are a subset of the 507 genes derived through the 10-fold cross validation procedure.

Gene Seq. ID	Gene Accession	Gene Seq. ID	Gene Accession
659	NM_001667	10538	NM_004911
719	NM_001685	10643	NM_020974
831	Contig43684	12259	NM_006544
838	Contig13548_RC	12416	U90904
1505	AF148505	12553	Contig64861_RC
1623	Contig17273_RC	12572	AF055033
1626	Contig35229_RC	13270	Contig5456_RC
2819	NM_003376	13343	AL355708
3224	NM_020120	13490	AK000365
3232	NM_020123	13800	Contig47544_RC
3742	Contig44713	13917	Contig65439
3786	NM_002779	14762	AF160213
3851	Contig6238_RC	15157	Contig11065_RC
4952	NM_005007	15813	Contig50013_RC
4966	AB018337	15874	Contig31312_RC
5293	NM_003607	16474	Contig14882_RC
6463	NM_012479	17778	NM_015984
7012	NM_005219	18425	Contig63102_RC
7126	NM_005243	19549	NM_000207
7797	NM_013306	19840	NM_000272
8071	NM_013360	19928	NM_000291
8162	NM_013376	20891	BE739817_RC
8910	NM_013438	23161	Contig41716_RC
8976	NM_004701	23744	NM_000797
8982	NM_004703	23889	Contig22253_RC
9235	Contig33814_RC	23978	NM_002208
9646	NM_006260	24016	NM_002224
9874	NM_005594	24169	NM_002268
10325	NM_007057		

Table 9: Sequence numbers and accessions of the 57-gene signature.

A linear neuron similar to that presented in section 3.3 was trained on the available 78 training samples, each described through the set of 200 ELOOCV genes and tested on the provided independent test set of 19 samples, described also by the same set of

genes. We emphasize that we focused on the 200 genes set since it is actually the intersection of the two supersets derived by 10-fold and ELOOCV procedures. One gene was eliminated per iteration and the set of surviving genes that gave the highest classification accuracy on the independent test set was selected as the final set of genes (signature). This procedure resulted at a set of 57 gene signature with an accuracy of 89.47% on the independent test set (missing only two samples), Van't Veer et al. derived a set of 70 markers with the same accuracy in [7]. The systematic gene names of the 57 markers are reported in Table 9

The learning procedure that was used at this stage is the variant learning rate scheme given below introduced in section 2.3.5

$$w_i(t+1) = w_i(t) + |d-y| \cdot sign\left(e \cdot f'(u)\right) \cdot sign(g_i) \cdot K_i \qquad (2.38)$$

The term $|d-y|$ plays the role of a variant learning rate, which is decreasing as y approaches the goal d. In other words, we take larger step towards the target when being away from it, but we slow down the update process when approaching it. With this learning scheme, the only parameter that needs to be tuned is the number of epochs, which was set to 500. The weight factor K_i for this set of experimental steps was adaptively set through equation (2.33).

3.6.4 Expression Profile Analysis of Selected Genes

In this experimental step we study the clustering ability of the derived set of 57 genes in three data sets available for BC: a) the training data set of the 78 samples, b) the independent test set of the 19 samples and c) the cohort of the 234 new cases published by Van De Vijver et al. in [3]. Complete-linkage hierarchical clustering with Pearson distance measure was used on all three data sets.

Clustering on Training Set

Clustering of the training set revealed two basic sets of clusters as depicted in Figure 26.A. Using the follow up times published in [1], these two clusters yielded the Kaplan-Meir survival curve illustrated in Figure 26.B, indicating a significant difference on the survival prediction of the two clusters, with the good prognosis group being the one on the left of Figure 26.A. As a conclusion, we state that the profile of the selected markers through our methodology is a sufficient survival predictor on the training set. In the next experimental step we study the behavior of the selected genes on the independent test set of the 19 samples.

Clustering on Test Set

Hierarchical clustering on the independent test set derives, as in the case of the training set, two basic clusters depicted in Figure 26.C. These two clusters correspond to good and poor prognosis groups as revealed by the Kaplan-Meier survival analysis conducted (Figure 26.D), resulting also into significant difference on survival prediction.

Clustering on 234 New Samples

We also tested the expression profile of the selected 57 markers on the new cohort of 234 samples published by Van De Vijver et al. in [3] . The clustering result on this new set of samples is illustrated in Figure 26.E. Two basic clustering groups were discovered corresponding, as in the previous two cases, to the two prognostic groups of good (green cluster) and poor (red cluster) clinical outcome. This fact is verified by the survival analysis depicted in Figure 26.F. The Kaplan-Meier curve derived in this experimental step is comparable to that of Van De Vijver et al. in [3], which further supports the valididty of the RFE-LNW method.Concluding this set of experimental step we emphasize that the proposed methodology manages to select a gene signature

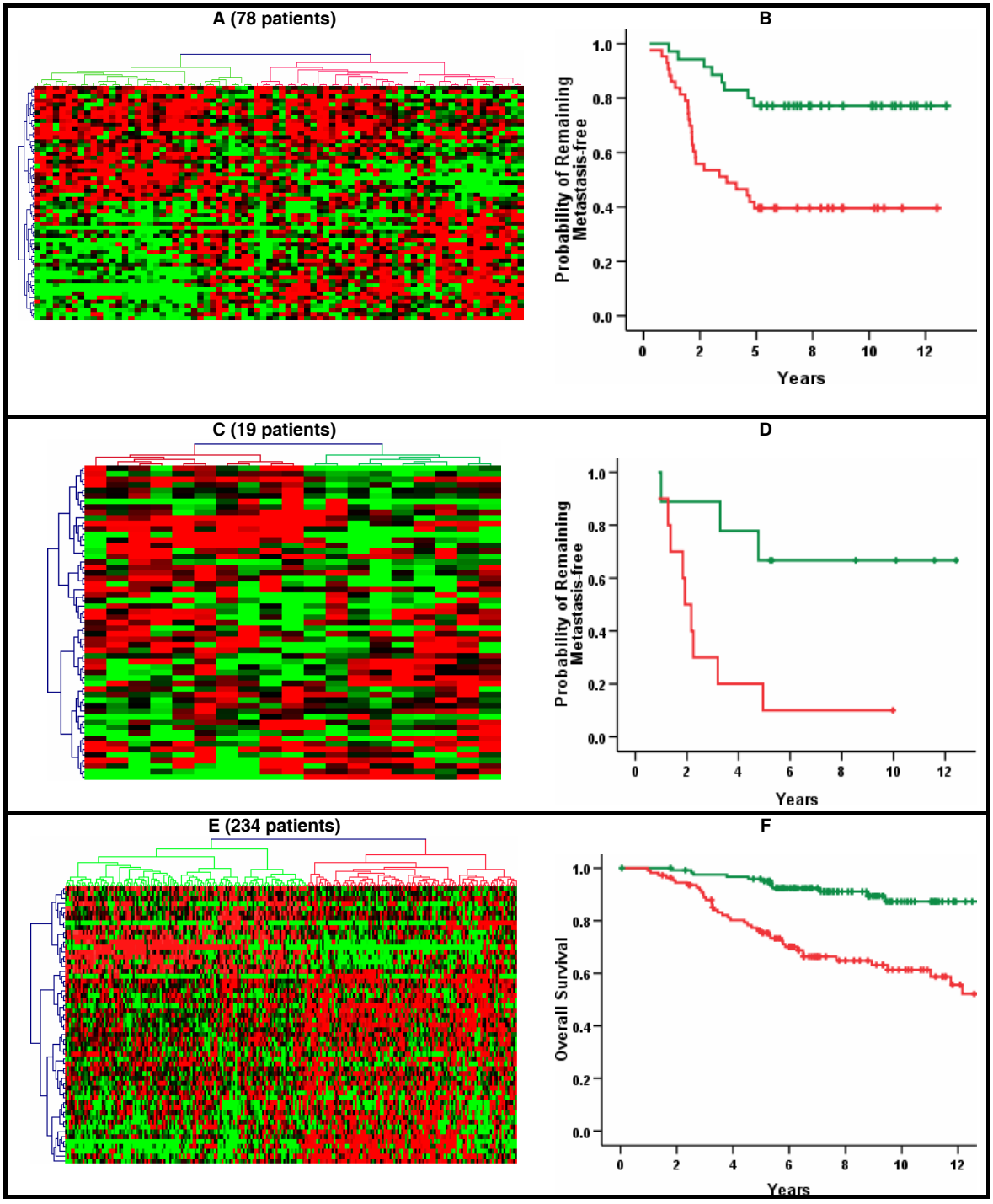

Figure 26: Clustering results and Kaplan-Meier curves of the 57-gene signature on the training set (A, B); on the independent test set (C, D) and on the 234 new patients (E, F). The derived signature serves the purpose of a sufficient survival predictor in all tested cases.

which, besides its high classification accuracy on the training as well as on the independent test set of 19 samples, enables the discovery of two well distinguished prognostic groups. This was verified both on the training and on the independent test set.

The fact that these markers could be used as successful survival predictors on the totally independent test set of 234 new is an encouraging result and demonstrates that the RFE-LNW complies well with clinical outcome, enhancing the conclusion that it is searching towards a meaningful biological path. We emphasize at this stage that no preprocessing step was applied to reduce the initial number of genes as was done in [1]. By employing such preprocessing we believe that the produced result could be further refined.

For all conducted experimental steps the tested methodologies were implemented on MatLab platform. For the SVMs implementation, the osu-svm MatLab toolbox [22] was used. The MEV Ver. 4.0 [23] was employed for gene visualization and clustering results, while the SPSS statistical software was used for the Kaplan-Meier survival analysis.

3.7 Study of Bias

Wrapper methods have been criticized for introducing a selection bias to the feature selection process. Ambroise and McLachlan in [6] address this issue and point out that wrapper methods introduce a great amount of bias if the test set is also used in the feature selection process. This bias, however, is corrected when the test set is excluded from the feature selection process. In this study three basic steps of accuracy measuring experiments were taken, namely ILOO, ELOO and 10-fold cross validations. As we pointed out in the introduction we are aware of the fact that ILOOCV introduces bias to the feature selection process, but we used it only as a test

to measure the learning ability of the underlined methodologies *on the training set* and not as an actual measure for the final feature selection process. For the marker selection process instead, we used external evaluation procedures (ELOO and 10-fold) in which case the test set is totally unknown to the feature selection process and thus any introduced bias is corrected. We address the above issues by conducting a special type of experiment similar to that presented in [6] with the aim of demonstrating the bias introduced by the various performed tests. We point out that for the specific experiment both RFE-LNW and RFE-SVM are used with the parameters of ELOO and 10-fold cross validation processes, since the final genes

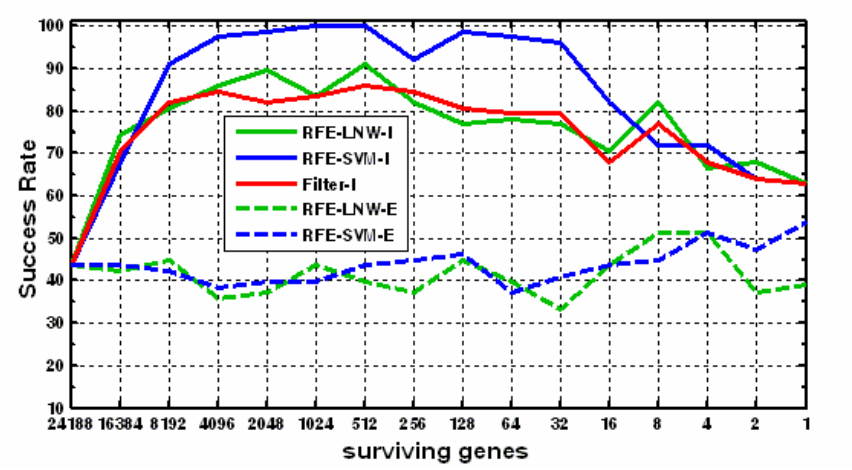

Figure 27: Bias variance of the tested methodologies. The RFE-LNW in the internal evaluation process approaches the bias of the Filter method, while both underlined methodologies correct for the selection bias when an external leave one out evaluation procedure is considered.

ignature was derived through those two procedures. For this experiment we focused on the BC data set, where each sample is randomly selected and assigned also at random a positive or a negative label. This process results to a totally random set, on which a totally unbiased feature selection process will give the accuracy of the random guess (approximately 50%). Then our tested methodologies were applied to this set using both ILOO and ELOO valuation criteria.

The results of the experiment are depicted in Figure 27, where we point out some interesting observations. First of all we observe that indeed both tested methods

introduce a large amount of bias in the case of the ILOO evaluation process, confirming the result of Ambroise and McLachlan. However, the bias introduced by RFE-LNW (green line) is significantly less, than that introduced by RFE-SVM (blue line) and almost approaches the bias introduced by the filter method (red line) that uses a variation of Fisher's ratio as the feature ranking criterion. On the other hand, both methods correct for the selection bias when an external evaluation procedure, such as ELOO is used (green and blue dashed lines), since their performance approaches around 50%, as expected.

As a concluding remark we emphasize two basic points: a) the proposed methodology, even though it falls into the category of the wrapper methods, introduces a bias that approaches that of the filter method and b) the evaluation criteria that were used as means of selecting the final set of markers genes correct for any selection bias.

3.8 Benchmark Comparison of Results in Breast Cancer

The fact that the strongest predictors for metastasis in breast cancer such as lymph node and histological grade fail to classify accurately breast tumors, lead the research community to DNA microarray analysis in order to identify a gene expression signature strongly predictive of the clinical outcome. One of the first studies that address this issue and has been considered as a benchmark is that of Van't Veer et al. in [1]. A signature of 70 genes is shown to be able to distinguish between poor and good prognosis patients with a success rate of 89.5% on an independent test set of 19 samples, while it gives a classification accuracy of 83% on the initial training set of 78 samples. Even though the study has received criticism for not cross validating the feature selection step [16] and thus it gives overoptimistic estimates, the 70-gene signature has nonetheless shown success in the cross validation study of Van't De

Vijver et al., [8]. The latter demonstrated that the 70-gene signature profile could be used as a sufficient survival predictor, outperforming classical clinical predictors, applied on a test cohort consisting of 234 new patients.

Following this result other studies were reported from which we selectively refer to [25] and [26]. In the first one, ridge regression succeeded to classify the independent test set with a success rate of 89.5% by selecting only 8 genes as markers. This result however, is not cross validated through a repetitive external procedure and is not tested on the cohort of the 234 samples. In the second study the authors derived a 44 gene signature achieving the same success rate on the independent test set. The result is cross validated through a repetitive 10-fold procedure; while the derived signature was tested using the 234 new samples producing also comparable Kaplan-Meier survival analysis as with [8]. Concluding we may state that the result derived through RFE-LNW is comparable to benchmark results, since the derived 57 gene signature gives a classification performance of 89.5% on the independent test set, while it produces comparable survival curve as with[8].

3.9 Discussion and Conclusions

In this work we study the performance of two linear approaches, namely a linear Support Vector Machine and a Linear Neuron, embedded within the Recursive Feature Elimination approach for gene selection. Three series of computational experiments were conducted. In the first one the well established RFE-SVM method scheme was compared with the proposed RFE-LNW approach using appropriate measures of quality and accuracy. Results have shown that the RFE-LNW method derives better quality clusters of selected genes than the RFE-SVM, while the ILOOCV performance of the two methods revealed a connection between cluster quality and success rate. In the second set of computational experiments, ELOO and

10-fold cross validation performances were measured on three different data sets. The RFE-LNW showed more stable performance on external leave one out and resulted at a set of 57 markers with 89.47% (two samples missed) on the independent test set in breast cancer. The third computational step showed that the selected markers could be used as survival predictors on three different data sets including 234 new samples.

The advantage introduced by the RFE-LNW is that it focuses and implicitly searches for differentially expressed genes, through the Fisher's metric embedded within its learning process. This idea actually allows a filter criterion to be applied in a wrapper manner and hence it hybridizes characteristics of both wrapper and filter methods. This hybridization leads to the positive side effect of bias reduction as demonstrated in APPENDIX I by an appropriate experimental setup. A disadvantage of RFE-LNW is that it requires a number of parameters to be fine tuned, such as the learning rate, the number of epochs and the feature ranking criterion to be used. Besides, since it is based on neural network theory it may be trapped into a local minimum, which is a known problem to the neural network community. On the other hand, RFE-SVM may require only one parameter to be fine tuned, but it focuses only on classification performance and hence neglects intrinsic data characteristics, such as the variation on the expression levels of the selected genes. This inadequacy of RFE-SVM was highlighted in the BC data set, where it was shown that it creates ill-defined clusters leading to poor generalization performance on the external evaluation tests.
Furthermore, the SVM classifier is founded on the support vectors which are actually the basic fundamental samples of learning and could be considered as class representatives. Hence, the existence of a few but well represented samples could be enough to build an effective and efficient classifier. Nevertheless, absence of such representative samples could lead to poor generalization and low quality performance.

In such situations we need to embed intrinsic data characteristics into some form of a "background knowledge" in order to assist the classification and as a consequence the feature selection task; such a "background knowledge" can be implanted as a form of Fisher's ratio (other metrics could be used as well) within the training procedure. We emphasize that our aim in this study is not to provide a method that would eventually replace already existing and successful approaches, such as RFE-SVM, but to show that the proposed methodology can provide a robust wrapper algorithmic scheme, especially when aiming implicitly at selecting differentially expressed genes.

Proceeding onto the next chapter we examine the biological significance of the 57-gene breast cancer signature and demonstrate that it points to useful biological knowledge, which in turn could be use to further refine derived results.

References

[1] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, Gene selection for cancer classification using Support vector machines, *machine learning*, 46 (2002) 389-422.

[2] R. T. Golub, K. D. Slonim, P. Tamayo, C. Huard, et al., Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, *Science*, 286 (1999) 531-536.

[3] U. Alon, N. Barkai, D. Notterman, S. Ybarra, et al., Broad patterns of gene expression revealed by clustering analysis of tumor and normal cancer tissues proposed by oligonucleotide arrays, *PNAS*, 96 (1999),6745-6750.

[4] S. G. Baker and B. S. Kramer, Identifying genes that contribute more to good classification in microarrays, *BMC Bioinformatics*, 7 (2006), 407.

[5] M. E. Blazadonakis, M. Zervakis, Wrapper Filter Criteria Via Linear Neuron and Kernel Approaches, *Computers in Biology and Medicine*, To appear.

[6] C. Ambroise and G. McLachlan , Selection bias in gene extraction on the basis of microarray gene-expression data, *PNAS*, 99 (2002), 6562-6566.

[7] L. J. Van't Veer, H. Dai, M. J. Van de Vijver, Y. D. He, et al., Gene expression profiling predicts clinical outcome of breast cancer, *Letters to Nature* 415 (2002) 530-536.

[8] M. J. Van De Vijver, Y. D. He, L. J. Van't Veer, et al., A gene expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine,* 347(2002), 1999-2009.

[9] A. Alizadeh, M. Eisen, RE. Davis, Ma C., I. Lossos, A. Rosenwal, J. Boldrick, H. Sabet, T. Tran, X. Yu et al., Distinct subtypes of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature* 403 (2000) 503-511.

[10] M. Riedmiler and H. Braun, A direct adoptive method for faster backpropagation learning: the RPROP algorithm, *IEEE International Conference on Neural Networks* (ICNN 1993*)* 586-591.

[11] F. Azuaje, A cluster validity frame work for genome expression data, *Bioinformatics*, 18 (2002) 319-320.

