Supervised classification and spatial dependency analysis in human cancer using high throughput data

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof. dr. ir. J.T. Fokkema,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op maandag 14 april 2008 om 10.00 uur
door

Carmen LAI

Ingegnere Elettronico, Universita' degli Studi di Cagliari
geboren te Nuoro, Sardegna, Italy.

Dit proefschrift is goedgekeurd door de promotor:

Prof. dr. ir. M.J.T. Reinders

Samenstelling promotiecommissie:

Rector Magnificus	voorzitter
Prof. dr. ir. M.J.T. Reinders	Technische Universiteit Delft, promotor
Dr. ir. L.F.A. Wessels	Nethelands Cancer Institute, Amsterdam
Prof. dr. M.J.van de Vijver	Universiteit van Amsterdam
Prof. dr. ir. T. Young	Technische Universiteit Delft
Prof. dr. ir. A.H.C. van Kampen	Universiteit van Amsterdam
Prof. dr. J.N. Kok	Universiteit Leiden
Prof. dr. ir. Y. Moreau	Katholieke Universiteit Leuven, Belgium
Prof. dr. ir. J. Biemond	Technische Universiteit Delft, reservelid

The work described in this thesis was financially supported by the Cancer Genomics Center and was partially carried out at The Netherlands Cancer Institute - Antoni van Leeuwenhoek Ziekenhuis (NKI-AVL).

ISBN 978-90-9022906-5

Cover: Dot-painting by Carmen Lai

Chapters 2, 6 and 7: Copyright © 2008 by C. Lai
Chapter 3: Copyright © 2006 by Elsever
Chapter 4: Copyright © 2006 by BioMed Central
Chapter 5: Copyright © 2007 by BioMed Central

All rights reserved. No part of the material protected by this copyright notice may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without the prior permission of the copyright owner.

Author e-mail: `carmen@prsysdesign.net`

Per Pavel,

per i miei genitori,

per nonna Zelinda

Table of Contents

1 Introduction 1
- 1.1 Background 1
- 1.2 Scope 4
- 1.3 Thesis outline 7

Part I 9

2 A study on univariate gene selection for classification of gene expression datasets: possible improvements on a state of the art method 11
- 2.1 Introduction 12
- 2.2 Methods 12
- 2.3 Experimental results 17
- 2.4 Conclusions 23

3 Random Subspace Method for multivariate feature selection 25
- 3.1 Introduction 26
- 3.2 Feature selection techniques 27
- 3.3 Datasets 31
- 3.4 Experimental results 32
- 3.5 Conclusions 37

4 A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets 39
- 4.1 Background 40
- 4.2 Methods 42
- 4.3 Results 47
- 4.4 Conclusions 49

Part II 51

5 SIRAC: Supervised Identification of Regions of Aberration in aCGH datasets 53
- 5.1 Background 54
- 5.2 Algorithm description 56
- 5.3 Experimental results 59
- 5.4 Discussion and conclusions 67

6 Integration of DNA copy number alterations and prognostic gene signatures to predict prognosis of patients with breast cancer 71
6.1 Introduction . . . 72
6.2 Material and Methods . . . 73
6.3 Results . . . 76
6.4 Conclusions . . . 85

7 A genome-wide search for spatial organization of copy-number and expression dependencies 87
7.1 Introduction . . . 88
7.2 Method description . . . 90
7.3 Results and Discussion . . . 92
7.4 Conclusions . . . 103

8 Conclusions 105
8.1 Concluding remarks . . . 105
8.2 Open issues . . . 108

Afterword 111

A Genome and gene array description 113

References 117

Summary 135

Samenvatting 137

Acknowledgments 139

Publications 143

Curriculum Vitae 145

1
Introduction

1.1 Background

Cancer consist of cells of the body which proliferate in an uncontrolled fashion. Cancer cells may develop the ability to leave their tissue of origin and survive in other tissue types, causing metastases [Albe 02]. The diagnosis and classification of cancer is a necessary step towards the treatment of the disease. Conventional methods used in the clinics are based on clinical, pathological and molecular parameters [Thie 06, Char 05]. The clinical parameters include the age of the patient, and the stage of the tumor, which describes the extension of the tumor locally or at a distance from the primary site (e.g. in case of metastases). Pathological parameters include the size of the tumor, the lymph node status and the grading, which reports the morphology and proliferative capacity of the primary tumor. Molecular markers are determined mostly by immunohistochemistry methods, examples for breast cancer are the presence of estrogen and progesterone receptors. Unfortunately, the conventional methods are not fully capable of precisely defining prognosis and predicting response to therapy [Thie 06, Char 05]. Moreover, human expertise is required, in the person of well-trained and experienced clinicians and pathologists. This is not a simple demand, also considering the increasing number of tests that doctors are requested to perform. An automated system could support the clinics with e.g. a second opinion, or tools for training pathologists.

The advent of high throughput biomolecular measurements, such as gene expression arrays, allows a close look at the molecular mechanisms of diseases. A gene expression array measures the expression of thousands of genes simultaneously (Appendix A). It provides detailed genomic information that may help to detect the heterogeneity in an otherwise homogeneous patient group. That is, often patients share similar clinical parameters, but still exhibit diverse survival or treatment responses. Gene expression datasets have raised a new range of possibilities and questions such as: Can array technology assist pathologists in tumor stratification, i.e. detecting homogeneous subtypes

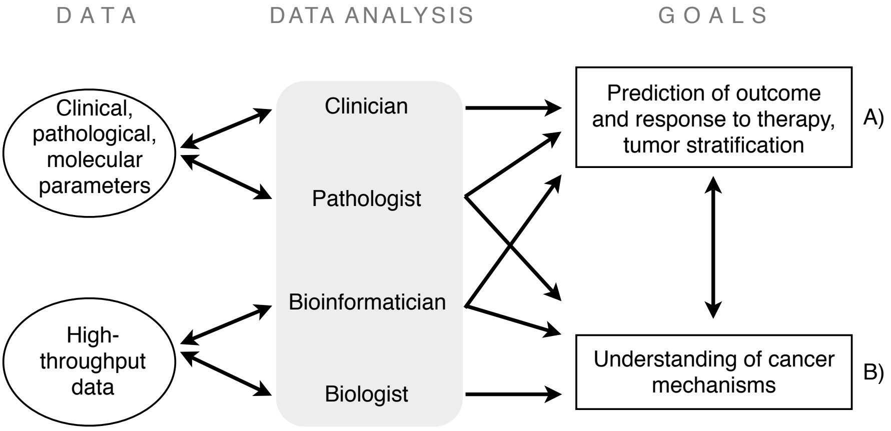

DATA
DATA ANALYSIS
GOALS
Clinical, pathological, molecular parameters
High-throughput data
Clinician
Pathologist
Bioinformatician
Biologist
Prediction of outcome and response to therapy, tumor stratification
A)
Understanding of cancer mechanisms
B)

Figure 1.1: Schema of major objectives, data types and people involved in cancer research and practice.

within the same disease? Can tumors that metastasize be distinguished from the ones that do not? Can the prediction of prognosis and response to therapy be improved?

Figure 1.1 depicts the data, people, and goals engaged in cancer research and practice nowadays. Generally speaking there are two main goals. First, the construction of methodologies that can predict outcome and response to therapy. This can be supported by reaching a stratification of the cancer types in seemingly homogeneous groups that are caused by the same mechanism and share the same clinical behavior. This is not only a research topic in itself but also a clinically relevant demand. A second major goal of cancer research is to improve our understanding on how cancer evolves. This would advance the development of better drugs, targeting the discovered biological mechanism, and improving patient health care. The new high-throughput datasets have opened new possibilities for addressing both of these goals. A new professional figure, the bioinformatician, is needed to develop complex data analysis techniques. The bioinformatician assists both the biologist in studying cancer mechanisms, and the clinician and pathologist in predicting outcome and response to therapy. Note that the two goals are not disconnected, as the progress towards one would help the other. For example, a better understanding of cancer mechanisms may lead to a refined stratification of samples that share the same clinical outcome. In the same way, the people involved form a multidisciplinary team that shares a wide range of expertise.

In the past years, lots of research has been performed concerning data analysis tools. First studies with gene expression data have used clustering algorithms to group patients into categories that shared common biological features [Pero 99, Ben 99]. This is an unsupervised way of analyzing the data, since no prior knowledge about the patients is used. No class labels assigning the patients to homogeneous group are exploited, but the classes are deduced from the results of the clustering algorithm [Spee 03, Jian]. Another way of approaching the problem of outcome or response to therapy prediction is to employ a supervised classification procedure. This implies to learn from examples (the given array data and the class labels) the patterns of the classes of interest defined by the labels assigned to the patients. A statistical model, called a classifier, is built (trained) to discriminate the classes of interest. This model should be able to generalize to data unseen during the training process, i.e. to new patients. In this way a new patient will be classified into one of the categories of interest, e.g. aggressive/non-aggressive cancer, or a tumor that will/will not respond to a treatment X.

Learning the classification model implies estimating its parameters from the data. A major difficulty encountered in the data analysis of expression arrays is the so called *small sample size* problem [Duin 95, Jain 97, Raud 91, Brag 07]. This arises from the heavy imbalance between the number of patients and the vast number of genes. Consequently, the estimated model may over-fit the data resulting in loss of generalization power.