[12] D. L. Davies and D. W. Bouldin, A cluster separation measure, *IEEE Trans. Patt. Anal. Machine Intell*. *PAMI*, 1 (1979) 224–227.

[13] S. Bandyopadhyay and U. Maulik, Nonparametric Genetic Clustering: Comparison of Validity Indicies, *IEEE transactions on systems, man, and cybernetics* 31 (2001) 120-125

[14] J. Wang, J. Delabie, H. Aashein, E. Smeland and O. Myklebost, Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study, *BMC Bioinformatics* (2002) 3:36.

[15] J. Vesanto, and E. Alhoniemi, Clustering of the Self Organizing Map, *IEEE transactions on neural networks*, 11 (2000) 586-600.

[16] R. Simon, M. D. Radmacher, K. Dobbin, L. M. McShane, Pitfalls in the use of DNA Microarray Data for Diagnostic and Prognostic Classification, *Journal of the National Cancer Institute* 95 **(**2003) 14–18.

[17] A. Little and D. Rubin, Statistical Analysis with Missing Data *(Wiley Series in Probability and Mathematical Statistics*, (1987).

[18] H. Jiang, C. Gomez-Manzano, H. Aoki, M. M. Alonso, S. Kondo, F. McCormick, J. Xu, Y. Kondo, B. N. Bekele, H. Colman, F. F. Lang and J. Fueyo, Examination of the Therapeutic Potentional of Delta-24-RGD in Brain Tumor Stem Cells: Role of Autophagic Cell Death, *Journal of National Cancer Institute* 99 (2007), 1410-1414.

[19] S. Armstrong, J. Staunton, L. Silverman, R. Pieters, et al., MLL translocations specify a distinct gene expression profile that distinguishes a unique leukaemia, *nature genetics*, 30 (2002), 41-47.

[20] K-A. Do, G. J. McLachlan, R. Bean, and S. Wen, Application of Gene Shaving and Mixture Models to Cluster Microarray Gene Expression Data, *Systems Biology Special Issue* 2 (2007) 25-43.

[21] X. Li, S. Rao, Y. Wang and B. Gong, Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling, *Nucleid Acids Research* 32 (2004) 2685-2694.

[22] http://sourceforge.net/projects/svm/

[23] http://www.tm4.org/mev.html

[24] G. J. McLachlan, R. W. Bean and L. Ben-Tovin Jones, A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays, *Bioinformatics* 22 (2006) 1608-1615.

[25] F. Li and Y. Yang, Analysis of recursive gene selection approaches from microarray data, *Bioinformatics* 21 (2005) 3741-3747.

[26] R. Shen, D. Ghosh, A. Chinnaiyan and Z. Meng, Eigengene-based linear discriminant model for tumor classification using gene expression microarray data, *Bioinformatics* 22 (2006) 2635-2642

CHAPTER 4

Revealing Significant Biological Knowledge via Gene Ontologies and Pathways

4.1 Abstract

Objective: Our objective in this study is to examine the biological significance of the 57-gene signature derived in chapter 3, using published biological knowledge on gene functionalities and biological processes.

Methods and Materials: Many scientific works in the field of bioinformatics and marker selection deal with the problem of deriving a gene signature using a variety of stochastic and/or pattern recognition approaches. Special focus is given on the statistical perspective of the derived result without paying much attention on its biological aspect and significance. We propose to use and rank significant biological knowledge, hidden behind the 57-gene signature, using hyper-geometric distribution probability, which in turn will be used for the generation of a new gene signature. We assess the statistical significance of the new derived signature using classical classification success rate measure, but we also assess the significance on the expression levels of marker genes through the global score (Q) statistic.

Results: Using gene ontologies and pathways we assess the biological 'meaning' of the 57-gene signature and show that it points to valid biological knowledge.

4.2 Introduction

Most prognostic gene signatures derived so far were evaluated by applying stringent statistical criteria on the performance of individual genes, regardless of their biological functions. A very recent study [1] has demonstrated that by taking into account previous biological knowledge provided through gene ontologies and pathways, significant results could be derived revealing the biological "mechanisms"

hidden behind ER-positive and ER-negative breast cancer patients. A significant result that has also been derived in the same study is that even though there is little or no overlap between the various published prognostic gene signatures when individual genes are taken into account, the situation is reversed when considering the biology hidden behind them.

Inspired by the work presented in [1], we demonstrate that the 57-gene breast cancer signature derived in chapter 3 and published in [2] points to significant biological knowledge, which when considered as the starting point of a new gene selection procedure, produces a more significant result in terms of outcome prediction than the initial one, verifying the biological validity of the initial 57-gene signature. In order to associate genes in the signature with biological significance embedded in GOBPs and PWs, use the concept of hypergeometric distribution. Based on such a biological ranking we select a new signature. For the purpose of evaluating the significance of the derived results at the various stages of the procedure we use the global test in [3] to elaborate on the relation between gene signature and clinical outcome. The test is a score test which is used to assess a signature's significance; the higher the score the more significant the signature is. In essence, the test assesses the correlation of gene expressions in the signature, with the clinical outcome on the available samples. It is build on the fact that genes in a signature are differentially expressed across classes and that genes with similar expression patterns point to the same outcome.

4.3 Methods

In this section we provide some background information on a set of criteria that will be used on this study and concern: a) the criteria that will be used to assess the significance of the biological knowledge hidden behind our 57-gene signature, b) the

criteria that will be used to assess the statistical significance of the derived gene signature(s) on the available data set and c) the classification approach that is used to assess prediction accuracy. Essential background knowledge on gene ontologies and pathways is provided in section 1.4.

4.3.1 The Hyper-geometric Probability Distribution

The Hyper Geometric Distribution Probability is a discrete probability distribution that describes the number of successes in a sequence of N draws, corresponding to the N genes constituting a specific PW or GOBP, from a finite population of M total genes without replacement:

$$p = f\left(x \middle| M, K, N\right) = \frac{\binom{K}{x}\binom{M-K}{N-x}}{\binom{M}{N}} \tag{3.1}$$

where x is the number of genes in the signature belonging in the pathway, M is the total number of genes in data, K the number of the pathway genes that exist in the data and N is the number of genes in the examined gene signature. In other words, it measures the probability of drawing among M total genes, x of them, such that they belong to a specific pathway of K genes, given that those x genes are also part of the N gene signature under consideration. The lower the value of p the most significant the examined pathway is, since it has a low probability of being selected at random. Note that for simplicity purposes we use the term pathway to designate both the pathway and the GOBP concepts.

Elaborating now on equation (3.1), we underline some interesting aspects. Suppose that a pathway contains only one gene $\left(K = 1\right)$ and that gene $\left(x = 1\right)$ has been selected in a signature consisting of only that one gene $\left(N = 1\right)$.Then for large M we get a p

(HGDP) value approximately zero. This implies that the specific pathway is an important one, since the probability of selecting genes of that pathway at random is close to zero.Consider now the inverse situation, let the number of genes constituting a pathway equals to the total number of genes $(K = M)$, restricting actually our selection to only one pathway that would lead to a signature of N genes all selected from the same pathway $(x = K)$. Then, for such a situation we get a p value of one, implying that the specific pathway is not a significant one since it gives a high probability of being a random guess. Let us considering a more realistic situation and focus in our gene signature of 57-genes being selected from a set 24188. Consider, for instance, the signal transduction pathway, which contains 3788 genes (during the period of our research), 2115 of which exist in our 24188 gene set $(K = 2115)$, which is pointed to by 14 different genes $(x = 14)$ in our 57-gene signature $(N = 57)$. This case yieldsa p value of 2.26E-04, ranking the specific pathway among the most important ones. Furthermore, indicating that a pathway could be important even if it consist of a large number of genes (provided that it contributes a large proportion of genes in the final gene signature), the fact that about 25% (14/57) genes of our gene signature are part of the signal transduction GOBP makes the specific ontology very important.

Using such a metric we can assess the statistical significance of a pathway as to whether its selection in the final gene signature is a random result or not. Thus, ranking then the pathways according to their corresponding p-values leads to a ranked list of pathway significance.

4.3.2 The Global Test

The global test [3] elaborates on the connection between gene expressions and clinical outcome. Suppose we are given gene expression measurements of n samples and m genes and we want to test if there is a close connection between gene expression patterns and clinical outcome. If a group of genes can be used to predict the clinical outcome, the gene expression patterns must differ for different clinical outcomes. Defining $X = \left[x_{ij} \right]$ as the $n \times m$ data matrix containing the m genes of interest and Y as the $n \times 1$ clinical outcome vector, we can model dependence of Y depends in X. In the model of [4] there is an intercept α, a length m vector of regression coefficients β and a function h (logit function) such that:

$$E\left(Y_i \middle| \beta\right) = h^{-1}\left(\alpha + \sum_{j=1}^{m} x_{ij}\beta_j \right) \tag{3.2}$$

Testing whether there is a predictive effect of the gene expression on the clinical outcome is equivalent to testing the hypothesis:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_m = 0 \tag{3.3}$$

It can be shown that if H_0 is true the test statistic Q is derived by:

$$Q = \frac{\left(Y - \mu\right)' R \left(Y - \mu\right)}{\mu_2} \tag{3.4}$$

where $R = (1/m) XX'$, $\mu = h^{-1}(\alpha)$ is the expectation of Y under H_0 and μ_2 is the second moment of Y under H_0. Hence, Q is a score test which can be interpreted in two alternative ways as follows:

$$Q = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{\mu_2} \left[X_i' \left(Y - \mu\right) \right]^2 \tag{3.5}$$

or as:

$$Q = \frac{1}{\mu_2} \sum_{i=1}^{n} \sum_{j=1}^{n} R_{ij} \left(Y_i - \mu\right)\left(Y_j - \mu\right) \tag{3.6}$$

Equation (3.5) indicates that genes with large variance have much more influence on the outcome of the test than genes with low variance. Since *R* is the gene covariance matrix and $(Y-\mu)(Y-\mu)'$ is the covariance matrix of the clinical outcomes, equation (3.6) focuses on samples and checks whether samples with similar outcomes share also similar gene expression patterns.

The global test is used in the experimental section of this chapter as a measure to assess the statistical significance on the expression of gene expressions within a gene signature on the outcome (label) of samples.

4.3.3 Nearest Centroid Classifier

For classification purposes we use the nearest centroid prediction rule [6]. Each patient is classified according to the distance between his/her signature and the two average profiles; the predicted class is the one closer to examined profile, by means of the Euclidean distance. Such a classifier can be formulated as follows:

$$f\left(x\right) = sign\left(\left(\boldsymbol{x} - \boldsymbol{c}\right) \cdot \boldsymbol{w}\right) \tag{3.7}$$

where

$$\boldsymbol{c} = \frac{\boldsymbol{c}_+ + \boldsymbol{c}_-}{2} \tag{3.8}$$

$$\boldsymbol{w} = \boldsymbol{c}_+ - \boldsymbol{c}_- \tag{3.9}$$

and $\boldsymbol{c}_+$, $\boldsymbol{c}_-$ are the centroids of the positive and negative classes respectively. This method is similar to that used by Michiels et al. [7]; we use Euclidean distance instead of Pearson correlation. It is also a close alternative to the classification method used by Van't Veer et al. [5] and Niijima et al. [8].

4.4 Experimental Setup

We consider the breast cancer data set published in [5] and examined extensively in chapters 2 and 3. We focus on the 57-gene signature derived in chapter 3 and published in [2], and assign to each one of its 57 genes the associated pathway (PW) [9] as well as the corresponding Gene Ontology Biological Process (GOBP) [10]. We compute the Hyper-Geometric Distribution Probability (HGDP) for each pathway and GOBP involved in the signature and rank GOBPs and PWs in ascending order according to their probability score, as a measure of assessing the randomness of selecting a specific GOBP/PW. In this way, we associate each individual gene in the signature with already known and published biological knowledge. In our study the most significant GOBP derived, which also provides a significant amount of genes is the "signal transduction" GOBP (GO:0007165) with HGDP of 2.26E-04, while the most significant pathway is the T-Cell receptor signalling pathway with HGDP of 3.24E-02.

4.4.1 Building a Gene Ontology and Pathway Signature

In our attempt to engage right from the initial step of marker selection the GOBPs and PWs pointed by the 57-gene signature, we collect all genes included in GOBPs with a HGDP of less than 0.2 and enrich them with all genes indicated by the most significant pathway, namely the T-Cell receptor. This collection results in an ensemble of 4197 genes, which is used as the basis of a new gene selection process. We apply the liner neuron methodology in the same manner as in section 3.3, keeping the same cut off scenario as the one used for the experiments in section 2.5.2 and is illustrated in more detail in Appendix I. For comparison purposes with the result of [5], where a 70-gene signature is derived, we select a gene signature of size closest to 70 that give the maximum classification accuracy on the independent test set. This

selection rule resulted in 71 genes with a classification accuracy of 84.21% on the independent set of 19 samples (3 missing cases). Nearest neighbor classification was applied as a mean of classifying unknown samples according to their gene expression profile. Thus, a new sample is assigned to the class with the closest (Euclidean distance) mean expression profile. To train the linear neuron, the same parameters as those used to derive the 57 gene signature were applied i.e., 500 epochs were used as long as the number of surviving genes was less than or equal to 1024, 300 epochs were used afterwards, while the learning rate was set to 10^{-4} .

4.4.2 Statistical Significance of the Derived Result

The global test is used as a measure to asses the statistical significance of the 71-gene signature derived in the previous section. We measure the test statistic Q which can be intuitively interpreted as the influence of genes and samples to the final result. Genes with large variance have much more influence on the outcome of the test statistic Q than genes with lower variance (equation (3.5)), which is a desirable property in microarray analysis[11]. At the same time, the test examines whether samples with similar gene-expression patterns are correlated with similar outcomes through equation (3.6). To assess the Q metric of the 71-gene signature, we applied a hierarchical clustering using Pearson Correlation Coefficient with complete linkage on the 234 new samples published by Van De Vijver et al. in [12], using those 71 genes that constitute the signature.

Hierarchical clustering revealed two main clusters with significant variation on the expression levels of genes (Figure 28), where rows correspond to genes and columns to patients. The Q statistic of the two main clusters returned a value of 488.3, while the same test yielded a value of 367.41 when measured on the initial 57 gene signature and 416.96 on the 70-gene expression profile derived by

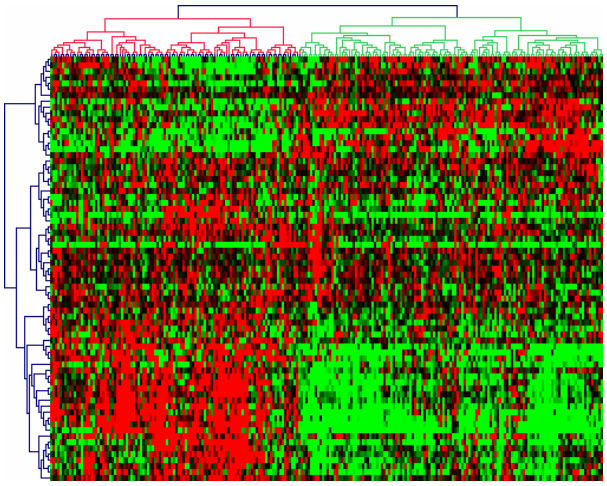

Figure 28: Hierarchical clustering applied on the 234 new patients (columns) of the 71-gene signatures (rows) reveals two main clusters with significance variation on the expression levels of selected genes.

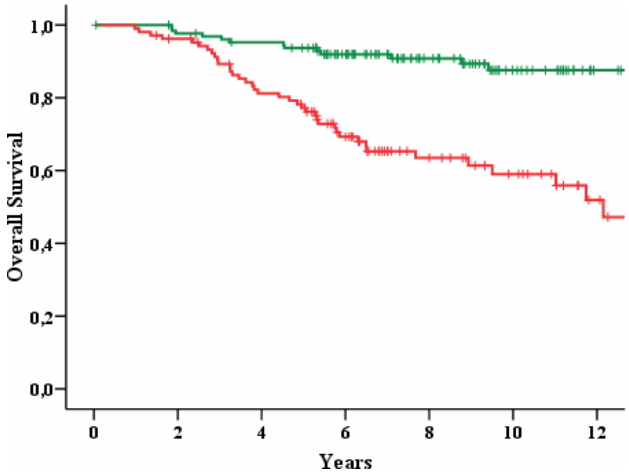

Figure 29: Kaplan-Meier survival analysis of clustering result using GOBP and PWs. It shows significant difference on survival prediction and reveals that the two clusters correspond to low (green) and high(red) risk groups.

Van't Veer et al. [5]. Hence, according to the test the 71-gene signature is more significant than the aforementioned signatures, so that the use of background biological knowledge in terms of GOPBs and PWs helps in improving the statistical significance of the derived result.

4.4.3 Clinical Prediction Outcome

Using the clustering result derived in the previous section we asses the survival prediction outcome of the two underlined clusters of patients using Kaplan-Meier analysis illustrated in Figure 29. The analysis revealed significant difference on the survival prediction, indicating that the two clusters actually correspond to poor (red) and good (green) prognostic patient groups.

4.4.4 Assessing Randomness of the Derived Result

One question of particular interest is how likely a produced result of approximately same or higher statistical significance would have been produced using any set of 4197 genes among the 24188 we are given initially? This question is a direct side effect of the random sample-prognosis correlation effect due to sparse coverage of the decision space. In a very high dimensionality space with a low coverage (few samples) the existence of more than one solution seems a logical consequence. Moreover, the existence of random gene combinations that may provide good correlation with the outcome is also possible. This issue has been addressed by Ein-Dor et al. [14], demonstrating the existence of multiple solutions in Van't Veer's data set. Even though our 71-gene signature has been derived through an evolution process taking into account additional stringent evaluation criteria [2] along with background biological knowledge provided through GOBPs and PWs, we proceed in an evaluation procedure similar to Ein-Dor's [14] and show how likely is to derive a 71-gene signature with approximately the same Q statistical significance starting from

any set of 4197 genes. To answer such a question we repeatedly and randomly selected 4197 genes from the initial given set of 24188. Among 80 runs, only in three cases the derived 71-gene signature was approximately as significant as the proposed one. This indicates that our starting point of the 4197 genes is unlikely to have happened by chance (probability of randomness 3/80), indicating also that the initial 57-gene signature, which was actually our information source, points indeed to significant biological knowledge, reducing substantially the uncertainty factor of the derived result.

4.5 Conclusions

In this chapter, we assess the problem of marker selection in a "reverse engineering" manner, closing a cycle in the process of selection. We usually start a marker selection procedure from a point of raw data and address it statistically as a dimensionality reduction paradigm. In this chapter, we take the derived result of such a process and proceed in a reverse way, in an attempt to improve its biological relevance by appropriately combining biological knowledge, while at the same time verifying the biological validity of our 57-gene signature. An added value of the proposed procedure is that it could be used to combine biological knowledge associated with different gene signatures, coming from different sources, aiming for further improvement of results.

Proceeding onto the next chapter, we combine biological knowledge derived from different gene signatures in an attempt to reveal alternative biological mechanisms which might be trigger breast cancer.

References

[1] J. X. Yu, A. M. Sieuwerts, Y. Zhang et al., "Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer", *BMC Cancer*, 2007, 7:182, http://www.biomedcentral.com/1471-2407/7/182.

[2] M. E. Blazadonakis and M. Zervakis, "The Linear Neuron as a Marker Selector and Clinical Predictor in Cancer Gene Analysis", *Computer Methods and Programs in Biomedicine,* ELSEVIER, Vol 91:1 (2008) pp 22-35.