To cope with the small sample size problem a reduction in the number of genes (dimensionality) is beneficial. This can be achieved in two main ways. One solution is based on feature extraction, which identifies a smaller number of dimensions (metagenes) than the original number of genes on the array, by combining information from all genes. Examples of such methods are: Principal Component Analysis [Bicc 03, Yeun 01], that projects the genes in the directions of maximal variance; or Independent Component Analysis [Lieb 02, Lee 03, Capo 06], which extends PCA to non-Gaussian data. Still all genes are required to classify a new samples. Another procedure to reduce the number of genes involves the selection of a small, yet informative subset. The gene subset is optimized according to a criterion, e.g. searching criteria such as the t-test, or the performance of the classifier itself [Tsam 03, Koha 97]. Due to the large number of genes, an exhaustive search strategy is not feasible. Several suboptimal searches have been proposed. Backward selection starts from a complete set of genes removing redundant or uninformative features according to a selection criterion, e.g. the classification performance of a Support Vector Machine (SVM) [Fure 00, West 00, Rako 03]. The forward feature selection starts with one gene and iteratively searches the informative genes amongst all available ones. A widely used search approach proposes individual gene selection based on univariate ranking according to a criterion [Golu 99, Ben 00, Veer 02, Vijv 02, Khan 01], or based on Markov blanket filtering [Xing 01]. Other search strategies use genetic algorithms [Li 01, Kiku 03], random search [Xion 01b], or pair-wise comparisons [Bo 02, Gema 04, Xu 05].

The goal of a feature selection method is to find an informative representation, which would increase the classification performance. Several studies have proposed signatures, i.e. small gene lists, for different purposes: to predict prognosis [Veer 02, Vijv 02, Chan 05a, Rama 03, Wang 05b, Mill 05, Cart 06], to improve cancer stratification in classes which share the same phenotype, i.e. the observable characteristic of the cancer [Pero 00, Sorl 01, Sorl 03], and, more recently, pilot studies on response to therapy [Ayer 04, Chan 05b, Hann 05, Ma 04]. However, a validation on larger patient cohorts is still needed to test the generalization power of these signatures, in such a way that they can be trustfully adopted in the clinics.

Understanding the mechanism of cancer would allow to learn the molecular causes of the disease. Existing subtypes may be refined, or new ones may be determined in order to group the patients in homogeneous categories that share the same mechanisms of cancer development. Unfortunately, the gene signatures defined with the statistical analysis of microarrays cannot always be directly related to the underlying biology of the cancer. In order to achieve more insight into the cancer biology, expression arrays are complemented with other types of information. For example, we can measure nowadays DNA copy number variations [Pink 05], single nucleotide polymorphisms (SNPs) [Shas 03], transcription factors binding sites [Jeff 07], and protein levels [Buck 04]. Other knowledge about the genes can be included as well, e.g. in the form of accurate annotations of genes' characteristics and functions [Stek 03]. A recent drive towards integration of these different sources of information can be observed [Edgr 06, Alli 06, Buss 07]. The aim is a better understanding of the mechanisms behind cancer development, by combining the different aspects of the same process. This would allow a better diagnosis and patient health care, and will stimulate the creation of more effective and patient tailored therapies.

1.2 Scope

Part I: dependencies in gene expression datasets

The first part of the thesis focuses on Goal A in Figure 1.1, especially with classifier building for outcome prediction. In order to reduce the dimensionality of the original datasets, we have concentrated on gene selection procedures. Identifying a limited number of genes compared with the number of genes on the array has an additional benefit: it provides the biologists with a tractable number of variables to be evaluated in order to gain understanding of the cancer mechanisms. Moreover, a small number of genes, e.g. in the order of few tens, would allow cheaper tests to be used routinely in the clinic. Many studies have proposed gene signatures to obtain more accurate classification. However, the overlap between these signatures is very limited. For example, to predict the ability of a primary breast tumor to metastasize two signatures were proposed: a 70 gene signature from the Netherlands Cancer Institute [Veer 02], and a 76 gene signature from the Rotterdam Medical Center [Wang 05b]. Although the goal is the same, i.e. outcome prediction, the overlap between the two gene lists is only three genes. This motivated us to investigate several ways to increase the robustness of a signature, not only in terms of classification performances, but especially concerning the stability of the genes selected independent of the sample cohort.

The relevance of a gene can be evaluated either individually (univariate approaches), or in a multivariate manner. Univariate approaches are simple and fast, therefore appealing and popular [Golu 99, Ben 00, Tibs 02, Veer 02, Khan 01, Xing 01]. However, they assume that the genes are independent. Multivariate approaches, on the contrary, evaluate the relevance of the genes considering how they function as a group, taking into account their dependencies [Xion 01b, Bo 02, Guyo 02, Bhat 03]. Genes are known to interact with each other, e.g. Gene *a* produces a transcription factor that binds to Gene *b*, activating its transcription. Therefore, a model that allows for dependencies, may capture more complex interactions between genes. Several limitations, however, restrict the use of multivariate approaches. Firstly, due to the higher complexity, they are more prone to over-training, especially in small sample size problems. Secondly, they may be computationally expensive, which prevents them from being applied to large feature spaces. In this thesis, we have investigated new ways to perform outcome prediction and have compared state of the art gene selection and classification procedures in a rigorous framework. The need for this stems from the fact that the first wave of research concerning the classification of gene expression datasets inappropriately adopted machine learning procedures, resulting in suboptimal classifiers [Cho 03, Chow 01, Khan 01, Xing 01, Jaeg 03, Ding 03, Bhat 03, Guyo 02, Silv 05, Bo 02, Ben 00]. Later research pointed out serious shortcomings in both the design and the evaluation of gene expression classifiers due to procedural errors and inadequate data validation [Ambr 02, Wess 05]. Our major contribution is the first consistent evaluation study on univariate and multivariate selection techniques, in order to identify the strong and weak characteristics of both approaches.

Part II: dependencies between DNA copy number and expression data

The second part of the thesis concerns Goal B in Figure 1.1, i.e. the use of high-throughput data to learn about cancer biology. Supervised classification based on gene expression data provides the possibility to construct generalizing classifiers using gene subsets. However, in order to gain biological insight into the mechanisms of cancer, the statistical analysis of expression arrays alone is not sufficient. This motivated us to integrate different sources of information, in particular copy number alteration data, expression data, and the genomic location of these measurements.

We have focused on DNA copy number since the genomic alterations are important events in cancer development [Leng 98]. A tumor suppressor gene can be disabled by its physical loss, or similarly an oncogene may be over-expressed via the amplification of the region where it is located. The identification of chromosomal aberrations is, therefore, a powerful instrument to study cancer [Bert 03, Pink 05]. Especially when coupled with the gene expression data, it may guide the identification of key genes, since, for example, the likelihood of a gene being involved in cancer is larger if it is both amplified and over-expressed. Our driving questions has been whether there are genomic aberrations that define the classes of interest, and what the influence of these aberrations is on gene expression.

The current approaches combining copy number and gene expression data can be organized in two main groups. The first group of methods employs copy number data to detect chromosomal aberrations, and then uses expression measurements to identify the genes that are correlated with the corresponding aberrated DNA-probes [Adle 06, Chin 06, Hyma 02, Frid 06, Poll 02]. The second group utilizes the expression data to evaluate the genetic and epigenetic effects (expression alterations due to DNA modification or other effects such as DNA methylation). These findings are then validated with copy number information [Reya 05, Stra 06, Furg 05].

Our work can be placed in the first group of approaches, since we start from the copy number data, and investigate the spatial local dependency between DNA-probes. We have built a systematic search across the complete genome to identify the copy number aberrations specific to the problem under study, i.e. cancer stratification and clinical outcome. More precisely, we exploited the class labels in the first step to identify spatial regions of DNA copy number alteration that are correlated with the class outcome. Then, we determined which genes were present in these areas.

So far, only the spatial local dependencies between the DNA-probes and the expression of genes have been investigated. The final contribution of our research is the study of both the local and the genome-wide spatial relationships between DNA alteration and changes in gene activities. We have aimed at pinpointing genome-wide dependencies via the identification of the correlation between a chromosomal aberration in a region and the expression on other locations on the genome.

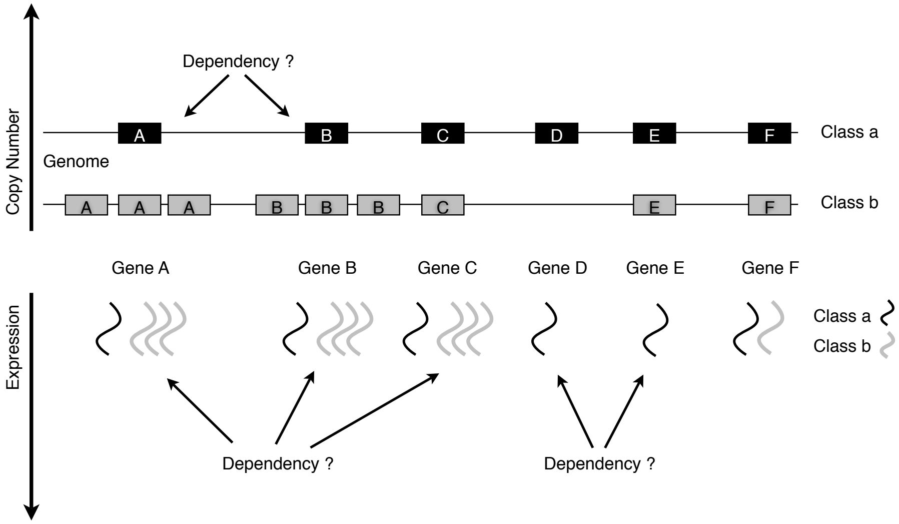


Figure 1.2: Schema of copy number and expression measurements for genes from A to F in two samples of different classes.

General theme

The underlying theme throughout this thesis, is the investigation of dependencies in gene expression and/or copy number measurements. Cancer is a complex disease. Therefore, one expects that multiple genes are affected simultaneously when a cell becomes tumorous. This in turn implies that the expression or the copy number of these genes will change in concordance with each other, or, in other words, that they are dependent on each other. Figure 1.2 schematizes the expression and the copy number of two samples from different classes, namely, Class *a* and *b* depicted in black and gray respectively. The hypothetical state of six genes, from A to F, is illustrated. In the upper part of Figure 1.2 the copy number for the 6 genes are depicted as filled rectangles along a line representing the genome location.. For example, Gene A shows a three copies in Class b, while there is only a single copy in Class a. The expression of the genes, in the form of mRNA transcripts, is depicted with curly segments in the lower part of Figure 1.2.