[3] J. J. Goeman, S. A. Van de Geer, F. de Kort, H.C. Van Houwelingen. "A global test for groups of genes: testing association with a clinical outcome", *Bionformatics*, 2004, 20, pp. 93-99.

[4] P. McCullagh and J. A. Nelder, "Generalized Linear Models", *Chapman and Hall*, Boca Raton, USA, 1989.

[5] L. J. Van't Veer, H. Dai, M. J. Van de Vijver, Y. D. He, et al., "Gene expression profiling predicts clinical outcome of breast cancer", *Letters to Nature*, 2002, 415, pp. 530-536.

[6] R. Simon, Diagnostic and prognostic prediction rule using gene expression profiles in high dimensional microarray data, *British Journal of Cancer* 89 (2003) pp. 1599-1604.

[7] Michiels S, Koscielny S and Hill C., "Prediction of cancer outcome with microarrays: a multiple random validation strategy", *Lancet*, 365 (2005), pp. 488-492.

[8] S. Niijima, S. Kuhara, Recursive gene selection based on maximum margin criterion, *BMC-Bioinformatics,* (2006) 7:543.

[9] http://www.netpath.org

[10] http://www.geneontology.org/

[11] R. Simon., M. D. Radmache, K. Dobbin. and L. M. McShane, "Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification", *J. Natl Cancer Inst.*95 (2003), pp. 14-18.

[12] M. J. Van De Vijver, Y. D. He, L. J. Van't Veer, et al., "A gene expression signature as a predictor of survival in breast cancer", *The New England Journal of Medicine,* 347 (2002), pp. 1999-2009.

[13] K. Yu, C. H. Lee, P. H. Tan, G. S. Hong, S. B. Wee, C. Y. Wong and P. Tan. "A Molecular Signature of the Nottingham Prognostic Index in Breast Cancer", *Cancer Research*, 64 (2004), pp. 2962-2968.

[14] L. Ein-Dor, I. Kela, G. Getz, D. Givol and E. Domany, Outcome signature genes in breast cancer: is there a unique set?, *Bioinformatics*, 21 (2005), pp. 171-178.

CHAPTER 5

Integrating Biological Knowledge for Marker Gene Selection in Breast Cancer

5.1 Abstract

Objectives: It is well known by now that published prognostic gene signatures share very few genes or no genes at all in common giving hand to criticism and disputation. An optimistic answer to the problem is that the solution may not be unique, while a pessimistic view leads to nihilism and depreciation. We, as well as others, believe in the positive side of things and view the derived solutions as part of more global one, where each individual signature intersects only a small part of it, sharing only crumbs with each other. We focus on such knowledge intersections by revealing common biological mechanisms hidden behind two or more gene signatures, which alternatively is used as a meta-knowledge towards a more global and unified solution.

Furthermore, it is widely recognized that, one group's signature does not perform well on another's group's data, due to incompatibilities of microarray technologies and the experimental design. We assess this cross-platform aspect, showing that searching for a more global and unified solution as above also help in overcoming such incompatibilities.

Methods and Materials: Different genes may point at the same biological mechanisms or complementary protein processes, thus carrying indirect relationships. In this study we look behind a gene signature and reveal the biological knowledge it covers by means of biological processes and pathways. Based on the concept of HGPD (section 4.3.1) we derive the significant biological mechanisms, by means of GOBPs and PWs, as in chapter 4. We perform this analysis on three BC signatures. We then merge the derived PWs and GOBPs and perform gene selection from the resulting set

of genes. The new signature is evaluated by means of its statistical significance (Q metric) and its predictive ability (survival analysis). Furthermore, the new signature derived from one data set is applied on a different dataset (cross-platform validation).

Results: By riddling this biological knowledge we demonstrate that initial results are significantly improved statistically-wise, while integration of biological knowledge from different sources (gene signatures) shows promise in unfolding the effects of biological mechanisms opening the road to more global solutions which are blocked by different experimental platforms or designs.

5.2 Introduction

Microarray technology has become a valuable tool in the hands of experts classifying breast tumors according to their prognosis, type, ER status, or response to treatment. An open problem for tackling breast cancer is the most appropriate treatment protocol that a specific patient should follow. Even though chemotherapy or hormonal therapy reduces the risk of distant metastasis by approximately 1/3, 70-80% of the patients receiving adjuvant treatment would have survived without it [1]. Additionally, insurmountable inconsistencies in histological grading constrained the American Joint Committee on Cancer to exclude histological tumor grading from its staging criteria [3]. Hence, increasing the prognostic value through the use of stable and robust markers is more than a necessity, a direction towards which microarray technology has contributed a lot.

It is quite evident by now that simply acquiring microarray data is not enough; one must be able to extract meaningful information marking the need for collaboration among various scientific fields such as medicine, biology, statistics and computer science. Besides, “an understanding of both the biology and the computational methods is essential for tackling the associated data mining task without being

distracted the abundant fool's gold" [4]. We extend this aspects to: simply generating a statistically significant results is not enough; any result should also be evaluated in terms of its biological significance, which in any case is more important than its statistical one. To establish this position we refer to the study of Van't Veer et al. [1] which has received harsh criticism [5], [6] from a statistical point of view, when considering stringent statistical criteria. On the other hand, taking into account additional biological and medical criteria, FDA (Food and Drug Association) acted counter to statistical critisism and approved the result of Van't Veer et al. for commercial production [7]. It is the first clear product that profiles genetic activity, measuring the likelihood of tumor recurrence. It may help doctors in planning appropriate therapy for a patient when used in accordance with other clinical criteria and laboratory tests. On the same line, the European Organization for Research and Treatment of Cancer (EORTC) launched the MINDACT (Microarray In Node Negative Disease may avoid ChemoTherapy) project opening the road for randomized trials of the Van't Veer's result [8]. This is obviously a typical example of contradiction between statisticians and medical doctors, each one counting and evaluating results from a different and possibly diverse perspective.

In this study we demonstrate that by appropriately adopting biological knowledge, statistical results are significantly improved. It is evident that published gene signatures have few or probably no genes at all in common. Recent studies however proved that even if there is no significant gene overlap between two different gene signatures there might be significant overlap when the biology hidden behind is taken into account [9]. Building on these steps, we collect significant biological information in terms of Gene Ontology Biological Processes (GOBPs) and Pathways (PWs) hidden behind the 57-gene signature (S1) published in [11], the derivation of which is

discussed extensively in chapter 3. The 70-gene signature (S2) published by Van't Veer et al. in [1] and the 76-gene signature (S3) published by Wang et al. [10]. First we show that the three signatures indeed demonstrate significant biological overlap when GOBPs and PWs are considered, in complete accordance to [9]. Secondly, we demonstrate that statistical results are substantially improved when integrating the biological knowledge derived from S1 and S2. The derived results of such knowledge integration are evaluated on the 234 new cases published in [12], as well as on the 286 cases published in [10], which are assessed using a different microarray platform and experimental design. The proposed approach actually tackles two major problems in cancer gene selection: a) The minor or no overlap issue between different gene signatures at the gene level [13], and b) the cross platform evaluation of results by testing a predictor that is derived using a specific microarray platform and experimental design on another group's data derived with a different experimental platform and protocol [15].

5.3 Gene Signature Overlap

The concept of gene signature overlap can be seen at a four level of abstraction as follows.

1. Gene-level overlap, assesses the number of common genes between two signatures. This iissue has been addressed before showing minimal or no overlap among the various signatures published in breast cancer [13], [15].
2. Pathway-level overlap, assesses the common pathways that exist between two gene signatures [14], [15].
3. Significant pathways overlap as measured by means of the HGDP introduced in chapter 4 section 4.3.1. Locating the pathways indicated to by the genes of a gene signature, we can produce a rank order list of significant pathways in

ascending order according to HGDP. We can then define pathway overlap by looking at the top ranked pathways derived by two or more signatures under consideration [14]. In this study we assess a pathway as significant when its HGDP is less than 0.05.

4. We go one step further and define gene overlap within the significant pathways of two or more signatures. At this level we focus on the significant pathways of the signatures as defined above and assess the overlap of the ensemble of genes they define. In this form we essentially consider overlap of all genes that are functionally related with the ones involved in the original gene signature. Thus, we consider overlap among groups of genes, instead of genes themselves, at abstract level of functionality instead of expression.

In the following sections we focus on the 4th level overlap, but we also consider overlap at the remaining three levels.

5.4 Data Sets

Two publicly available breast cancer data sets are used to assess and validate results. The data set of Van't Veer et al. [1] described in more detail earlier in chapters 2 and 3. The 78 sample set of the data is used for training, the 19 additional samples were used for independent test set evaluation and the 234 new patients cohort is used as an extra validation set. In addition, the data published by Wang et al. [10] on BC is also employed for further validation purposes. It consists of 286 patients and 22283 genes; the whole data set is used for validation purposes, and for testing the cross platform prediction ability of a signature derived based on different experimental design and application platform. Based on this data set, we analyze our results according to disease recurrence. Hence, those patients for which no recurrence occurred are

characterized as the good prognosis group and belong to the negative class, while those for which a reoccurrence occurred belong to the positive class.

5.5 Associating Gene Signatures and Pathways

In this section we associate each one of the S1 and S2 signatures with significant GOBPs [19] and PWs [20] they are associated with, by assembling the GOBP and PW indicated by each of the genes in each signature. We rank in ascending order the associated GOBPs and PWs according to their p values (given by equation (3.1) and preserve only those with a p value of less than 0.05. Even though there exists minor overlap when absolute gene identifiers are taken into account between the two signatures (only 5 genes in common), the situation is significantly reversed when we focus on the biological knowledge provided through the GOBPs and PWs associated with the underlined signatures. Thus, by assembling the genes of GOBPs and PWs we end up with a collection of 5899 genes for the S1 signature, and 5860 genes for the S2 signature. The interesting result is that the two signatures show a significant overlap of 52% (4031 genes in common). This demonstrates that the two signatures share significant biological knowledge overlap even though they reflect very small gene overlap. The GOBPs and PWs with associated information are given for S1 in Table 10, Table 11, and for S2 in Table 12, Table 13. We also point out that even though there is only one significant GOBP in common between the two signatures (3^{rd} level overlap, see section 5.3), i.e. the Signal Transduction (GO:0007165), this reflects a dominating overlap since it contributes 3788 genes in the final ensemble, while the remaining gene ontologies are contributing significantly less; note that one gene may contribute to more than one GOBPs and PWs. This fact highlights yet another aspect in biological knowledge overlap. Even though two signatures may seem to share an

GOBP - ID	GOBP - Name	GOBP - Num of Genes	GOBP - Genes in Data	Genes in Signature	HGPD
GO:0051084	de novo' posttranslational protein folding	10	6	2	0.000081
GO:0007165	signal transduction	3788	2115	14	0.000226
GO:0051649	establishment of cellular localization	1027	542	7	0.000236
GO:0006886	intracellular protein transport	556	314	5	0.000766
GO:0006573	valine metabolic process	2	1	1	0.002400
GO:0019859	thymine metabolic process	1	1	1	0.002400
GO:0030949	positive regulation of vascular endothelial growth	1	1	1	0.002400
GO:0051000	positive regulation of nitric-oxide synthase activity	1	1	1	0.002400
GO:0030036	actin cytoskeleton organization and biogenesis	216	135	3	0.003700
GO:0009966	regulation of signal transduction	543	293	4	0.004400
GO:0006101	citrate metabolic process	2	2	1	0.004700
GO:0045908	negative regulation of vasodilation	2	2	1	0.004700
GO:0045909	positive regulation of vasodilation	3	3	1	0.007000
GO:0007632	visual behavior	6	4	1	0.009400
GO:0031532	actin cytoskeleton reorganization	8	4	1	0.009400
GO:0045429	positive regulation of nitric oxide process	8	4	1	0.009400
GO:0007049	cell cycle	915	588	5	0.009800
GO:0009615	response to virus	98	65	2	0.009800
GO:0006091	generation of precursor metabolites and energy	602	387	4	0.010900
GO:0042177	negative regulation of protein catabolic process	7	5	1	0.011700
GO:0050715	positive regulation of cytokine secretion	14	5	1	0.011700
GO:0006006	glucose metabolic process	114	84	2	0.015800
GO:0050930	induction of positive chemotaxis	9	7	1	0.016300
GO:0007021	tubulin folding	8	8	1	0.018500
GO:0051056	regulation of small GTPase mediated signal transduction	215	92	2	0.018600
GO:0007051	spindle organization and biogenesis	22	9	1	0.020800
GO:0007159	leukocyte adhesion	11	9	1	0.020800

GOBP - ID	GOBP - Name	GOBP - Num of Genes	GOBP - Genes in Data	Genes in Signature	HGPD
GO:0007212	dopamine receptor signaling pathway	12	9	1	0.020800
GO:0006892	post-Golgi vesicle-mediated transport	26	10	1	0.023100
GO:0042994	cytoplasmic sequestering of transcription factor	12	10	1	0.023100
GO:0001570	vasculogenesis	23	11	1	0.025300
GO:0016310	phosphorylation	791	516	4	0.026000
GO:0000070	mitotic sister chromatid segregation	30	12	1	0.027600
GO:0050708	regulation of protein secretion	25	12	1	0.027600
GO:0006607	NLS-bearing substrate import into nucleus	14	14	1	0.032000
GO:0045664	regulation of neuron differentiation	25	15	1	0.034200
GO:0048167	regulation of synaptic plasticity	20	15	1	0.034200
GO:0046631	alpha-beta T cell activation	18	18	1	0.040800

Table 10: The 38 Significant GOBPs pointed to by the 57-Gene-Signature (S1).

Pathway	Genes in Pathway	Genes in Data	Genes in Signature	HGPD
Tcell	359	318	3	0.0324
Hedg	26	19	1	0.0429

Table 11: The 2 significant pathways pointed to by the 57-Gene-Signature (S1)

GOBP - ID	GOBP - Name	GOBP - Num of Genes	GOBP - Genes in Data	Genes in Signature	HGPD
GO:0009653	anatomical structure morphogenesis	1068	740	8	0.0010
GO:0006260	DNA replication	219	168	4	0.0013
GO:0043627	response to estrogen stimulus	25	20	2	0.0015
GO:0007165	signal transduction	3788	2115	14	0.0017
GO:0001501	skeletal development	235	182	4	0.0017
GO:0031333	negative regulation of protein complex assembly	1	1	1	0.0029
GO:0031274	positive regulation of pseudopodium formation	5	1	1	0.0029
GO:0000022	mitotic spindle elongation	2	1	1	0.0029
GO:0030198	extracellular matrix organization and biogenesis	50	39	2	0.0055
GO:0008065	establishment of blood-nerve barrier	2	2	1	0.0058
GO:0006271	DNA strand elongation during DNA replication	3	2	1	0.0058
GO:0009887	organ morphogenesis	390	290	4	0.0085
GO:0006562	proline catabolic process	4	3	1	0.0086
GO:0009113	purine base biosynthetic process	5	3	1	0.0086
GO:0015012	heparan sulfate proteoglycan biosynthetic process	10	6	1	0.0171
GO:0007586	digestion	97	76	2	0.0191
GO:0006024	glycosaminoglycan biosynthetic process	21	10	1	0.0282
GO:0030225	macrophage differentiation	11	10	1	0.0282
GO:0006890	retrograde vesicle-mediated transport, Golgi to ER	16	12	1	0.0337
GO:0006940	regulation of smooth muscle contraction	17	13	1	0.0364
GO:0001836	release of cytochrome c from mitochondria	11	14	1	0.0390
GO:0001837	epithelial to mesenchymal transition	12	14	1	0.0390
GO:0016525	negative regulation of angiogenesis	18	14	1	0.0390
GO:0006631	fatty acid metabolic process	185	123	2	0.0440

Table 12: The 24 significant GOBPs pointed to by the 70-Gene-Signature (S2)

Pathway	Genes in Pathway	Genes in Data	Genes in Signature	HGDP
Wnt	114	112	3	0.0039
TGF	1556	1182	9	0.0048
AR	603	468	5	0.0091
IL6	105	100	2	0.0310
IL4	307	267	3	0.0349

Table 13: Significant PWs pointed to by the 70-Gene-Signature (S2)

insignificant number of common pathways, this overlap may prove to be significant when the genes of the overlapped pathways are taken into account.

5.6 Building a Unified Pathway Signature

Focusing first on S1, the signature succeeded an 89.47% success rate on the independent test while it demonstrated a significant distinction on survival prediction [11]. Classification on the 234 new cases published in [12] yields 66% AUC, which induces 74% sensitivity (true positive) and 59% specificity (true negative), as illustrated in Table 14. Considering now the collection of genes derived by the appropriately refined biological knowledge, i.e., taking those genes contained in the 38 GOBPs and 2 PWs (Table 10, Table 11) but also exist in the data from the design

	AUC	SEN	SPE
S1	66%	74%	59%
S1-BK (70-GS)	69%	80%	59%
S2	69%	76%	62%
S2-BK (70-GS)	71%	78%	63%
S1S2-BK (69-GS)	73%	81%	64%

Table 14: Results evolution using appropriately refined biological knowledge pointed to by GOBPs and PWs.

of the experiment, we derived a collection of 3535 genes. By applying the filter selection method as described in section 2.3.3 equation (2.11), a 70-gene Biological Knowledge Signature (S1-BK) is extracted. Notice that there is a significant gene overlap (50 genes in common) between S1-BK and the 71-gene signature derived in chapter 4. This is despite the fact that the two signatures were derived using different

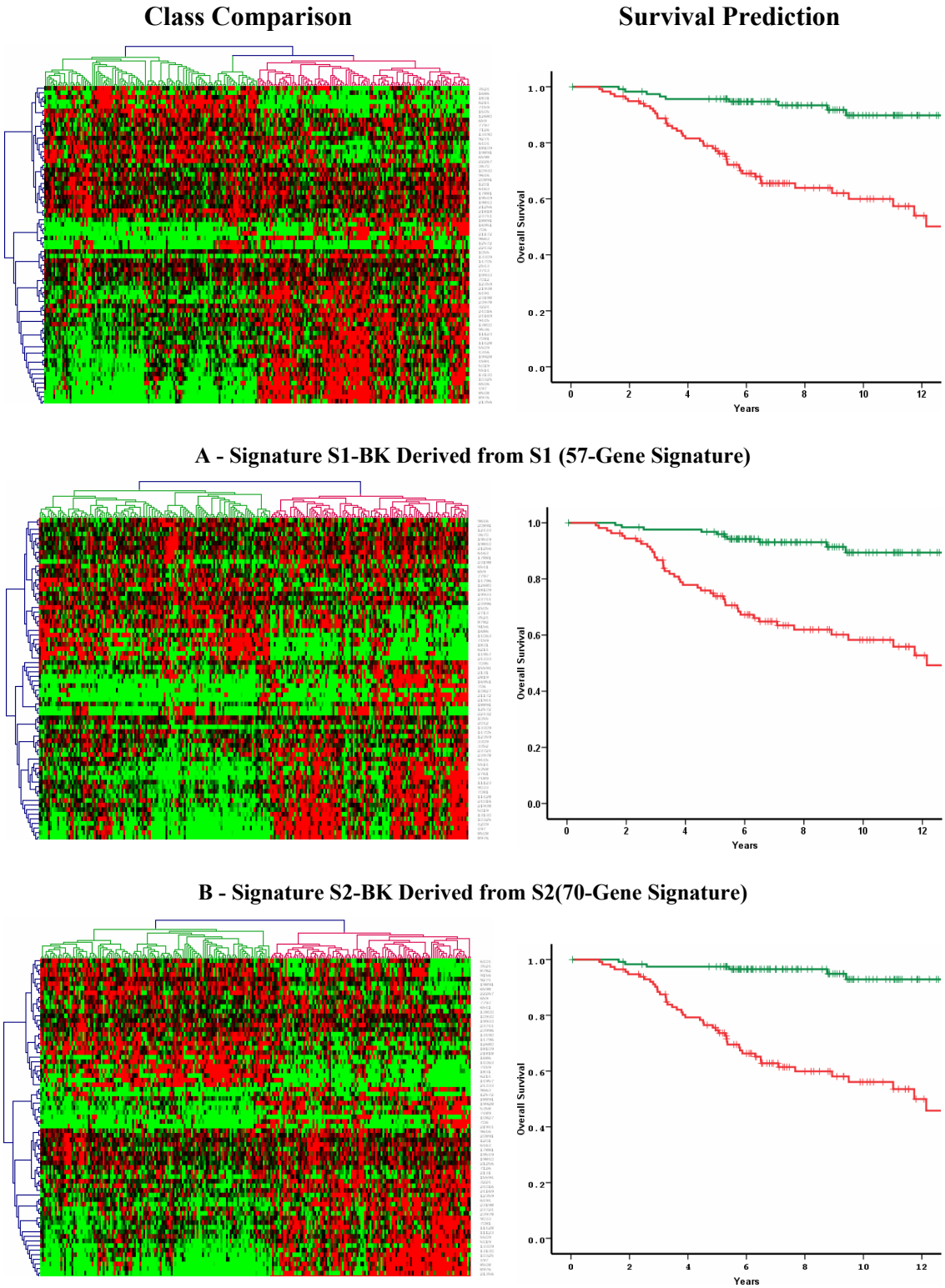

Figure 30: Derived pathway signatures with the corresponding survival prediction curves based on our 57 gene signature (panel A) and on the 70-gene Van't Veer's signature (panel B). In panel C the derived integrated signature is presented, along with the corresponding Kaplan Meier curve.