In Part I of this thesis, we focus on gene expression measurements only. The lower part of Figure 1.2 shows for example that only Gene F is non informative, since the same amount of mRNA transcript is produced in the samples of both classes. On the contrary, in Class b Genes A, B and C are over-expressed (i.e. there is three times more mRNA in Class b than in Class *a*), and Genes D and E are down-regulated (no mRNA is produced). One may hypothesize that the genes A,B,C or D,E are dependent. We introduce several ways to capture these possible dependencies. More specifically, since a multivariate gene selection strategy takes into account the dependencies between genes, we strive to apply multivariate techniques to analyze expression data. We investigate the role of univariate and multivariate gene selection algorithms, and extensively compared the two approaches.

Part II of this thesis studies the dependency between copy number alterations, exemplified e.g. by Genes A and B in the upper part of Figure 1.2, and the dependencies with their respective expression levels. Examples of dependency between the two data types are the amplification in Genes A and B, which produces their over-expression, or the deletion of Gene D, that determines the loss of the corresponding mRNA transcript. Note that other mechanisms (epigenetic effects) may produce cases such as Genes C and E, where the expression is not dependent on the copy number of the corresponding gene. We proposed a new statistical method to identify local spatial dependencies between copy number and gene expression measurements, and further searched for genome-wide dependencies.

1.3 Thesis outline

The thesis is divided into two parts. Part I deals with gene (or feature) selection techniques, and classification mainly applied to expression datasets. Part II of the thesis focuses on the integration of copy number and expression data, with the aim of studying correlations between copy number aberration and gene expression changes.

Part I: Chapter 2 addresses the problem of obtaining a robust gene signature. We have considered an existing method, the univariate gene selection, and have explored ways to improve it. In univariate gene selection, first, the genes are ranked individually according to a criterion that should asses their relevance in discriminating between the two classes of interest. Then the top k genes are selected to train the low complex Nearest Mean Classifier (NMC). The parameter k plays an important role and needs to be optimized. We have studied several ways to perform this optimization aiming both at an improved classification performance and at a small, yet robust, number of selected genes.

Chapter 3 presents the *Random Subspace Method* (RSM) that we have developed to perform feature selection in a multivariate manner. While univariate selection assesses the relevance of a feature on an individual basis, multivariate feature selection aims at identifying a number of features that, taken together, capture relevant information. The benefits of the RSM algorithm are illustrated on an artificial dataset, which provides ground truth information, and on a real dataset, that consists of autofluorescence spectra measured in the oral cavity of healthy and diseased patients. This work was published in *Pattern Recognition Letters* [Lai 06b].

In Chapter 4, we have performed an extensive comparison of several gene selection techniques, both univariate and multivariate. While many studies claimed good performance, the procedural errors made their results inconclusive. This motivated us to study in an unbiased protocol several state of the art techniques in order to understand the benefits and limitations of those techniques. This work was published in *BMC Bioinformatics* [Lai 06a].

Part II: Chapter 5, concentrates on the copy number data. We have developed an algorithm (SIRAC) that exploits spatial dependencies in order to identify regions of chromosomal aberrations, which are correlated with the classes of interest. In particular, the focus has been on the characterization of copy number aberrations in the cancer subtypes identified by Sorlie and Perou [Pero 00,Sorl 01,Sorl 03]. This work was published in *BMC Bioinformatics* [Lai 07].

Chapter 6 extensively describes the implications of the genome copy number alterations in breast cancer. Special attention is devoted to 68 samples selected from the NKI cohort [Vijv 02], for which both copy number and expression data were available. The regions of aberrations identified with SIRAC have been further investigated by analyzing the expression of the genes on the same genomic location. The objective has been to identify the genes that were affected by the copy number alterations and have major functional involvement in breast cancer development. This work will be submitted to *Cancer Research* [Horl ed].

In Chapter 7 our interest has been on the detection of causal spatial dependencies and interactions between copy number and expression alterations across the whole genome. An unsupervised extension of the SIRAC algorithm has been developed to highlight the patterns of correlation between the two data types. The new algorithm (IGDam, Identification of Genome-wide Dependencies between aCGH and mRNA data) extends the search for spatial dependencies from the one dimensional space of the copy number data to the two dimensions of the combined copy number and expression data.

A final discussion of the research is summarized in Chapter 8.

Part I

2

A study on univariate gene selection for classification of gene expression datasets: possible improvements on a state of the art method

This chapter addresses the problem of obtaining a robust gene signature. We have considered an existing method, the univariate gene selection, and have explored ways to improve it. In univariate gene selection, first, the genes are ranked individually according to a criterion that should asses their relevance in discriminating between the two classes of interest. Then the top k genes are selected to train the low complex Nearest Mean Classifier (NMC). The parameter k plays an important role and needs to be optimized. We have studied several ways to perform this optimization aiming both at an improved classification performance and at a small, yet robust, number of selected genes.

2.1 Introduction

Gene expression arrays enable the measurement of the activity levels of thousands of genes on a single microscope slide. An important application of this technology is the prediction of disease state of a patient based on a signature of the gene activities. Such a diagnostic signature is typically derived from a dataset consisting of the gene expression measurements of a series of patients. Since typically hundreds of patients and thousands of gene activities are measured, analysis of these data sets is a challenging manifestation of the small sample size problem in pattern recognition. The primary objective is to build a classifier which assigns a new sample as accurately as possible into one of the diagnostic categories, for example tumor/normal tissue, or benign/malignant tumor. A secondary objective is to find a small number of genes, i.e. a signature, which the diagnostic classifier employs as input, and which consequently carries the information relevant for the diagnostic task. This process of identifying the genes relevant to the classification task is known as feature selection.

Due to the high dimensionality of the feature space suboptimal strategies are needed to search for the most informative genes to add to an informative list. A widely used approach carries out informative gene selection by 1) performing a univariate ranking according to a single gene-based criterion [Golu 99, Ben 00, Veer 02, Vijv 02, Khan 01] and then adding the genes in the order of informativeness, or 2) ranking the genes based on Markov blanket filtering [Xing 01], and then adding the pairs in the ranked order. Xiong *et al.* [Xion 01b] and Bo *et al.* [Bo 02] suggested a pair-based method that evaluates the relevance of pairs of genes, and sequentially add pairs in the order of informativeness. Other approaches that aim at identifying informative genes are based on random searches of the gene population, also referred to as the Monte Carlo method [Xion 01b], and genetic algorithms [Li 01, Kiku 03].

We decided to focus on a univariate gene selection procedure, due to its popularity ([Golu 99, Ben 00, Veer 02, Vijv 02, Khan 01, Xing 01]) and efficiency. We refer to this method as *average individual ranking*, and describe our implementation in Section 2.2.1. We propose different possible extensions of the *average individual ranking*, such as gene or classifier combining (also in Section 2.2.1). In order to understand the behavior of the methods, we have designed an artificial dataset which provides ground truth information. The artificial dataset is described in detail in Section 2.2.2. Our experiments suggest that combining genes or classifiers strategies that we investigated do not yield an improvement over the simple *average individual ranking* method (Section 2.3.3). We hypothesize that the weak element is the individual gene selection procedure and focus on this question (Section 2.3.4). The further experiments reveal the inherent weakness of this selection strategy.

2.2 Methods

2.2.1 Gene selection and classification schemes

When designing a classification system, two steps need to be taken. The first one is a classifier training, and the second one is the estimation of a classifier performance. Due to the small number of samples, the cross-validation procedure is the preferred approach to estimate the classification error [Ambr 02, Koha 95]. In order to have an

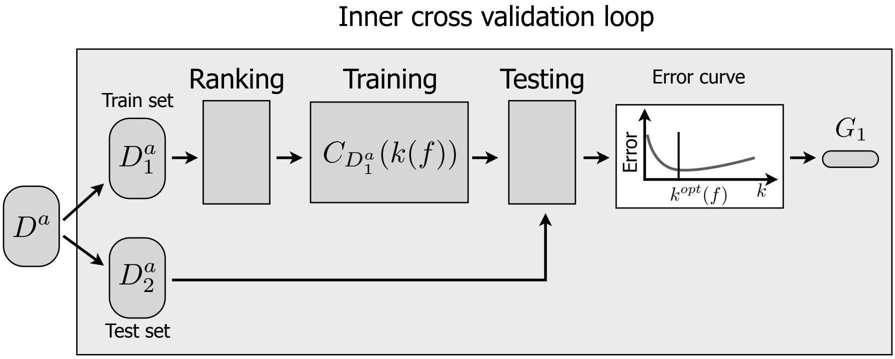
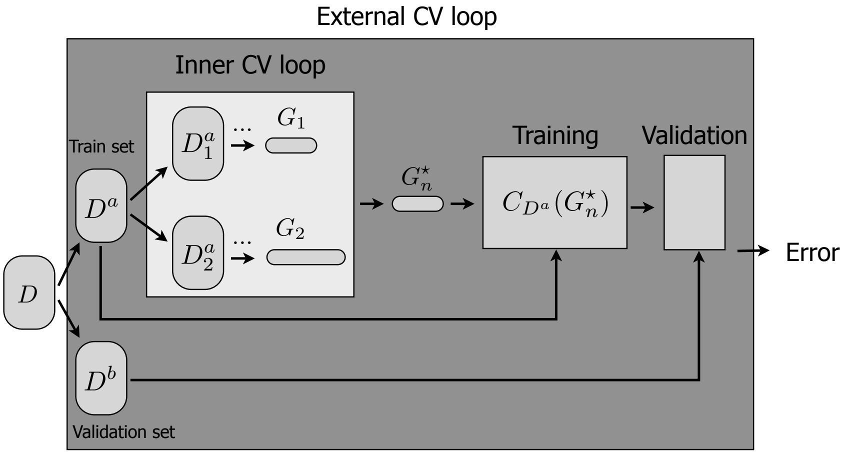


Figure 2.1: The inner loop of a double loop cross validation strategy in case of gene ranking according to a criterion.