Gene ID	Assecion	Symbol	Gene ID	Assecion	Symbol
21256	NM_000419	ITGA2B	7489	NM_003875	GMPS
23724	NM_000788	DCK	12359	NM_014584	ERO1L
706	D25328	PFKP	9663	NM_006265	RAD21
14957	AJ224741	MATN3	9646	NM_006260	DNAJC3
10930	U17327	NOS1	8976	NM_004701	CCNB2
10325	NM_007057	ZWINT	7126	NM_005243	EWSR1
18891	NM_000127	EXT1	1831	NM_003239	TGFB3
5358	NM_020386	HRASLS	5509	NM_002914	RFC2
13130	NM_005915	MCM6	24169	NM_002268	KPNA4
14796	Contig34634_RC	GCN1L1	19933	NM_000296	PKD1
8782	NM_006117	PECI	19549	NM_000207	INS
11123	NM_007111	TFDP1	6494	NM_005192	CDKN3
15594	NM_006931	SLC2A3	18109	NM_000017	ACADS
497	NM_002358	MAD2L1	13800	Contig47544_RC	ATP5E
2131	X05610	COL4A2	9156	AF257175	PECI
6214	NM_012429	SEC14L2	3524	NM_004163	RAB27B
9274	NM_004798	KIF3B	8508	NM_003981	PRC1
19928	NM_000291	PGK1	23744	NM_000797	DRD4
1686	AF201951	MS4A7	12680	U82987	BBC3
22267	NM_001282	AP2B1	17881	D13540	PTPN11
23978	NM_002208	ITGAE	10827	NM_004994	MMP9
659	NM_001667	ARL2	24016	NM_002224	ITPR3
9033	NM_005444	RQCD1	6404	NM_005176	ATP5G2
21356	NM_001168	BIRC5	13490	AK000365	ACO2
12572	AF055033	IGFBP5	19891	NM_000286	PEX12
24333	NM_000849	GSTM3	6541	NM_003748	ALDH4A1
14063	NM_006763	BTG2	20891	BE739817_RC	IFNAR1
1201	NM_003163	STX1B	7459	NM_003862	FGF18
23996	NM_002217	ITIH3	13309	NM_006681	NMU
19840	NM_000272	NPHP1	6463	NM_012479	YWHAG
21818	AB023216	KIAA0999	23198	Contig25991	ECT2
7081	NM_004504	HRB	3224	NM_020120	UGCGL1
6598	NM_003766	BECN1	7797	NM_013306	SNX15
5019	NM_012310	KIF4A	21944	NM_001216	CA9
11428	NM_005721	ACTR3			

Table 15: The 69 genes of the S1S2-BK gene signature

selection methods, while significant GOBPs and PWs were assessed using alternative HGPD criteria. For the extraction of S1-BK, the initial 78 samples were used for training and the additional 19-samples were used for testing, onto which S1-BK achieved a success rate of 84.47%. Evaluating now the performance of S1-BK on the 234 samples cohort, we note a significant improvement over its predecessor S1. More specifically, the AUC measure is improved by 3 units (from 66% to 69%) while the sensitivity rate achieved an 80% success rate, improving substantially by 6 units, while the specificity rate on the other hand remained on the same level. We emphasize the increase in sensitivity, since the percentage of success rate in the true positive cases is very crucial to medical doctors; the cost of misclassifying a positive patient is much higher than the cost of misclassifying a negative one. In addition, high sensitive percentage enables doctors to decide more accurately on the therapeutic protocol, which is very important in cancer since patients may avoid unnecessary toxic treatment. Returning to S1-BK signature depicted in Figure 30 (panel A), we also perform a class comparison of the derived classification result on the 234 new cases [12], in order to reveal expression level differences of the gene signature, if such exist, between the two classes. The class comparison is performed through a labeled clustering. Thus, the class label, derived through the classification process, is given as input to the clustering algorithm in order to avoid inter-mingling of samples between the two classes and at the same time reveal differences in the expression levels of selected genes. Such a clustering designates significant expression level differences of marker genes between the two classes (green and red part of the tree) as demonstrated in Figure 30 (panel A); rows correspond to genes columns to patients. Additionally, this clustering yields Kaplan-Meier survival prediction curves distinguishing significantly the two prognostic groups between each other, with the good prognosis

(green curve), associated to the green cluster result and the poor prognosis (red curve) associated with the red cluster result.

Focusing on S2, the underlined gene signature achieves a success rate of 89.47% on the independent test set [1], while it gives 0.69 AUC translated to 76% sensitivity and 62% specificity on the 234 new patients data set [12]. Following the same scenario as in the case of S1 to exploit significant biological knowledge, i.e., the genes pointed by the top 24 GOBPs (Table 12) and the top 5 pathways (Table 13), we derive a collection of 3854 genes, which are again candidates for the derivation of a new biological knowledge signature. Applying the filter method as in the case of S1, the new derived signature consists of 70 genes that give an 89.47% success rate; we refer to this signature as S2-BK. When tested on the 234 cases, S2-BK improves its performance compared to its predecessor signature S2. More, specifically, AUC improves from 0.69 to 0.71, sensitivity from 0.76 to 0.78 and specificity from 0.62 to 0.63, as illustrated in Table 14. Even though performance increase is not as significant as in the case of S1, improvement is still substantial indicating that biological knowledge is indeed useful for signature refinement and performance prediction . The corresponding class comparison and Kaplan-Meier analysis are depicted in panel B of Figure 30, leading to similar conclusions as in the case of S1 signature.

Integrating now the biological knowledge reflected by both signatures results to a total ensemble of 4852 genes. Proceeding in a similar manner as before through the application of the filter method, a 69-gene signature is derived succeeding an 89.47% success rate on the independent test set of the 19 samples. This signature achieves an AUC rate of 0.73 with sensitivity and specificity rates of 81% and 64% respectively on the 234 new cases data set; we refer to this signature as S1S2-BK. Notice that by integrating the biological knowledge hidden behind the initial given gene signatures

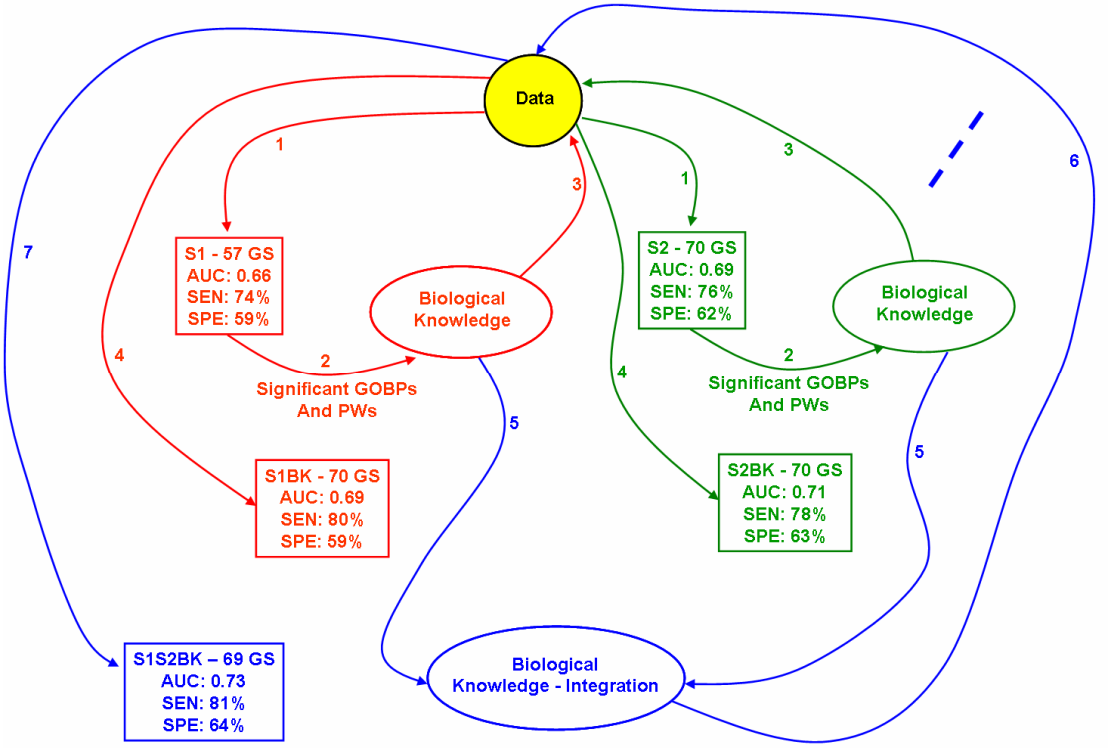

Figure 31: Biological knowledge evolution process followed to derive the S1S2-BK gene signature; red – green part represents independent focus on a single signature, blue part represents biological knowledge integration.

(S1 and S2), all measures are significantly increased compared to the ones derived by the initial S1 and S2 signatures.

A class comparison and Kaplan Meier study of S1S2- BK is presented in Figure 30 (panel C), indicating significant differentiation on the expression level of selected genes on the 234 cases cohort, while a clear difference on survival prediction curves corresponding to the two groups is also verified. We observe that the probability of survival for the good prognosis profile lies above 95% for a time interval of twelve years, which may increase doctor's confidence on the therapeutic decision. Analyzing the result of the classification performance, S1S2-BK classified correctly 44/54 positive cases and 115/180 negative ones. Focusing on the classification of negative patients we notice that 115 true negative patients are classified correctly, while 10 positive are classified as false negatives. This implies that a doctor may decide with a

92% certainty (115/125) on not recommending chemotherapy to a negative patient that does not really needs it. The 69 genes constituting the S1S2-BK signature are listed in Table 15.

To measure the statistical significance of the classification results on the derived gene signature, we use the global score (Q) test [18]. The higher the score of the test the more significant the underlined gene signature. The integrated gene signature is statistically more significant (Q = 451.98), while S1 gives a score of 367.41 and S2 a score of 416.96. Another interesting result that needs to be highlighted is, unlike their predecessors, the derived genes signatures that are extracted from biological knowledge share a significant amount of overlap. Specifically S1-BK and S2-BK share 45 genes in common (48% overlap), while S1S2-BK shares an average overlap of 75% with S1-BK and S2-BK, owing to the fact that both initial signatures share significant biological knowledge overlap. The process of deriving an integrated solution is graphically depicted in Figure 31 where each individual signature contributes to knowledge evolution (red – green iteration cycles). Dashed lines in the upper right corner indicate that the process may repeatedly evolve adopting signatures from different sources, which in turn contribute additional complementary knowledge.

In such an evolving scheme we also considered the 76-gene Wang's signature (S3) [10] and analyzed it the same manner as we did for S1 and S2. We found a 20% and a 15% biological knowledge overlap with S1 and S2 respectively. Furthermore, integration of the biological knowledge between the three signatures resulted in a decrease of performance indicating that the biological knowledge offered by S3 is not adding complementary biological knowledge towards a statistical

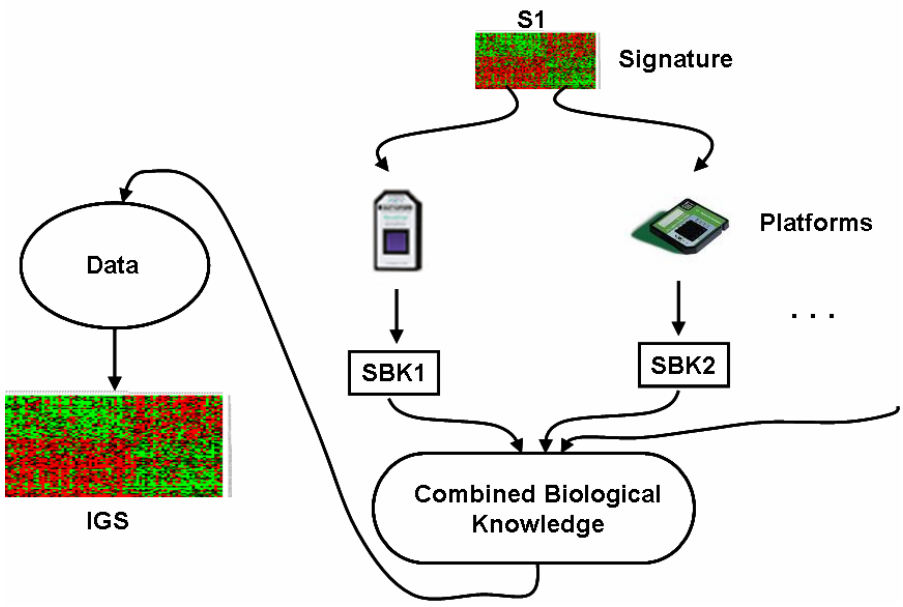

Figure 32: Knowledge integration based on different data sets.

improvement. This probably implies that the specific signature points to an alternative biological aspect, which when considered along with other signatures of similar perspective, could reveal complementary biological knowledge. This hypothesis could be further supported by the fact that the specific gene signature was not assessed explicitely through the expression profile of genes as S1 and S2, but rather through a Cox's regression model based on ER tumor status , imposing a different philosophy on prediction and probably on the biology of the outcome [10]. Another aspect playing important role is the incompatibilities between microarray platforms. Signatures S1 and S2 were derived using AGILENT microarray technology while S3 was derived using AFFYMETRIX chips. Chips are produced using genes from different PWs and GOBPs, so that different data sets may lead to different significant GOBPs and PWs assessing different aspects of biological relevance. Such differences may pave the way for an alternative integration approach, not assessed in this study, where biological significance is assessed on all available data sets and subsequent integration involves the total set of significant PWs and GOBPs. This integration approach is depicted in Figure 32, where significant biological knowledge (SBK) is

derived from different microarray platforms based on the same input signature. All significant PWs and GOBPs are then integrated resulting, resulting in in a pool of genes tested on the data, in order to derive a new integrated gene signature (IGS). In our case the different microarray platforms represent the Agilent Hu25K microarray that was used by the group of Van't Veer, and the Affymetrix U133a used by Wang's team. By appropriate application of HGPD (section 4.3.1) we can locate significant biological knowledge in terms of GOBPs and PWs derived from the genes in the signature and each gene chip. Focusing next on all genes associated with the joint set of significant PWs and GOBPs and returning to the initial dataset we derive a new integrated gene signature (IGS) by reapplying the feature selection method.

5.7 Cross Platform Validation of Integrated Signature

In this section we proceed by validating the S1S2-BK signature on the data set of 286 patients published by Wang et al. [10] towards a multi-center study that addresses problems with transferring results to different platforms, data set and experimentation designs. Before however proceeding in such a validation, we test the survival prediction capability of the initial signatures S1 and S2 on these 286 new patients. We searched for the genes of those signatures in Wang's data set. In the case of S1, we managed to locate 44 out of the 57 signature genes in the data while in the case of S2 we located 52 genes out of 70 in the signature. Applying SOM clustering [23] on these common genes we tried to discover the two prognostic patients groups and performed Kaplan-Meier survival analysis on the derived groups, in order to test the significance of the clustering result, and thus, the significance of the signatures. The results of these tests are depicted in Figure 33, where we notice that the underlined signatures are not able to differentiate significantly between the two prognostic groups. This outcome shows that results are tightly bounded by the array platforms

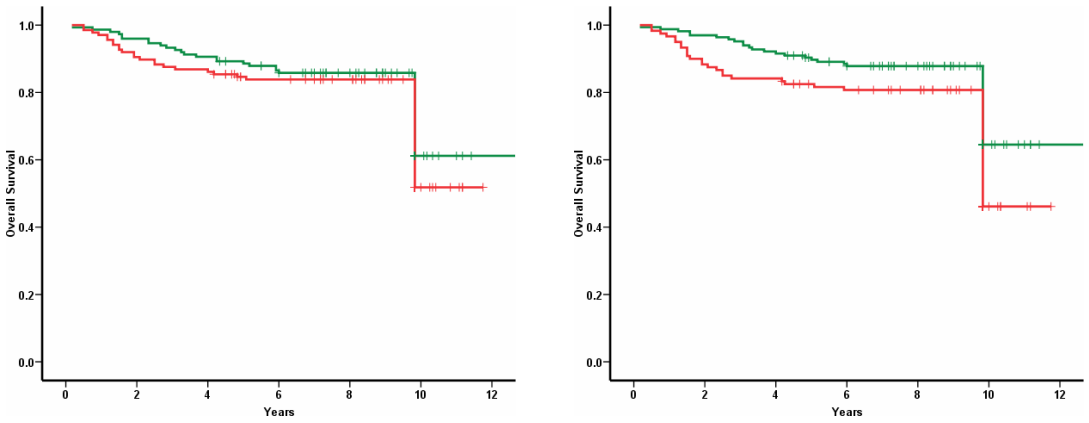

Figure 33: Kaplan Meier survival analysis of initial signatures (S1 – left, S2 – right) on the 286 cases of wang's data set.