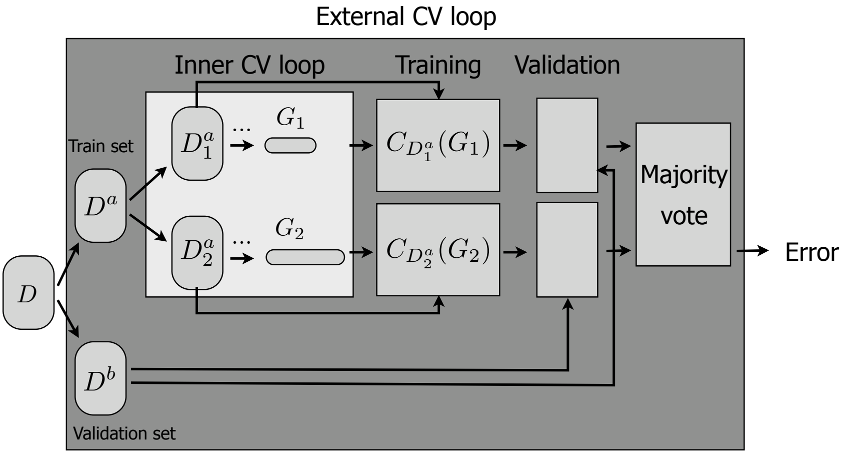
unbiased error estimate the two steps should be performed independently. Therefore, we employ a double loop of the cross-validation procedure [Wess 05]. In the inner loop the gene selection is performed, and the classifier is trained, while in the external loop the performance of the classifier is estimated.

Let us focus on the inner cross-validation loop, which is used to perform the gene selection. Figure 2.1 describes one fold of a two-fold cross-validation strategy. The data D^a is split into two parts D_1^a and D_2^a, used respectively for training and testing purposes. Based on a criterion, the informativeness of each gene in the training set D_1^a is evaluated individually. The genes are ranked accordingly, from the most to the least informative, we refer to the ordered list as R. A classifier is then trained, starting with the best gene, and is tested on the same genes in the independent test set D_2^a. The procedure repeatedly expands this set, with five genes at a time in the order established by the ranking R, until k_{max} genes are selected. The step of five genes at a time is choose only for computational reasons, instead of a smaller one gene step. Each time the features are added, the error is computed on the test set. As a result, we can plot the error of the classifier as a function of the number of genes used for classification purposes (error curve). The curve in the right part of Figure 2.1 illustrates the expected result. Typically this curve will show that a small number of genes gives large error rates, due to insufficient information. The (initial) addition of relevant genes lowers the error, reaching a minimum. Further addition of genes, however, degrades the classifier performance. Minimizing the error provides a selection of relevant genes ($k^{opt}(f)$ in Figure 2.1), i.e. a signature. In the second fold, the same procedure will be repeated using D_2^a as the training set and D_1^a as the test set.

The outcome of the inner loop is F sets of genes G_f's, where F is the number of the folds in the inner cross validation procedure, and G_f denotes a set of genes selected in the f^{th} fold ($G_f = \{g_i \,|\, i \in R(f), 1 \leq i \leq k^{opt}(f)\}$). Note that the gene sets can have different sizes and may be composed of different genes. The *average individual ranking* procedure estimates the size of the optimum signature by evaluating the average of the error curves of all folds. The gene size $k^\star$ that minimizes the average error, is considered to define the optimum size. Therefore, the first $k^\star$ genes of the ranked list computed based on the complete dataset D^a will be considered as the final signature.



(a) Combining gene sets. The F genesets outcome of the inner CV loop are combined into the set $G_n^\star$ which is used to train the classifier.



(b) Combining classifiers. A classifier is trained for each fold of the inner CV loop and the results are combined according to majority vote.

Figure 2.2: Double loop cross validation schema. The two proposed combining strategies are illustrate in a 2 fold cross validation case.

The *average individual ranking* procedure is a frequently used approach to combine the results obtained during the inner cross-validation loop. This approach uses cross-validation only to estimate the optimum size of the gene set. Since it is desirable to extract as much information as possible from the data, one could exploit the results of each fold, by combining either the information of each gene set (*combining gene sets*), or the classifiers trained in each fold (*combining classifiers*). The aim is to obtain in this way a more reliable, robust, and stable gene set and, consequently, a better classification result. Figure 2.2 presents a schema of the two implemented combining strategies.

Figure 2.2 (a) shows our proposal to combine gene sets. The inner loop provides F gene sets G_f of variable sizes and genes, with $F = 2$. Our hypothesis is that the informative genes will be often present in the different folds, while the uninformative ones have low probability of being selected multiple times. In order to count the number of times that a gene g is selected, we introduce the following function c:

$$c(g) = \sum_{f=1}^{F} \mathcal{I}(g \in G_f) \tag{2.1}$$

where $\mathcal{I}$ is an indicator function with value 1 if the argument is true and 0 otherwise. Three alternative signatures $G^\star$ can be built with:

$$G^\star_n = \{g_i | c(g_i) \geq n\}_{i=1..N} \tag{2.2}$$

where :

$n = 1$, i.e. a gene is present in at least 1 fold,

$n = \frac{1}{2}F$, i.e. a gene must be present in at least half of the F cross-validation folds,

$n = F$, i.e. a gene must be selected in all F folds.