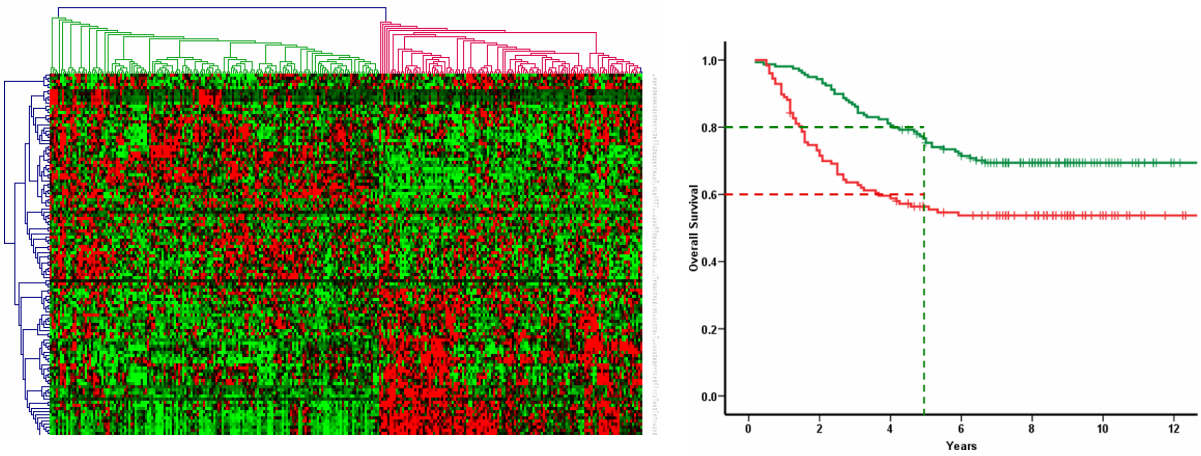

Figure 34: Class comparison (left) and Kaplan-Meier analysis (right) of the integrated signature on Wang's data set.

and the design of the experiment [22]; Van't Veer's group performed a double array experiment on Agilent chips, while Wang's group applied a single array experiment using Affymetrix chips. Performing the exact same experiment, but using the integrated gene signature (S1S2-BK) derived earlier, we managed to locate 116 genes in Wang's data; notice that some gene symbols appear more than once in Wang's data set. Examining the integrated signature in the same manner as before, SOM clustering revealed two groups of patients corresponding to the two prognostic groups.

The clustered result is depicted in Figure 34 (left panel); to preserve the two groups derived by SOM but better organize the classes according to gene expressions, we depict the labelled hierarchical clustering of the SOM groups. The two clusters reveal significant difference on the gene expression profiles. These two clusters map a good (green) and a poor (red) prognostic group as verified by the Kaplan-Meier survival analysis in Figure 34 (right panel). Notice the substantial improvement of the integrated gene signature, derived by appropriatly coupling of biological knowledge, over the initial predecessor signatures S1 and S2 (Figure 33) . Also notice that, even though the difference on the survival prediction may not be as significant as in the case of Van't Vijver data Figure 30, the difference at the crucial period of 5 years is still substantial with the good prognosis group achieving approximately an 80% probability of survival, while for the poor prognosis group the probability is below 60%.

5.8 Conclusions

In this chapter we addressed the problem of overlap between derived gene signatures as well as that of testing on datasets from different centers. We begin with the assumption that even though two solutions may seem to share minor or no overlap, they are probably part of a more global solution, each one focusing on a different,

complementary or overlapping part of the biological aspect of the problem. Experimental results demonstrated that when taking into account the biological knowledge behind gene signatures commonalities may be revealed, even though this may not be visible at gene level. Focusing on two such gene signatures, our signature consisting of 57 genes (S1) and the 70 gene Van't Veer's signature (S2), we showed that they share significant gene overlap (69%) at the biological level. Additionally, integration of the derived biological knowledge produces a statistically more significant signature, improving significantly the performance of the initial ones demonstrating reliable clinical prediction outcome on a dataset that was derived using different experimental design and microarray platform. This further enhances the value of these two gene signatures implying that the appropriately integrated signature does not entirely dependent on the platform or experimental design, but it is associated directly with probable biological origin of the disease itself.

Finally, we point out that the proposed approach is not competitive to different solutions in revealing or criticizing their weak or strong points but is rather taking advantage of the valuable crumbs of knowledge each one is contributing targeting at an evolution which will hopefully lead to a more global and unified solution. Until then we propose to view each individual result from a positive perspective and look for those valuable crumbs it probably has to offer.

References

[1] L. J. Van't Veer, H. Dai, M. J. Van de Vijver, Y. D. He, et al., Gene expression profiling predicts clinical outcome of breast cancer, *Letters to Nature* 415 (2002) 530-536.

[2] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, et al. Gene Expression Profiling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade to Improve Prognosis, *Journal of the National Cancer Institute*, 98 (2006) pp. 262-272.

[3] S. E. Singletary, C. Allred, P. Ashley, L. W. Basset, D. Berry, K. I. Bland, et al. Revision of the American Joint Committee on Cancer Staging System for Breast Cancer, *Journal of Clinical Oncology* 20 (2002) pp. 3628-36.

[4] D. K. Slonin, From patterns to pathways: gene expression data analysis comes of age, nature genetics supplement, 32 (2002) pp. 502-508.

[5] R. Simon, M. D. Radmacher, K. Dobbin, L. M. McShane, Pitfalls in the use of DNA Microarray Data for Diagnostic and Prognostic Classification, *Journal of the National Cancer Institute* 95 **(**2003) pp.14–18.

[6] Michiels S, Koscielny S and Hill C., "Prediction of cancer outcome with microarrays: a multiple random validation strategy", *Lancet*, 365 (2005), pp. 488-492.

[7] http://medgadget.com/archives/2007/02/mammaprint_a_br.html

[8] http://www.eortc.be/services/unit/mindact/MINDACT_websiteii.asp

[9] Yu J. X., Sieuwerts A. M., Zhang Y. et al., Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer, *BMC Cancer*, 7: 182 (2007), http://www.biomedcentral.com/1471-2407/7/182

[10] Wang Y., Klijn J. G. M., Zhang Y., Sieuerts A. M. et al., Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer 365 (2005) pp. 671-679.

[11] Blazadonakis M. E. and Zervakis M., The Linear Neuron as Marker Selector and Clinical Predictor, *Computer Methods and Programs in Biomedicine* 91:1 (2008) pp .22-35.

[12] M. J. Van De Vijver, Y. D. He, L. J. Van't Veer et al., A gene expression signature as a predictor of survival in breast cancer, *The New England Journal of Medicine,* 347(2002), 1999-2009.

[13] L. Ein-Dor, I. Kela, G. Getz, D. Givol and E. Domany, Outcome signature genes in breast cancer: is there a unique set?, *Bioinformatics*, 21 (2005), pp. 171-178.

[14] J. X. Yu, A. M. Sieuwerts, Y. Zhang et al., "Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer", *BMC Cancer*, 2007, 7:182, http://www.biomedcentral.com/1471-2407/7/182.

[15] R. S. Tuma, Multiple Gene Signatures Aim to Qualify Risk in Breast Cancer, *Journal of the National Cancer Institute*, 97 (2005), pp. 332.

[16] R. Simon, Diagnostic and prognostic prediction rule using gene expression profiles in high dimensional microarray data, *British Journal of Cancer* 89 (2003) pp. 1599-1604.

[17] S. Niijima, S. Kuhara, Recursive gene selection based on maximum margin criterion, *BMC-Bioinformatics,* (2006) 7:543.

[18] J. J. Goeman, S. A. Van de Geer, F. de Kort, H.C. Van Houwelingen. "A global test for groups of genes: testing association with a clinical outcome", *Bionformatics*, 2004, 20, pp. 93-99.

[19] http://www.geneontology.org/

[20] http://www.netpath.org

[21] J. J. Goeman, S. A. Van de Geer, F. de Kort, H.C. Van Houwelingen, A global test for groups of genes: testing association with a clinical outcome, *Bionformatics*, 20 (2004) pp. 93-99.

[22] R. S. Tuma, Multiple Gene Signatures Aim to Qualify Risk in Breast Cancer, *Journal of National Cancer Institute*, 97:5 (2005), pp. 332.

[23] T. Kohonen, Self Organizing Maps (Springer-Verlag Berlin, 2001).

Overall Conclusions and Open Research Directions

In this thesis we assess the problem of gene marker selection in various domains of interest. We focused on the dynamic approach of wrapper method and enrich it with an appropriately adopted filter criterion i.e. a variation of Fisher's ratio. This attempt resulted in an integration of both methods incorporating their advantages under a unified umbrella, producing more compact and distinct clusters of marker genes.

Integration was achieved by appropriately adapting the learning procedure of a linear neuron embedding within its learning process a Fisher's correlation criterion to assess intrinsic characteristics of data (RFE-LNW scheme). Even though a linear approach was applied through a linear neuron due to the characteristics of application domain, the method could be easily expanded to a multilayer perceptron and applied to non-linear problems. In addition, other correlation coefficients assessing intrinsic characteristics of data could be used in place of Fisher's ratio. In addition, focusing on the margin of separation defined by an SVM classifier and marked by the support vectors, we considered the critical boundary region which dominates the classification of the good and poor prognosis patients. Following a criticism on the SVM operation we proposed an alternative gene ranking criterion on support vectors which define yet a second integrated approach (RFE_FSVs).

Using such an integration approach, through the application of a linear neuron, along with an appropriate evaluation scheme we derived a promising breast cancer gene signature. Furthermore, we validated produced results using statistical criteria and compared it with other bench mark results in the field. In the last two chapters we also integrated the biological relevance of gene signatures in an attempt to validate them through the use of organized and published biological knowledge. This attempt, demonstrated that solutions appearing different at a first glance could actually show a

significant degree of biological overlap. This approach leads to a knowledge evolution process where biological information derived from different solutions could be integrated under a unified framework helping in unfolding the biological mechanisms hidden behind a disease. As a general conclusion we state that the present thesis leads to a methodological and biological integration of results.

It seems though that an integrated approach has benefits over filter and wrapper techniques. However, none of the techniques can be claimed to be superior of others for all pathology areas where the data distributions and decision spaces may change the ranking of algorithms. There is a need to set up conditions for using each algorithm dependent on problem specifications.

Further directions to our research emerge from important implications of Figure 31, where we search for those signatures which add complementary biological knowledge to the already accumulated, in a hope to understand and discover the biological mechanisms that trigger cancer. This knowledge, however, even though it has been associated with known biological processes does not carry biological evidence yet (except statistical evidence) that those processes are actually linked to the disease itself. This opens another research direction for validating the derived pathways outcome of cancer, leading to the design and implementation of appropriate biological experiments which may reveal possible new biomedical knowledge. Obviously this emphasizes the need of collaboration among different research fields and specialties for an effective tackling of the problem. On the other hand, aiming on specific and already known biological processes or pathways associated with the disease itself, is yet another search path worthwhile to address Additionally, despite research efforts, there is still the open problem, of associating histopathological level findings to a molecular level signature, which addresses a major challenge in cancer

histopathology, since it can overcome insurmountable inconsistencies in histological grading between institutions, assisting in avoiding unnecessary toxic treatments. Thus, integration of genomic with histopathological markers, along with their evaluation on the clinical benefits above and beyond established tests and indices aims, at the core of the problem and at the cellural mechanisms that trigger cancer.

APPENDIX I

Accuracy Evaluation Criteria

Accuracy evaluation is still an open question in the problem of marker selection. Various evaluation criteria have been used and reported in studies related to marker selection, each one serving its purpose within its context of use. In this section, we present an overview of the evaluation criteria that are used in our study with the aim of providing a systematic interpretation of their scope and effeciency within the field of marker selection. These measures assess the ability of the selected marker set to correctly separate the classes of interest. However, they do not reflect any clustering characteristics of the selected markers that could indicate gene expression differentiation among the classes of interest (intra class similarity or between class distance, similar to Fisher's measure). Given a prediction rule R, the following accuracy criteria are assessed:

Apparent Success Rate (ASR): It is the percentage of training features correctly classified by R. The use this metric must be with caution, since it reflects how well the given prediction rule has learned the training set without providing any insights into the generalization ability of the rule. It is also referred to as self test and is expected to be an over optimistic estimate of the algorithm's true prediction accuracy.

Leave One Out Success Rate (LOO-SR): For the computation of this metric one sample is left out of the training set. The prediction rule is then derived with the remaining $n-1$ samples and tested with the left out sample. The process is repeated until every sample is used as a test sample and finally the total mean accuracy is evaluated. In the problem of marker selection two different schemes of LOO success

rate can be assessed: The External (ELOO) and the Internal (ILOO) success rates [14]. In the ELOO, the removal of the left-out sample takes place before the selection of differentially expressed genes and application of the prediction rule. Alternatively, in the ILOO scheme the removal of the left-out sample takes place after the selection process but before the application of the prediction rule. In the ILOO scheme the entire training set is considered in the feature selection process and the LOO strategy is applied only in the evaluation phase. By these means, the ELOO evaluation methodology is an almost unbiased estimator, since the tested sample is totally unknown not only to the classifier, but to the entire selection process.

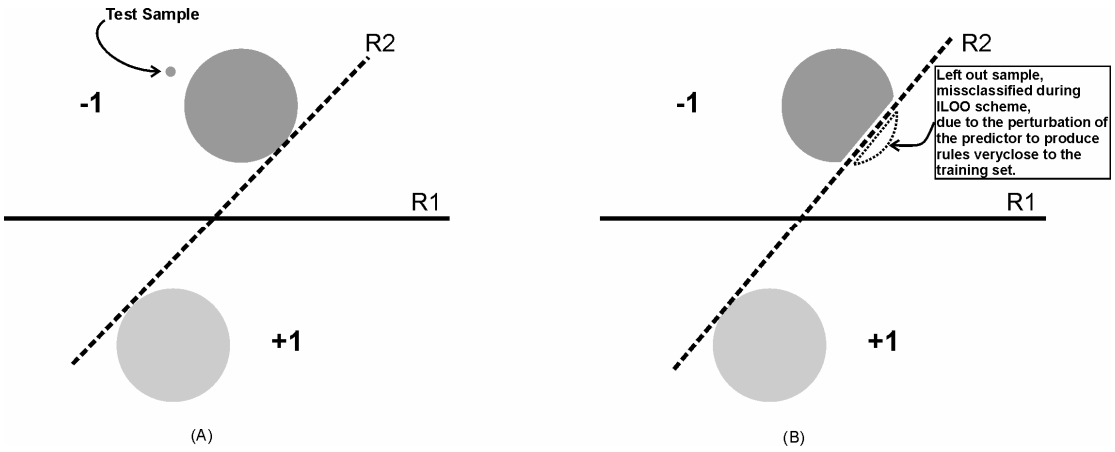

Figure 35: ILOO success rate measures the generalization ability of the prediction rule on the training set, which could reflect its generalization ability on an independent test set.

However, the use of different training sets at each iteration might lead to different sets of marker genes, where a post processing step is necessary to derive the optimal set of marker genes. On the other hand, ILOO evaluates a set of marker genes and a measure of how well a prediction rule can generalize on the training set. In other words it indicates how strict (close to the training set) or loose (far from the training set) the derived prediction rule is on its learning environment, i.e. the training set. To further illustrate the performance of ILOO we demonstrate the following scenario. Suppose we are given two different linear rules R1 and R2 produced by two different learning systems, which are used to classify a training set consisting of positive (+1)

and negative (-1) samples, as in Figure 35 (A). We are also given a test sample that is left out from the training procedure. Both derived rules classify correctly the training set (perfect ASR) as well as the left out sample. Nevertheless, R2 has a tendency to over learn the training set by being very close to the two classes (a strict rule), while R1 tries to maximize its distance between the two classes (a loose rule). Obviously R1 has an advantage over R2 which can be revealed by the ILOO scheme in Figure 35 (B). Indeed, since R2 has the tendency to be tightly bound to the border of the classes, when border line samples (white region of negative class in Figure 35 (B)), are left out and excluded from the training set, they will be misclassified in the testing phase, leading to lower ILOO measures for the R2 rule than the R1 rule. By this means, ILOO can be viewed as a measure of evaluating the learning performance of two or more different systems according to the prediction rule they derive on the training set.

Independent Test Set Success Rate (ITS-SR): The prediction rule is induced using a given training set and is evaluated on an independent test set which is totally unknown to the process of deriving the prediction rule.

Sensitivity (SN) and Specificity (SP): Reflects the percentage of True Positive (TP) and True Negative (TN) samples respectively which in case of a medical test are very essential. In ideal situations we expect these metrics to be as high as possible indicating that a classifier can distinguish between TP and TN cases effectively. In realistic situations however absolute success is not usually achieved rendering a high sensitivity results as a more desirable than a high specificity one. The cost of misclassifying a true positive patient is significantly higher than the cost of misclassifying a true negative one. Besides high sensitivity could assist doctors for deciding with higher confidence on the treatment protocol to be followed.

ROC and Precision Recall (PR) Curves, have been used to provide yet another detailed evaluation criterion for the overall performance of classifiers. Nevertheless, they are not directly applied to this work. SN and SP could be used instead as an indirect evaluator of the ROC or PR curves.

In the following tables the above presented measures are used to evaluate the performance of the underlined methodologies in leukemia (Tables 1 and 2) and breast cancer data set (Tables 3 and 4). The first two columns of the following tables refer to the cut off scenario (number of surviving genes per iteration) that were used in the backward elimination process, while the remaining columns refer to the performance of each tested methodology. We provide a list of the acronyms that are used to describe the various measures used:

ILOO (Internal Leave One Out)

ILOO – SR (Internal Leav One Out – Success Rate)

ASR (Apparent Success Rate), i.e., self test estimation.

SEN (Sensitivity)

SP (Specificity)

ITS – SR (Independent TestSet Success Rate)

Table 1 ILOO accuracy performance of RFE-LNW and RFE-FSVs, using a 4^{th} and a combination of RBF and a 7^{th} degree polynomial kernel (Leukemia data set).

		RFE-LNW				RFE-FSVs-4DK				RFE-FSVs-RBF7DK			
ITERATION	GENES	ILOO-SR	ASR	SEN	SP	ILOO-SR	ASR	SEN	SP	ILOO-SR	ASR	SEN	SP
1	7129	94.74	94.74	0.82	1.00	94.74	100.00	0.82	1.00	94.74	100.00	0.82	1.00
2	4096	97.37	94.74	0.91	1.00	97.37	100.00	0.91	1.00	97.37	100.00	0.91	1.00
3	2048	100.00	97.37	1.00	1.00	97.37	100.00	0.91	1.00	97.37	100.00	0.91	1.00
4	1024	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
5	900	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
6	800	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
7	700	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
8	600	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
9	500	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
10	400	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
11	300	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
12	200	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
13	190	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
14	180	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
15	170	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
16	160	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
17	150	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
18	140	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
19	130	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
20	120	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
21	110	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
22	100	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
23	99	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
24	98	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
25	97	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
26	96	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
27	95	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
28	94	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
29	93	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00

		RFE-LNW				RFE-FSVs-4DK				RFE-FSVs-RBF7DK			
ITERATION	**GENES**	**ILOO-SR**	**ASR**	**SEN**	**SP**	**ILOO-SR**	**ASR**	**SEN**	**SP**	**ILOO-SR**	**ASR**	**SEN**	**SP**
30	92	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
31	91	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
32	90	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
33	89	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
34	88	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
35	87	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
36	86	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
37	85	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
38	84	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
39	83	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
40	82	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
41	81	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
42	80	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
43	79	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
44	78	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
45	77	100.00	100.00	1.00	1.00	97.37	100.00	0.91	1.00	100.00	100.00	1.00	1.00
46	76	100.00	100.00	1.00	1.00	97.37	100.00	0.91	1.00	100.00	100.00	1.00	1.00
47	75	100.00	100.00	1.00	1.00	97.37	100.00	0.91	1.00	100.00	100.00	1.00	1.00
48	74	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
49	73	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
50	72	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
51	71	100.00	100.00	1.00	1.00	97.37	100.00	0.91	1.00	100.00	100.00	1.00	1.00
52	70	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
53	69	100.00	100.00	1.00	1.00	97.37	100.00	0.91	1.00	100.00	100.00	1.00	1.00
54	68	100.00	100.00	1.00	1.00	97.37	100.00	0.91	1.00	100.00	100.00	1.00	1.00
55	67	100.00	100.00	1.00	1.00	97.37	100.00	0.91	1.00	100.00	100.00	1.00	1.00
56	66	100.00	100.00	1.00	1.00	97.37	100.00	0.91	1.00	100.00	100.00	1.00	1.00
57	65	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
58	64	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
59	63	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
60	62	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
61	61	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
62	60	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00

		RFE-LNW				RFE-FSVs-4DK				RFE-FSVs-RBF7DK			
ITERATION	**GENES**	**ILOO-SR**	**ASR**	**SEN**	**SP**	**ILOO-SR**	**ASR**	**SEN**	**SP**	**ILOO-SR**	**ASR**	**SEN**	**SP**
63	59	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
64	58	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
65	57	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
66	56	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
67	55	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
68	54	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
69	53	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
70	52	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
71	51	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
72	50	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
73	49	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
74	48	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
75	47	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
76	46	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
77	45	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
78	44	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
79	43	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
80	42	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
81	41	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
82	40	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
83	39	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
84	38	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
85	37	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
86	36	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
87	35	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
88	34	100.00	100.00	1.00	1.00	97.37	100.00	0.91	1.00	100.00	100.00	1.00	1.00
89	33	100.00	100.00	1.00	1.00	97.37	100.00	0.91	1.00	100.00	100.00	1.00	1.00
90	32	100.00	100.00	1.00	1.00	97.37	100.00	0.91	1.00	100.00	100.00	1.00	1.00
91	31	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
92	30	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
93	29	100.00	100.00	1.00	1.00	97.37	100.00	0.91	1.00	100.00	100.00	1.00	1.00
94	28	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
95	27	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00

		RFE-LNW				RFE-FSVs-4DK				RFE-FSVs-RBF7DK			
ITERATION	**GENES**	**ILOO-SR**	**ASR**	**SEN**	**SP**	**ILOO-SR**	**ASR**	**SEN**	**SP**	**ILOO-SR**	**ASR**	**SEN**	**SP**
96	26	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
97	25	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
98	24	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	97.37	100.00	0.91	1.00
99	23	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	97.37	100.00	0.91	1.00
100	22	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
101	21	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
102	20	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
103	19	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
104	18	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
105	17	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
106	16	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	97.37	100.00	0.91	1.00
107	15	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	97.37	100.00	0.91	1.00
108	14	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
109	13	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
110	12	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
111	11	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
112	10	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
113	9	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
114	8	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
115	7	97.37	100.00	0.91	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
116	6	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
117	5	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
118	4	92.11	100.00	0.82	0.96	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
119	3	92.11	100.00	0.82	0.96	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00
120	2	97.37	97.37	0.91	1.00	97.37	100.00	0.91	1.00	94.74	100.00	0.82	1.00
121	1	97.37	97.37	0.91	1.00	97.37	100.00	0.91	1.00	97.37	100.00	0.91	1.00
Average		**99.74**	**99.85**	**0.99**	**1.00**	**99.61**	**100.00**	**0.99**	**1.00**	**99.76**	**100.00**	**0.99**	**1.00**

Table 2 Accuracy performance of the proposed methodologies on the independent test set in Leukemia data set, the 3rd degree polynomial kernel misses only one sample m(Leukemia data set).

		RFE-LNW				RFE-FSVs-4DK				RFE-FSVs-RBF7DK			
ITERATION	GENES	ITS-SR	ASR	SEN	SP	ILOO-SR	ASR	SEN	SP	ILOO-SR	ASR	SEN	SP
1	7129	79.41	94.74	0.50	1.00	79.41	100.00	0.50	1.00	79.41	100.00	0.50	1.00
2	4096	82.35	94.74	0.57	1.00	82.35	100.00	0.57	1.00	82.35	100.00	0.57	1.00
3	2048	82.35	97.37	0.57	1.00	82.35	100.00	0.57	1.00	82.35	100.00	0.57	1.00
4	1024	85.29	100.00	0.64	1.00	82.35	100.00	0.57	1.00	82.35	100.00	0.57	1.00
5	900	85.29	100.00	0.64	1.00	82.35	100.00	0.57	1.00	82.35	100.00	0.57	1.00
6	800	85.29	100.00	0.64	1.00	82.35	100.00	0.57	1.00	82.35	100.00	0.57	1.00
7	700	85.29	100.00	0.64	1.00	82.35	100.00	0.57	1.00	82.35	100.00	0.57	1.00
8	600	85.29	100.00	0.64	1.00	82.35	100.00	0.57	1.00	82.35	100.00	0.57	1.00
9	500	88.24	100.00	0.71	1.00	85.29	100.00	0.64	1.00	85.29	100.00	0.64	1.00
10	400	88.24	100.00	0.71	1.00	82.35	100.00	0.57	1.00	82.35	100.00	0.57	1.00
11	300	88.24	100.00	0.71	1.00	88.24	100.00	0.71	1.00	88.24	100.00	0.71	1.00
12	200	88.24	100.00	0.71	1.00	88.24	100.00	0.71	1.00	88.24	100.00	0.71	1.00
13	190	88.24	100.00	0.71	1.00	88.24	100.00	0.71	1.00	88.24	100.00	0.71	1.00
14	180	85.29	100.00	0.64	1.00	88.24	100.00	0.71	1.00	88.24	100.00	0.71	1.00
15	170	85.29	100.00	0.64	1.00	88.24	100.00	0.71	1.00	88.24	100.00	0.71	1.00
16	160	85.29	100.00	0.64	1.00	88.24	100.00	0.71	1.00	88.24	100.00	0.71	1.00
17	150	85.29	100.00	0.64	1.00	88.24	100.00	0.71	1.00	88.24	100.00	0.71	1.00
18	140	85.29	100.00	0.64	1.00	88.24	100.00	0.71	1.00	88.24	100.00	0.71	1.00
19	130	88.24	100.00	0.71	1.00	88.24	100.00	0.71	1.00	88.24	100.00	0.71	1.00
20	120	91.18	100.00	0.79	1.00	88.24	100.00	0.71	1.00	91.18	100.00	0.79	1.00
21	110	91.18	100.00	0.79	1.00	91.18	100.00	0.79	1.00	94.12	100.00	0.86	1.00
22	100	91.18	100.00	0.79	1.00	88.24	100.00	0.71	1.00	94.12	100.00	0.86	1.00
23	99	88.24	100.00	0.71	1.00	88.24	100.00	0.71	1.00	94.12	100.00	0.86	1.00
24	98	88.24	100.00	0.71	1.00	88.24	100.00	0.71	1.00	94.12	100.00	0.86	1.00
25	97	88.24	100.00	0.71	1.00	91.18	100.00	0.79	1.00	94.12	100.00	0.86	1.00
26	96	88.24	100.00	0.71	1.00	91.18	100.00	0.79	1.00	94.12	100.00	0.86	1.00
27	95	88.24	100.00	0.71	1.00	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00
28	94	88.24	100.00	0.71	1.00	91.18	100.00	0.79	1.00	94.12	100.00	0.86	1.00
29	93	88.24	100.00	0.71	1.00	91.18	100.00	0.79	1.00	94.12	100.00	0.86	1.00

		RFE-LNW				RFE-FSVs-4DK				RFE-FSVs-RBF7DK			
ITERATION	**GENES**	**ITS-SR**	**ASR**	**SEN**	**SP**	**ILOO-SR**	**ASR**	**SEN**	**SP**	**ILOO-SR**	**ASR**	**SEN**	**SP**
30	92	88.24	100.00	0.71	1.00	91.18	100.00	0.79	1.00	94.12	100.00	0.86	1.00
31	91	88.24	100.00	0.71	1.00	88.24	100.00	0.71	1.00	94.12	100.00	0.86	1.00
32	90	88.24	100.00	0.71	1.00	88.24	100.00	0.71	1.00	94.12	100.00	0.86	1.00
33	89	88.24	100.00	0.71	1.00	88.24	100.00	0.71	1.00	94.12	100.00	0.86	1.00
34	88	88.24	100.00	0.71	1.00	91.18	100.00	0.79	1.00	94.12	100.00	0.86	1.00
35	87	88.24	100.00	0.71	1.00	88.24	100.00	0.79	0.95	94.12	100.00	0.86	1.00
36	86	88.24	100.00	0.71	1.00	88.24	100.00	0.79	0.95	94.12	100.00	0.86	1.00
37	85	88.24	100.00	0.71	1.00	88.24	100.00	0.79	0.95	94.12	100.00	0.86	1.00
38	84	88.24	100.00	0.71	1.00	82.35	100.00	0.64	0.95	94.12	100.00	0.86	1.00
39	83	88.24	100.00	0.71	1.00	82.35	100.00	0.64	0.95	94.12	100.00	0.86	1.00
40	82	88.24	100.00	0.71	1.00	85.29	100.00	0.71	0.95	94.12	100.00	0.86	1.00
41	81	88.24	100.00	0.71	1.00	85.29	100.00	0.71	0.95	94.12	100.00	0.86	1.00
42	80	88.24	100.00	0.71	1.00	88.24	100.00	0.71	1.00	94.12	100.00	0.86	1.00
43	79	91.18	100.00	0.79	1.00	88.24	100.00	0.71	1.00	94.12	100.00	0.86	1.00
44	78	91.18	100.00	0.79	1.00	88.24	100.00	0.71	1.00	94.12	100.00	0.86	1.00
45	77	91.18	100.00	0.79	1.00	91.18	100.00	0.79	1.00	94.12	100.00	0.86	1.00
46	76	91.18	100.00	0.79	1.00	91.18	100.00	0.79	1.00	94.12	100.00	0.86	1.00
47	75	91.18	100.00	0.79	1.00	91.18	100.00	0.79	1.00	94.12	100.00	0.86	1.00
48	74	94.12	100.00	0.86	1.00	91.18	100.00	0.79	1.00	94.12	100.00	0.86	1.00
49	73	94.12	100.00	0.86	1.00	91.18	100.00	0.79	1.00	94.12	100.00	0.86	1.00
50	72	94.12	100.00	0.86	1.00	91.18	100.00	0.79	1.00	94.12	100.00	0.86	1.00
51	71	88.24	100.00	0.71	1.00	91.18	100.00	0.79	1.00	94.12	100.00	0.86	1.00
52	70	88.24	100.00	0.71	1.00	91.18	100.00	0.79	1.00	94.12	100.00	0.86	1.00
53	69	94.12	100.00	0.86	1.00	91.18	100.00	0.79	1.00	94.12	100.00	0.86	1.00
54	68	94.12	100.00	0.86	1.00	91.18	100.00	0.79	1.00	94.12	100.00	0.86	1.00
55	67	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00
56	66	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00
57	65	94.12	100.00	0.86	1.00	91.18	100.00	0.79	1.00	94.12	100.00	0.86	1.00
58	64	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00
59	63	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00
60	62	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00
61	61	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00
62	60	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00
63	59	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00

		RFE-LNW				RFE-FSVs-4DK				RFE-FSVs-RBF7DK			
ITERATION	**GENES**	**ITS-SR**	**ASR**	**SEN**	**SP**	**ILOO-SR**	**ASR**	**SEN**	**SP**	**ILOO-SR**	**ASR**	**SEN**	**SP**
64	58	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00
65	57	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00
66	56	94.12	100.00	0.86	1.00	91.18	100.00	0.79	1.00	94.12	100.00	0.86	1.00
67	55	94.12	100.00	0.86	1.00	91.18	100.00	0.79	1.00	94.12	100.00	0.86	1.00
68	54	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00
69	53	94.12	100.00	0.86	1.00	91.18	100.00	0.79	1.00	91.18	100.00	0.79	1.00
70	52	94.12	100.00	0.86	1.00	91.18	100.00	0.79	1.00	91.18	100.00	0.79	1.00
71	51	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00
72	50	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00
73	49	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00
74	48	94.12	100.00	0.86	1.00	91.18	100.00	0.79	1.00	94.12	100.00	0.86	1.00
75	47	94.12	100.00	0.86	1.00	91.18	100.00	0.79	1.00	91.18	100.00	0.79	1.00
76	46	94.12	100.00	0.86	1.00	88.24	100.00	0.71	1.00	94.12	100.00	0.86	1.00
77	45	94.12	100.00	0.86	1.00	91.18	100.00	0.79	1.00	91.18	100.00	0.79	1.00
78	44	91.18	100.00	0.79	1.00	91.18	100.00	0.79	1.00	94.12	100.00	0.86	1.00
79	43	91.18	100.00	0.79	1.00	91.18	100.00	0.79	1.00	94.12	100.00	0.86	1.00
80	42	94.12	100.00	0.86	1.00	91.18	100.00	0.79	1.00	94.12	100.00	0.86	1.00
81	41	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00	91.18	100.00	0.79	1.00
82	40	94.12	100.00	0.86	1.00	91.18	100.00	0.79	1.00	91.18	100.00	0.79	1.00
83	39	94.12	100.00	0.86	1.00	91.18	100.00	0.79	1.00	88.24	100.00	0.79	0.95
84	38	94.12	100.00	0.86	1.00	91.18	100.00	0.79	1.00	88.24	100.00	0.79	0.95
85	37	94.12	100.00	0.86	1.00	94.12	100.00	0.86	1.00	88.24	100.00	0.79	0.95
86	36	94.12	100.00	0.86	1.00	85.29	100.00	0.64	1.00	88.24	100.00	0.79	0.95
87	35	85.29	100.00	0.64	1.00	85.29	100.00	0.64	1.00	88.24	100.00	0.79	0.95
88	34	85.29	100.00	0.64	1.00	85.29	100.00	0.64	1.00	88.24	100.00	0.79	0.95
89	33	85.29	100.00	0.64	1.00	85.29	100.00	0.64	1.00	91.18	100.00	0.86	0.95
90	32	85.29	100.00	0.64	1.00	85.29	100.00	0.64	1.00	91.18	100.00	0.86	0.95
91	31	85.29	100.00	0.64	1.00	85.29	100.00	0.64	1.00	88.24	100.00	0.79	0.95
92	30	88.24	100.00	0.71	1.00	85.29	100.00	0.64	1.00	91.18	100.00	0.86	0.95
93	29	85.29	100.00	0.64	1.00	85.29	100.00	0.64	1.00	91.18	100.00	0.86	0.95
94	28	85.29	100.00	0.64	1.00	85.29	100.00	0.64	1.00	91.18	100.00	0.86	0.95
95	27	85.29	100.00	0.64	1.00	85.29	100.00	0.64	1.00	88.24	100.00	0.79	0.95
96	26	82.35	100.00	0.57	1.00	85.29	100.00	0.64	1.00	88.24	100.00	0.79	0.95
97	25	82.35	100.00	0.57	1.00	85.29	100.00	0.64	1.00	91.18	100.00	0.86	0.95

		RFE-LNW				RFE-FSVs-4DK				RFE-FSVs-RBF7DK			
ITERATION	**GENES**	**ITS-SR**	**ASR**	**SEN**	**SP**	**ILOO-SR**	**ASR**	**SEN**	**SP**	**ILOO-SR**	**ASR**	**SEN**	**SP**
98	24	82.35	100.00	0.57	1.00	85.29	100.00	0.64	1.00	94.12	100.00	0.86	1.00
99	23	82.35	100.00	0.57	1.00	85.29	100.00	0.64	1.00	94.12	100.00	0.86	1.00
100	22	82.35	100.00	0.57	1.00	85.29	100.00	0.64	1.00	94.12	100.00	0.86	1.00
101	21	82.35	100.00	0.57	1.00	88.24	100.00	0.71	1.00	91.18	100.00	0.86	0.95
102	20	82.35	100.00	0.57	1.00	88.24	100.00	0.71	1.00	91.18	100.00	0.86	0.95
103	19	82.35	100.00	0.57	1.00	88.24	100.00	0.71	1.00	88.24	100.00	0.79	0.95
104	18	82.35	100.00	0.57	1.00	82.35	100.00	0.57	1.00	91.18	100.00	0.86	0.95
105	17	85.29	100.00	0.64	1.00	76.47	100.00	0.50	0.95	88.24	100.00	0.79	0.95
106	16	82.35	100.00	0.57	1.00	76.47	100.00	0.50	0.95	88.24	100.00	0.79	0.95
107	15	85.29	100.00	0.64	1.00	76.47	100.00	0.50	0.95	88.24	100.00	0.79	0.95
108	14	85.29	100.00	0.64	1.00	82.35	100.00	0.57	1.00	91.18	100.00	0.79	1.00
109	13	85.29	100.00	0.64	1.00	82.35	100.00	0.57	1.00	91.18	100.00	0.79	1.00
110	12	85.29	100.00	0.64	1.00	82.35	100.00	0.57	1.00	91.18	100.00	0.79	1.00
111	11	85.29	100.00	0.64	1.00	82.35	100.00	0.57	1.00	91.18	100.00	0.79	1.00
112	10	82.35	100.00	0.57	1.00	82.35	100.00	0.57	1.00	91.18	100.00	0.79	1.00
113	9	82.35	100.00	0.57	1.00	82.35	100.00	0.57	1.00	91.18	100.00	0.79	1.00
114	8	82.35	100.00	0.57	1.00	82.35	100.00	0.57	1.00	85.29	100.00	0.64	1.00
115	7	91.18	100.00	0.79	1.00	88.24	100.00	0.71	1.00	82.35	100.00	0.57	1.00
116	6	94.12	100.00	0.86	1.00	85.29	100.00	0.64	1.00	79.41	100.00	0.50	1.00
117	5	82.35	100.00	0.64	0.95	88.24	100.00	0.71	1.00	82.35	100.00	0.57	1.00
118	4	88.24	100.00	0.71	1.00	88.24	100.00	0.79	0.95	79.41	100.00	0.50	1.00
119	3	88.24	100.00	0.71	1.00	88.24	100.00	0.79	0.95	**97.06**	**100.00**	**0.93**	**1.00**
120	2	**94.12**	**97.37**	**0.93**	**0.95**	**94.12**	**100.00**	**0.93**	**0.95**	94.12	100.00	0.93	0.95
121	1	91.18	97.37	0.93	0.90	91.18	100.00	0.93	0.90	91.18	100.00	0.93	0.90
Average		**88.89**	**99.85**	**0.73**	**1.00**	**88.19**	**100.00**	**0.72**	**0.99**	**91.03**	**100.00**	**0.80**	**0.99**

Table 3 ILOO accuracy results of the tested methodologies on BC domain.

		RFE-LNW				RFE-SVM				RFE-FSVs-7DK			
ITERATION	GENES	ILOO-SR	ASR	SEN	SP	ILOO-SR	ASR	SEN	SP	ILOO-SR	ASR	SEN	SP
1	24188	62.82	97.44	0.59	0.66	62.82	100.00	0.59	0.66	62.82	100.00	0.59	0.66
2	16384	87.18	100.00	0.82	0.91	82.05	100.00	0.76	0.86	71.79	100.00	0.65	0.77
3	8192	98.72	100.00	0.97	1.00	93.59	100.00	0.91	0.95	84.62	100.00	0.76	0.91
4	4096	98.72	100.00	0.97	1.00	98.72	100.00	0.97	1.00	84.62	100.00	0.79	0.89
5	2048	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	85.90	100.00	0.79	0.91
6	1024	97.44	100.00	0.94	1.00	98.72	100.00	0.97	1.00	84.62	100.00	0.79	0.89
7	900	97.44	100.00	0.94	1.00	100.00	100.00	1.00	1.00	82.05	100.00	0.74	0.89
8	800	98.72	100.00	0.97	1.00	100.00	100.00	1.00	1.00	82.05	100.00	0.76	0.86
9	700	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	82.05	100.00	0.76	0.86
10	600	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	84.62	100.00	0.79	0.89
11	500	100.00	100.00	1.00	1.00	98.72	100.00	0.97	1.00	85.90	100.00	0.79	0.91
12	400	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	87.18	100.00	0.85	0.89
13	300	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	83.33	100.00	0.85	0.82
14	200	97.44	100.00	0.94	1.00	100.00	100.00	1.00	1.00	83.33	100.00	0.82	0.84
15	190	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	80.77	100.00	0.82	0.80
16	180	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	83.33	100.00	0.82	0.84
17	170	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	83.33	100.00	0.82	0.84
18	160	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	84.62	100.00	0.82	0.86
19	150	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	85.90	100.00	0.82	0.89
20	140	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	88.46	100.00	0.82	0.93
21	130	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	92.31	100.00	0.88	0.95
22	120	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	91.03	100.00	0.88	0.93
23	110	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	89.74	100.00	0.88	0.91
24	100	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	91.03	100.00	0.91	0.91
25	99	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	89.74	100.00	0.88	0.91
26	98	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	89.74	100.00	0.88	0.91
27	97	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	91.03	100.00	0.91	0.91
28	96	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	91.03	100.00	0.91	0.91
29	95	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	91.03	100.00	0.91	0.91
30	94	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	91.03	100.00	0.91	0.91
31	93	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	91.03	100.00	0.91	0.91
32	92	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	91.03	100.00	0.91	0.91

		RFE-LNW				RFE-SVM				RFE-FSVs-7DK			
ITERATION	**GENES**	**ILOO-SR**	**ASR**	**SEN**	**SP**	**ILOO-SR**	**ASR**	**SEN**	**SP**	**ILOO-SR**	**ASR**	**SEN**	**SP**
33	91	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	89.74	100.00	0.91	0.89
34	90	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	91.03	100.00	0.91	0.91
35	89	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	91.03	100.00	0.91	0.91
36	88	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	91.03	100.00	0.91	0.91
37	87	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	91.03	100.00	0.91	0.91
38	86	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	91.03	100.00	0.91	0.91
39	85	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	91.03	100.00	0.91	0.91
40	84	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	91.03	100.00	0.91	0.91
41	83	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	89.74	100.00	0.88	0.91
42	82	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	85.90	100.00	0.85	0.86
43	81	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	84.62	100.00	0.85	0.84
44	80	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	85.90	100.00	0.85	0.86
45	79	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	85.90	100.00	0.85	0.86
46	78	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	85.90	100.00	0.85	0.86
47	77	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	88.46	100.00	0.85	0.91
48	76	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	84.62	100.00	0.82	0.86
49	75	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	88.46	100.00	0.85	0.91
50	74	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	88.46	100.00	0.85	0.91
51	73	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	87.18	100.00	0.85	0.89
52	72	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	87.18	100.00	0.85	0.89
53	71	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	87.18	100.00	0.82	0.91
54	70	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	85.90	100.00	0.82	0.89
55	69	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	87.18	100.00	0.82	0.91
56	68	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	85.90	100.00	0.82	0.89
57	67	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	85.90	100.00	0.82	0.89
58	66	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	85.90	100.00	0.82	0.89
59	65	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	85.90	100.00	0.82	0.89
60	64	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	85.90	100.00	0.82	0.89
61	63	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	87.18	100.00	0.82	0.91
62	62	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	88.46	100.00	0.82	0.93
63	61	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	89.74	100.00	0.85	0.93
64	60	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	89.74	100.00	0.85	0.93
65	59	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	89.74	100.00	0.85	0.93
66	58	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	91.03	100.00	0.85	0.95

		RFE-LNW				RFE-SVM				RFE-FSVs-7DK			
ITERATION	**GENES**	**ILOO-SR**	**ASR**	**SEN**	**SP**	**ILOO-SR**	**ASR**	**SEN**	**SP**	**ILOO-SR**	**ASR**	**SEN**	**SP**
67	57	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	89.74	100.00	0.85	0.93
68	56	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	88.46	100.00	0.85	0.91
69	55	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	87.18	100.00	0.85	0.89
70	54	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	91.03	100.00	0.85	0.95
71	53	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	91.03	100.00	0.85	0.95
72	52	100.00	100.00	1.00	1.00	98.72	100.00	0.97	1.00	91.03	100.00	0.85	0.95
73	51	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	87.18	100.00	0.85	0.89
74	50	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	87.18	100.00	0.85	0.89
75	49	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	87.18	100.00	0.85	0.89
76	48	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	84.62	100.00	0.79	0.89
77	47	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	91.03	100.00	0.88	0.93
78	46	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	91.03	100.00	0.88	0.93
79	45	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	88.46	100.00	0.88	0.89
80	44	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	88.46	100.00	0.82	0.93
81	43	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	89.74	100.00	0.85	0.93
82	42	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	88.46	100.00	0.85	0.91
83	41	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	87.18	100.00	0.85	0.89
84	40	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	87.18	100.00	0.85	0.89
85	39	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	85.90	100.00	0.85	0.86
86	38	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	88.46	100.00	0.85	0.91
87	37	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	87.18	100.00	0.85	0.89
88	36	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	88.46	100.00	0.85	0.91
89	35	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	88.46	100.00	0.85	0.91
90	34	100.00	100.00	1.00	1.00	97.44	100.00	1.00	0.95	89.74	100.00	0.85	0.93
91	33	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	87.18	100.00	0.85	0.89
92	32	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	88.46	100.00	0.85	0.91
93	31	100.00	100.00	1.00	1.00	98.72	100.00	1.00	0.98	88.46	100.00	0.85	0.91
94	30	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	88.46	100.00	0.85	0.91
95	29	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	88.46	100.00	0.85	0.91
96	28	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	89.74	100.00	0.88	0.91
97	27	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	88.46	100.00	0.88	0.89
98	26	100.00	100.00	1.00	1.00	98.72	100.00	1.00	0.98	84.62	100.00	0.82	0.86
99	25	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	83.33	100.00	0.82	0.84
100	24	100.00	100.00	1.00	1.00	98.72	100.00	1.00	0.98	87.18	100.00	0.88	0.86

		RFE-LNW				RFE-SVM				RFE-FSVs-7DK			
ITERATION	**GENES**	**ILOO-SR**	**ASR**	**SEN**	**SP**	**ILOO-SR**	**ASR**	**SEN**	**SP**	**ILOO-SR**	**ASR**	**SEN**	**SP**
101	23	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	88.46	100.00	0.85	0.91
102	22	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	92.31	100.00	0.88	0.95
103	21	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	85.90	100.00	0.82	0.89
104	20	100.00	100.00	1.00	1.00	100.00	100.00	1.00	1.00	85.90	100.00	0.82	0.89
105	19	100.00	100.00	1.00	1.00	98.72	100.00	0.97	1.00	85.90	100.00	0.82	0.89
106	18	100.00	100.00	1.00	1.00	98.72	100.00	0.97	1.00	85.90	100.00	0.82	0.89
107	17	100.00	100.00	1.00	1.00	94.87	100.00	0.94	0.95	82.05	100.00	0.76	0.86
108	16	100.00	100.00	1.00	1.00	97.44	100.00	0.97	0.98	83.33	100.00	0.79	0.86
109	15	98.72	100.00	1.00	0.98	98.72	100.00	1.00	0.98	83.33	100.00	0.82	0.84
110	14	100.00	100.00	1.00	1.00	97.44	100.00	0.94	1.00	85.90	100.00	0.85	0.86
111	13	100.00	100.00	1.00	1.00	98.72	100.00	0.97	1.00	87.18	100.00	0.85	0.89
112	12	100.00	100.00	1.00	1.00	92.31	100.00	0.88	0.95	83.33	100.00	0.82	0.84
113	11	97.44	100.00	0.97	0.98	97.44	100.00	0.97	0.98	82.05	100.00	0.82	0.82
114	10	98.72	100.00	0.97	1.00	97.44	100.00	0.97	0.98	76.92	100.00	0.76	0.77
115	9	98.72	100.00	0.97	1.00	97.44	100.00	0.97	0.98	82.05	100.00	0.79	0.84
116	8	100.00	100.00	1.00	1.00	88.46	96.15	0.85	0.91	80.77	100.00	0.79	0.82
117	7	92.31	98.72	0.91	0.93	84.62	91.03	0.76	0.91	83.33	100.00	0.85	0.82
118	6	91.03	93.59	0.88	0.93	83.33	88.46	0.71	0.93	78.21	100.00	0.79	0.77
119	5	88.46	92.31	0.85	0.91	82.05	84.62	0.74	0.89	76.92	100.00	0.71	0.82
120	4	85.90	91.03	0.79	0.91	74.36	82.05	0.71	0.77	82.05	92.31	0.79	0.84
121	3	82.05	85.90	0.79	0.84	74.36	78.21	0.68	0.80	79.49	89.74	0.74	0.84
122	2	73.08	76.92	0.68	0.77	76.92	78.21	0.65	0.86	74.36	82.05	0.76	0.73
123	1	66.67	67.95	0.50	0.80	73.08	73.08	0.65	0.80	69.23	71.79	0.59	0.77
Average		**98.47**	**99.22**	**0.98**	**0.99**	**97.83**	**98.96**	**0.97**	**0.98**	**86.42**	**99.48**	**0.84**	**0.88**

Table 4 Performance of the tested methodologies on an independent test set on BC domain, RFE-LNW misses only one sample with 31 genes and 2 samples with 7 genes.

		RFE-LNW				RFE-SVM				RFE-FSVs-7DK			
ITERATION	GENES	ITS-SR	ASR	SEN	SP	ILOO-SR	ASR	SEN	SP	ILOO-SR	ASR	SEN	SP
1	24188	68.42	97.44	0.92	0.29	68.42	100.00	0.92	0.29	68.42	100.00	0.92	0.29
2	16384	68.42	100.00	0.92	0.29	68.42	100.00	0.92	0.29	68.42	100.00	0.92	0.29
3	8192	63.16	100.00	0.92	0.14	68.42	100.00	0.92	0.29	73.68	100.00	0.92	0.43
4	4096	57.89	100.00	0.92	0.00	63.16	100.00	0.75	0.43	68.42	100.00	0.92	0.29
5	2048	68.42	100.00	0.92	0.29	57.89	100.00	0.67	0.43	73.68	100.00	1.00	0.29
6	1024	63.16	100.00	0.92	0.14	57.89	100.00	0.67	0.43	73.68	100.00	1.00	0.29
7	900	63.16	100.00	0.92	0.14	57.89	100.00	0.67	0.43	84.21	100.00	1.00	0.57
8	800	63.16	100.00	0.92	0.14	52.63	100.00	0.67	0.29	78.95	100.00	1.00	0.43
9	700	73.68	100.00	1.00	0.29	57.89	100.00	0.67	0.43	78.95	100.00	1.00	0.43
10	600	73.68	100.00	1.00	0.29	57.89	100.00	0.75	0.29	78.95	100.00	1.00	0.43
11	500	68.42	100.00	1.00	0.14	63.16	100.00	0.75	0.43	78.95	100.00	1.00	0.43
12	400	68.42	100.00	1.00	0.14	63.16	100.00	0.75	0.43	73.68	100.00	1.00	0.29
13	300	68.42	100.00	1.00	0.14	63.16	100.00	0.75	0.43	68.42	100.00	1.00	0.14
14	200	68.42	100.00	1.00	0.14	68.42	100.00	0.67	0.71	73.68	100.00	0.92	0.43
15	190	68.42	100.00	1.00	0.14	73.68	100.00	0.67	0.86	84.21	100.00	1.00	0.57
16	180	68.42	100.00	1.00	0.14	68.42	100.00	0.58	0.86	84.21	100.00	1.00	0.57
17	170	68.42	100.00	1.00	0.14	68.42	100.00	0.67	0.71	78.95	100.00	1.00	0.43
18	160	63.16	100.00	0.92	0.14	68.42	100.00	0.67	0.71	73.68	100.00	0.92	0.43
19	150	63.16	100.00	0.92	0.14	68.42	100.00	0.67	0.71	68.42	100.00	0.92	0.29
20	140	63.16	100.00	0.92	0.14	68.42	100.00	0.58	0.86	63.16	100.00	0.92	0.14
21	130	63.16	100.00	0.92	0.14	68.42	100.00	0.67	0.71	68.42	100.00	0.92	0.29
22	120	63.16	100.00	0.92	0.14	68.42	100.00	0.67	0.71	63.16	100.00	0.92	0.14
23	110	63.16	100.00	0.92	0.14	68.42	100.00	0.67	0.71	63.16	100.00	0.92	0.14
24	100	63.16	100.00	0.92	0.14	68.42	100.00	0.67	0.71	73.68	100.00	0.92	0.43
25	99	63.16	100.00	0.92	0.14	68.42	100.00	0.67	0.71	78.95	100.00	0.92	0.57
26	98	63.16	100.00	0.92	0.14	68.42	100.00	0.67	0.71	78.95	100.00	0.92	0.57
27	97	63.16	100.00	0.92	0.14	68.42	100.00	0.67	0.71	78.95	100.00	0.92	0.57
28	96	63.16	100.00	0.92	0.14	73.68	100.00	0.75	0.71	78.95	100.00	0.92	0.57
29	95	63.16	100.00	0.92	0.14	73.68	100.00	0.75	0.71	78.95	100.00	0.92	0.57

		RFE-LNW				RFE-SVM				RFE-FSVs-7DK			
ITERATION	**GENES**	**ITS-SR**	**ASR**	**SEN**	**SP**	**ILOO-SR**	**ASR**	**SEN**	**SP**	**ILOO-SR**	**ASR**	**SEN**	**SP**
30	94	63.16	100.00	0.92	0.14	73.68	100.00	0.75	0.71	78.95	100.00	0.92	0.57
31	93	63.16	100.00	0.92	0.14	73.68	100.00	0.75	0.71	78.95	100.00	0.92	0.57
32	92	63.16	100.00	0.92	0.14	78.95	100.00	0.83	0.71	78.95	100.00	0.92	0.57
33	91	63.16	100.00	0.92	0.14	73.68	100.00	0.75	0.71	78.95	100.00	0.92	0.57
34	90	63.16	100.00	0.92	0.14	73.68	100.00	0.75	0.71	78.95	100.00	0.92	0.57
35	89	63.16	100.00	0.92	0.14	73.68	100.00	0.75	0.71	78.95	100.00	0.92	0.57
36	88	63.16	100.00	0.92	0.14	73.68	100.00	0.75	0.71	78.95	100.00	0.92	0.57
37	87	68.42	100.00	0.92	0.29	73.68	100.00	0.75	0.71	78.95	100.00	0.92	0.57
38	86	63.16	100.00	0.92	0.14	73.68	100.00	0.75	0.71	84.21	100.00	0.92	0.71
39	85	63.16	100.00	0.92	0.14	73.68	100.00	0.75	0.71	84.21	100.00	0.92	0.71
40	84	63.16	100.00	0.92	0.14	73.68	100.00	0.75	0.71	84.21	100.00	0.92	0.71
41	83	68.42	100.00	0.92	0.29	73.68	100.00	0.75	0.71	84.21	100.00	0.92	0.71
42	82	68.42	100.00	0.92	0.29	73.68	100.00	0.75	0.71	89.47	100.00	0.92	0.86
43	81	68.42	100.00	0.92	0.29	73.68	100.00	0.75	0.71	89.47	100.00	0.92	0.86
44	80	68.42	100.00	0.92	0.29	73.68	100.00	0.75	0.71	84.21	100.00	0.92	0.71
45	79	73.68	100.00	0.92	0.43	73.68	100.00	0.75	0.71	84.21	100.00	0.92	0.71
46	78	73.68	100.00	0.92	0.43	73.68	100.00	0.75	0.71	84.21	100.00	0.92	0.71
47	77	73.68	100.00	0.92	0.43	73.68	100.00	0.75	0.71	84.21	100.00	0.92	0.71
48	76	78.95	100.00	0.92	0.57	73.68	100.00	0.75	0.71	89.47	100.00	0.92	0.86
49	75	78.95	100.00	0.92	0.57	78.95	100.00	0.83	0.71	89.47	100.00	0.92	0.86
50	74	78.95	100.00	0.92	0.57	78.95	100.00	0.83	0.71	89.47	100.00	0.92	0.86
51	73	73.68	100.00	0.92	0.43	68.42	100.00	0.67	0.71	**94.74**	**100.00**	**0.92**	**1.00**
52	72	78.95	100.00	0.92	0.57	68.42	100.00	0.67	0.71	89.47	100.00	0.92	0.86
53	71	78.95	100.00	0.92	0.57	73.68	100.00	0.75	0.71	84.21	100.00	0.92	0.71
54	70	78.95	100.00	0.92	0.57	73.68	100.00	0.75	0.71	73.68	100.00	0.92	0.43
55	69	78.95	100.00	0.92	0.57	63.16	100.00	0.58	0.71	78.95	100.00	0.92	0.57
56	68	78.95	100.00	0.83	0.71	63.16	100.00	0.58	0.71	84.21	100.00	0.83	0.86
57	67	78.95	100.00	0.83	0.71	57.89	100.00	0.58	0.57	84.21	100.00	0.92	0.71
58	66	78.95	100.00	0.83	0.71	63.16	100.00	0.67	0.57	73.68	100.00	0.83	0.57
59	65	78.95	100.00	0.83	0.71	68.42	100.00	0.75	0.57	73.68	100.00	0.83	0.57
60	64	78.95	100.00	0.83	0.71	78.95	100.00	0.75	0.86	73.68	100.00	0.83	0.57
61	63	73.68	100.00	0.83	0.57	68.42	100.00	0.67	0.71	68.42	100.00	0.83	0.43
62	62	73.68	100.00	0.83	0.57	68.42	100.00	0.67	0.71	68.42	100.00	0.83	0.43
63	61	73.68	100.00	0.92	0.43	68.42	100.00	0.58	0.86	68.42	100.00	0.83	0.43

		RFE-LNW				RFE-SVM				RFE-FSVs-7DK			
ITERATION	**GENES**	**ITS-SR**	**ASR**	**SEN**	**SP**	**ILOO-SR**	**ASR**	**SEN**	**SP**	**ILOO-SR**	**ASR**	**SEN**	**SP**
64	60	78.95	100.00	0.92	0.57	63.16	100.00	0.50	0.86	68.42	100.00	0.83	0.43
65	59	73.68	100.00	0.92	0.43	63.16	100.00	0.50	0.86	68.42	100.00	0.83	0.43
66	58	73.68	100.00	0.92	0.43	63.16	100.00	0.50	0.86	68.42	100.00	0.92	0.29
67	57	68.42	100.00	0.83	0.43	63.16	100.00	0.50	0.86	68.42	100.00	0.83	0.43
68	56	73.68	100.00	0.92	0.43	57.89	100.00	0.50	0.71	68.42	100.00	0.75	0.57
69	55	68.42	100.00	0.83	0.43	63.16	100.00	0.58	0.71	68.42	100.00	0.75	0.57
70	54	78.95	100.00	0.83	0.71	68.42	100.00	0.58	0.86	73.68	100.00	0.83	0.57
71	53	78.95	100.00	0.83	0.71	63.16	100.00	0.58	0.71	78.95	100.00	0.83	0.71
72	52	78.95	100.00	0.83	0.71	68.42	100.00	0.58	0.86	78.95	100.00	0.83	0.71
73	51	78.95	100.00	0.83	0.71	68.42	100.00	0.67	0.71	68.42	100.00	0.83	0.43
74	50	78.95	100.00	0.83	0.71	63.16	100.00	0.58	0.71	68.42	100.00	0.83	0.43
75	49	78.95	100.00	0.83	0.71	63.16	100.00	0.58	0.71	68.42	100.00	0.83	0.43
76	48	84.21	100.00	0.83	0.86	68.42	100.00	0.58	0.86	68.42	100.00	0.83	0.43
77	47	89.47	100.00	0.92	0.86	78.95	100.00	0.75	0.86	73.68	100.00	0.75	0.71
78	46	89.47	100.00	0.92	0.86	68.42	100.00	0.58	0.86	63.16	100.00	0.75	0.43
79	45	89.47	100.00	0.92	0.86	73.68	100.00	0.67	0.86	68.42	100.00	0.75	0.57
80	44	**89.47**	**100.00**	**0.92**	**0.86**	68.42	100.00	0.58	0.86	63.16	100.00	0.67	0.57
81	43	78.95	100.00	0.83	0.71	73.68	100.00	0.67	0.86	63.16	100.00	0.67	0.57
82	42	78.95	100.00	0.83	0.71	68.42	100.00	0.67	0.71	68.42	100.00	0.67	0.71
83	41	78.95	100.00	0.83	0.71	78.95	100.00	0.75	0.86	68.42	100.00	0.67	0.71
84	40	78.95	100.00	0.83	0.71	73.68	100.00	0.67	0.86	63.16	100.00	0.67	0.57
85	39	73.68	100.00	0.75	0.71	78.95	100.00	0.75	0.86	78.95	100.00	0.75	0.86
86	38	78.95	100.00	0.83	0.71	78.95	100.00	0.75	0.86	78.95	100.00	0.75	0.86
87	37	78.95	100.00	0.83	0.71	78.95	100.00	0.75	0.86	73.68	100.00	0.75	0.71
88	36	78.95	100.00	0.83	0.71	78.95	100.00	0.75	0.86	73.68	100.00	0.75	0.71
89	35	84.21	100.00	0.92	0.71	78.95	100.00	0.75	0.86	73.68	100.00	0.75	0.71
90	34	84.21	100.00	0.92	0.71	78.95	100.00	0.75	0.86	68.42	100.00	0.75	0.57
91	33	78.95	100.00	0.83	0.71	73.68	100.00	0.75	0.71	68.42	100.00	0.67	0.71
92	32	78.95	100.00	0.83	0.71	**78.95**	**100.00**	**0.75**	**0.86**	63.16	100.00	0.58	0.71
93	31	78.95	100.00	0.83	0.71	73.68	100.00	0.75	0.71	63.16	100.00	0.58	0.71
94	30	78.95	100.00	0.83	0.71	68.42	100.00	0.67	0.71	63.16	100.00	0.58	0.71
95	29	73.68	100.00	0.83	0.57	68.42	100.00	0.67	0.71	68.42	100.00	0.67	0.71
96	28	73.68	100.00	0.75	0.71	73.68	100.00	0.67	0.86	68.42	100.00	0.58	0.86
97	27	73.68	100.00	0.83	0.57	73.68	100.00	0.67	0.86	52.63	100.00	0.67	0.29

		RFE-LNW				RFE-SVM				RFE-FSVs-7DK			
ITERATION	**GENES**	**ITS-SR**	**ASR**	**SEN**	**SP**	**ILOO-SR**	**ASR**	**SEN**	**SP**	**ILOO-SR**	**ASR**	**SEN**	**SP**
98	26	73.68	100.00	0.75	0.71	73.68	100.00	0.67	0.86	57.89	100.00	0.67	0.43
99	25	73.68	100.00	0.75	0.71	73.68	100.00	0.67	0.86	57.89	100.00	0.67	0.43
100	24	68.42	100.00	0.75	0.57	68.42	100.00	0.58	0.86	63.16	100.00	0.67	0.57
101	23	68.42	100.00	0.75	0.57	68.42	100.00	0.58	0.86	57.89	100.00	0.67	0.43
102	22	68.42	100.00	0.75	0.57	63.16	100.00	0.50	0.86	63.16	100.00	0.67	0.57
103	21	73.68	100.00	0.83	0.57	63.16	100.00	0.50	0.86	63.16	100.00	0.67	0.57
104	20	73.68	100.00	0.83	0.57	68.42	100.00	0.58	0.86	57.89	100.00	0.67	0.43
105	19	73.68	100.00	0.83	0.57	63.16	100.00	0.50	0.86	57.89	100.00	0.67	0.43
106	18	73.68	100.00	0.83	0.57	63.16	100.00	0.50	0.86	68.42	100.00	0.83	0.43
107	17	68.42	100.00	0.75	0.57	68.42	100.00	0.58	0.86	68.42	100.00	0.75	0.57
108	16	73.68	100.00	0.83	0.57	63.16	100.00	0.50	0.86	57.89	100.00	0.83	0.14
109	15	68.42	100.00	0.75	0.57	57.89	100.00	0.42	0.86	57.89	100.00	0.83	0.14
110	14	68.42	100.00	0.75	0.57	57.89	100.00	0.42	0.86	57.89	100.00	0.75	0.29
111	13	63.16	100.00	0.58	0.71	68.42	100.00	0.58	0.86	52.63	100.00	0.67	0.29
112	12	52.63	100.00	0.42	0.71	68.42	100.00	0.58	0.86	52.63	100.00	0.67	0.29
113	11	57.89	100.00	0.50	0.71	73.68	100.00	0.67	0.86	52.63	100.00	0.67	0.29
114	10	52.63	100.00	0.50	0.57	68.42	100.00	0.58	0.86	68.42	100.00	0.75	0.57
115	9	47.37	100.00	0.42	0.57	73.68	100.00	0.67	0.86	63.16	100.00	0.75	0.43
116	8	47.37	100.00	0.42	0.57	78.95	96.15	0.75	0.86	63.16	100.00	0.75	0.43
117	7	47.37	98.72	0.42	0.57	68.42	91.03	0.58	0.86	63.16	100.00	0.75	0.43
118	6	47.37	93.59	0.42	0.57	63.16	88.46	0.50	0.86	63.16	100.00	0.75	0.43
119	5	42.11	92.31	0.33	0.57	63.16	84.62	0.50	0.86	52.63	100.00	0.58	0.43
120	4	52.63	91.03	0.33	0.86	68.42	82.05	0.58	0.86	47.37	92.31	0.50	0.43
121	3	42.11	85.90	0.33	0.57	68.42	78.21	0.58	0.86	42.11	89.74	0.42	0.43
122	2	57.89	76.92	0.42	0.86	73.68	78.21	0.75	0.71	47.37	82.05	0.42	0.57
123	1	36.84	67.95	0.42	0.29	73.68	73.08	0.67	0.86	63.16	71.79	0.58	0.71
Average		**69.88**	**99.22**	**0.83**	**0.47**	**69.32**	**98.96**	**0.67**	**0.74**	**71.46**	**99.48**	**0.82**	**0.53**

APPENDIX II

Marker Genes

Table 1: Breast Cancer Domain: The 73-gene signature selected by RFE-FSVs-7DK.

Accession	Symbol	Description
AB002324	KIAA0326	zinc finger protein 629
Contig51464_RC	FLJ22477	F-box protein 31
NM_001685	ATP5J	ATP synthase, H+ transporting …
NM_013361	ZNF223	zinc finger protein 223
AL049689	LOC63923	
NM_000797	DRD4	dopamine receptor D4
NM_001667	ARL2	ADP-ribosylation factor-like 2
NM_015849	ELA2B	elastase 2B
NM_014489	FRAG1	FGF receptor activating protein 1
AB002297	DOCK3	dedicator of cytokinesis 3
NM_016017	LOC51630	CGI-70 protein
NM_003674	CDK10	cyclin-dependent kinase 10
Contig43684	FLJ23312	hypothetical protein FLJ23312
AF052087	LOC58509	NY-REN-24 antigen
Contig32125_RC		
Contig22253_RC	FLJ21062	hypothetical protein FLJ21062
Contig23356_RC		
NM_003862	FGF18	fibroblast growth factor 18
Contig54742_RC		
NM_000127	EXT1	exostoses (multiple) 1
NM_019028	ZDHHC13	zinc finger, DHHC-type containing 13
AF160213	LOC56889	transmembrane 9 superfamily member 3
NM_004703	RABEP1	rabaptin, RAB GTPase binding effector protein 1
NM_000158	GBE1	glucan (1,4-alpha-), branching enzyme 1
AF148505	ALDH6A1	aldehyde dehydrogenase 6 family, member A1
U82987	BBC3	BCL2 binding component 3
NM_000272	NPHP1	nephronophthisis 1 (juvenile)
NM_013306	SNX15	sorting nexin 15
NM_005176	ATP5G2	ATP synthase, H+ transporting, mitochondrial F0 complex…
NM_016359	NUSAP1	nucleolar and spindle associated protein 1
AL133603		
Contig49512_RC		
NM_002896	RBM4	RNA binding motif protein 4
NM_013438	UBQLN1	ubiquilin 1
Contig33814_RC		
NM_013360	ZNF222	zinc finger protein 222
AJ011306	DKFZP586J0119	translation initiation factor eIF-2b delta subunit
NM_005243	EWSR1	Ewing sarcoma breakpoint region 1
NM_000436	OXCT1	3-oxoacid CoA transferase 1
NM_003748	ALDH4A1	aldehyde dehydrogenase 4 family, member A1
Contig52554_RC		
NM_019086	FLJ20674	hypothetical protein FLJ20674

Accession	Symbol	Description
AF257175	PECI	peroxisomal D3,D2-enoyl-CoA isomerase
AA555029_RC		
NM_003079	SMARCE1	SWI/SNF related, matrix associated …
Contig63102_RC	FLJ11354	
Contig48328_RC		
NM_005219	DIAPH1	diaphanous homolog 1 (Drosophila)
NM_014675	CROCC	ciliary rootlet coiled-coil, rootletin
Contig26388_RC		
NM_018433	JMJD1A	jumonji domain containing 1A
NM_005774	ZNF255	zinc finger protein 224
Contig42933_RC		
Contig20217_RC		
BE739817_RC	IFNAR1	interferon (alpha, beta and omega) receptor 1
Contig11065_RC		
NM_020120	UGCGL1	UDP-glucose ceramide glucosyltransferase-like 1
NM_016444	ZNF226	zinc finger protein 226
NM_003239	TGFB3	transforming growth factor, beta 3
M26880	UBC	ubiquitin C
Contig14882_RC		
NM_020974	SCUBE2	signal peptide, CUB domain, EGF-like 2
AL080059		
Contig31312_RC		
NM_001661	ARL4D	ADP-ribosylation factor-like 4D
Contig47544_RC	ATP5E	ATP synthase, H+ transporting, mitochondrial F1 complex…
NM_013376	SERTAD1	SERTA domain containing 1
NM_018089	ANKZF1	ankyrin repeat and zinc finger domain containing 1
NM_006544	EXOC5	exocyst complex component 5
Contig44278_RC	DKFZP434K114	WD repeat domain 21A
Contig6238_RC		
Contig65439	FLJ21939	
NM_016448	DTL	denticleless homolog (Drosophila)

Table 2: Breast Cancer Domain: The 44-gene signature selected by RFE-LNW.

Accession	Symbol	Description
NM_006544	EXOC5	exocyst complex component 5
NM_013360	ZNF222	zinc finger protein 222
NM_020123	TM9SF3	transmembrane 9 superfamily member 3
NM_004953	EIF4G1	eukaryotic translation initiation factor 4 gamma, 1
NM_004604	STX4	syntaxin 4
Contig42746_RC		
NM_004721	MAP3K13	mitogen-activated protein kinase kinase kinase 13
BE739817_RC	IFNAR1	interferon (alpha, beta and omega) receptor 1
NM_019886	CHST7	carbohydrate (N-acetylglucosamine 6-O) sulfotransferase 7
X89657	ADAM3A	ADAM metallopeptidase domain 3A (cyritestin 1)
AJ011306	DKFZP586J0119	ranslation initiation factor eIF-2b delta subunit
NM_005371	METTL1	methyltransferase like 1
NM_001204	BMPR2	bone morphogenetic protein receptor, type II (serine/threonine kinase)
Contig47544_RC	ATP5E	ATP synthase, H+ transporting, mitochondrial F1 complex, epsilon subunit
NM_000207	INS	insulin
Contig57801_RC		
Contig412_RC	FLJ22233	olute carrier family 24 (sodium/potassium/calcium exchanger), member 6
Contig37160		
Y18643	METTL1	methyltransferase like 1
NM_018042	SLFN12	schlafen family member 12
NM_005243	EWSR1	Ewing sarcoma breakpoint region 1
AL080059		
NM_002896	RBM4	RNA binding motif protein 4
NM_005258	GCHFR	GTP cyclohydrolase I feedback regulator
NM_002833	PTPN9	protein tyrosine phosphatase, non-receptor type 9
Contig44278_RC	DKFZP434K114	WD repeat domain 21A
NM_002704	PPBP	pro-platelet basic protein (chemokine (C-X-C motif) ligand 7)
NM_016448	DTL	denticleless homolog (Drosophila)
Contig15164_RC		
Contig40185_RC		
Contig14882_RC		
AF160213	LOC56889	transmembrane 9 superfamily member 3
AB033036	RP13-347D8.3	KIAA1210 protein
AL050204		
NM_003862	FGF18	fibroblast growth factor 18
NM_013438	UBQLN1	ubiquilin 1
Contig43684	FLJ23312	hypothetical protein FLJ23312
AA555029_RC		
NM_005774	ZNF255	zinc finger protein 224
NM_018964	SLC37A1	solute carrier family 37 (glycerol-3-phosphate transporter), member 1
Contig56217_RC		
Contig32125_RC		
NM_005217	DEFA3	defensin, alpha 3, neutrophil-specific
AF043324	NMT1	N-myristoyltransferase 1

Table 3: Breast Cancer Domain: The 32-gene signature selected by RFE-SVM.

Accession	Symbol	Description
Contig10750_RC		
Contig31000_RC		
Contig48328_RC		
AL080059		
NM_006398	UBD	ubiquitin D
NM_019851	FGF20	fibroblast growth factor 20
NM_001062	TCN1	transcobalamin I (vitamin B12 binding protein, R binder family)
NM_000067	CA2	carbonic anhydrase II
NM_020974	SCUBE2	signal peptide, CUB domain, EGF-like 2
NM_001756	SERPINA6	serpin peptidase inhibitor, clade A …
NM_014665	LRRC14	leucine rich repeat containing 14
Contig50950_RC		
Contig11072_RC		
NM_006551	SCGB1D2	secretoglobin, family 1D, member 2
NM_001615	ACTG2	actin, gamma 2, smooth muscle, enteric
Contig53371_RC		
AF131741	LOC441052	hypothetical gene supported by AF131741
NC_001807	ND1	mitochondrially encoded NADH dehydrogenase 1
AF221520	PRKCBP2	oligodendrocyte lineage transcription factor 2
NM_005794	HEP27	dehydrogenase/reductase (SDR family) member 2
Contig14836_RC		
Contig23399_RC		
AB033065	KIAA1239	KIAA1239
Contig29617_RC		
Contig50396_RC		
Contig16202_RC		
Contig24609_RC		
NM_002809	PSMD3	proteasome (prosome, macropain) 26S subunit, non-ATPase, 3
NM_003147	SSX2	synovial sarcoma, X breakpoint 2
NM_003283	TNNT1	troponin T type 1 (skeletal, slow)
Contig7755_RC	MGC5395	AHNAK nucleoprotein
Contig46304_RC		

List of Author Publications Related to PhD Thesis

A) Journal Papers

1. **M. E. Blazadonakis** and M. Zervakis, "Wrapper Filtering Criteria via Linear Neuron and Kernel Approaches". *Computers in Biology and Medicine*, to appear.

2. **M. E. Blazadonakis** and M. Zervakis, "The Linear Neuron as Marker Selector and Clinical Predictor", *Computer Methods and Programs in Biomedicine*, Vol 91 (2008) pp 22-35.

Under submission

3. M. Zervakis, **M. E. Blazadonakis**, D. Kafetzopoulos, V. Danilatou, and M. Tsiknakis "Decision-Support Based on Microarray Analysis: A Critical Comparison of Methods", *under submission in BMC-Bioinformatics.*

4. **M. E. Blazadonakis** and M. Zervakis, Integrating Biological Knowledge for Marker Gene Selection in Breast Cancer, *under submission in BMC-Cancer.*

B) Book Chapters

1. **M. E. Blazadonakis** and M.Zervakis, "Support Vector Machines and Neural Networks as Marker Selectors in Cancer Gene Analysis", book chapter in "Intelligent Techniques and Tools for Novel System Architectures", Series of Computational Intelligence, Springer Verlag, to be published end of July 2008.

C) Conference Proceedings

1. **M. E. Blazadonakis** and M. Zervakis, "Comparison and Cross Platform Evaluation of Genomic Signatures in Breast Cancer", *submitted on the 7th International Conference of Machine Learning and Applications (ICMLA 2008), 11-13 Dec. 2008, San Diego, California*.

2. M. Zervakis, and **M. E. Blazadonakis**, A. Banti, D. Kafetzopoulos, V. Danilatou, and M. Tsiknakis, "Performance Validation of Microarray Analysis Methods", *submitted on the 8th IEEE International Conference on Bioinformatics and Biomedical Engineering (BIBE 2008), 8-10 Oct. 2008, Athens*.

3. **M. E. Blazadonakis** and M. Zervakis, "Revealing Significant Biological Knowledge via Gene Ontologies and Pathways", *2008 International Conference in Biomedical Engineering and Informatics (BMEI 2008)*. Sanya, Hainan , China, May 2008**,** pp 169-172.

4. **M. E. Blazadonakis** and M. Zervakis, "Polynomial and RBF Kernels as Marker Selection Tools – A Breast Cancer Case Study", *"Proceedings of the 6th International Conference on Machine Learning and Applications"*. Cincinnati Ohio 13-15 Dec. 2007, pp. 488-493.

5. **M. E. Blazadonakis**, A. Perperoglou and M. Zervakis, "Using a Single Neuron as a Marker Selector – A Breast Cancer Case Study", *"Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society"*. Lyon France Aug. 23-26 2007, pp. 4219-4222.

6. **M E. Blazadonakis** and M Zervakis, "Improving Filter Methods Through Wrapper Approaches - A Breast Cancer Case Study", *"Proceedings of the 3rd International Conference on Computational Intelligence in Medicine"*. Plymouth England 25-27 July 2007.

7. **M. E. Blazadonakis** and M. Zervakis, "Support Vector Machines and Neural Networks as Marker Selectors for Cancer Gene Analysis", *Proceeding of 3rd International IEEE Conference on Intelligent Systems*. London Sept. 4-6 2006, pp. 626-630.

8. M. Kounelakis, **M. E. Blazadonakis**, M. Zervakis, X. Kotsiakis , 2005, The Impact Of Bioinformatics In Clinical Practices On Human Cancer, *Proceedings of the European Conference on Emergent Aspects in Clinical Data Analysis (EACDA 2005), September 28-30, 2005, Pisa, Italy.*