The set $G^\star_1$ contains the genes that are selected at least once, i.e. all the genes present in the F sets. The sets $G^\star_{\frac{1}{2}F}$ and $G^\star_F$ contain the genes that were selected at least in half or in all the F gene sets, respectively. It is to be expected that the set $G^\star_F$ will contain the smallest number of genes of all three gene sets. Our hypothesis is that these genes are also the more relevant ones. Therefore, by combining genes, the inner loop of the cross-validation could provide valuable information. The gene set $G^\star_{\frac{1}{2}F}$ will increase with respect to $G^\star_F$, and the relevance of genes decreases. The set $G^\star_1$ will have the largest size. A classifier, $C^a_D(G^\star_n)$ in Figure 2.2 (a), will be re-trained on D^a and the classifier will be validated on the independent set D^b. The procedure will be repeated using D^b as training set and D^a as validation one. By averaging the classification errors of the different folds, we obtain an estimate of the classifier performance in the three cases, and therefore, we can evaluate the gene selection approach. Ultimately, the best selection procedure will be run on all available set D and the trained classifier will be the classifier proposed to be used in the clinics for the classification of new patients.

A second approach is to combine classifiers, instead of gene sets. Due to sampling effects, the performance of a single classifier can be very poor. Our aim is to compensate for this effect by combining all the F classifiers $C_{D^a_f}(k^{opt}(f))$. Figure 2.2 (b) illustrates this strategy, again in the two fold cross-validation case. For each fold of the inner cross-validation procedure, a classifier is trained using the best selected gene set of that fold, i.e. G_f. The F trained classifiers are applied on the independent test set D^b. The classification is based on majority vote, and the classification error is computed. Also in this case, the procedure will be repeated using D^b as the training set and D^a as the test one. By averaging the two errors, we obtain the classification error estimate of the combining classifier approach.

2.2.2 Artificial dataset

In order to investigate the challenges posed by typical gene expression data, we generate a comparable artificial problem. Our goal is not to simulate the real data set, as proposed by [Hube 03, Chil 02, Newt 01], but to have a controlled environment with roughly the same complexity, without having to deal with other sources of variation. Simple models proposed in the literature [Pudi 94, Jain 97, Trun 79] are based on normally distributed classes, but don't have comparable complexity with the gene expression datasets, since the feature size considered is not higher than 20. Therefore, a model with a larger complexity is needed. To study the effect of the small number of training samples on the univariate feature filtering procedure, we generate a dataset for which feature filtering (e.g. ranking) would be able to retrieve the correct feature sets, giving enough data. The artificial dataset can be summarized with a matrix $M \times N$ with M samples and N features. Each feature vector is sampled from the following two-class conditional densities:

$$p(X|\omega_1) \backsim N(\mu(i), 1) \qquad\qquad p(X|\omega_2) \backsim N(-\mu(i), 1) \tag{2.3}$$

where $\mu(i)$, is a function of the feature indicator i according to the following:

$$\mu(i) = \begin{cases} \mu_0(1 - \frac{i}{I}), & \text{if } 1 \leqslant i \leqslant I; \\ 0, & \text{if } I < i \leqslant N. \end{cases} \tag{2.4}$$

The most informative features are the ones with the smallest index value i. The distance between the means of both normal distributions, i.e. the class separation, linearly decreases from $2\mu_0$ for the first feature towards zero at the I-th feature. Therefore, the informativeness of a feature is defined by its index value i. All features with an index i larger than I are not informative, since the two normal distributions overlap completely. Note that each feature vector is generated independently, therefore the univariate ranking is a proper evaluation criterion (provided that there are enough training samples).

We are interested in a reliable gene selection procedure to reach a good classification performance. In order to evaluate the proposed methodologies, we would like to estimate the relevance of a selected gene set. Since we know that the lower the feature index i, the more informative gene it represents, we assign a score to each gene accordingly. The score function s_{g_i} of the generic gene g_i, is defined as follows:

$$s_{g_i} = \begin{cases} s_0(1 - \frac{i}{I}), & \text{if } 1 \leqslant i \leqslant I; \\ 0, & \text{if } I < i \leqslant N. \end{cases} \tag{2.5}$$

The value s_0 denotes the highest score which can be assigned. Each gene receives a score proportional to its informativeness, given by the position i in the original gene set. The higher the score, the more relevant the gene. A gene set $G^\star$ selected by one of the described methodologies will have a score $S = \frac{\sum_{i=1}^{k^*} s_{g_i}}{k^\star}$. Note that, even if the uninformative genes have score 0, their presence in the gene set will be penalized by the division for the total length $k^\star$ of the gene set. A comparison of the gene set score S of different methods will allow an evaluation of the gene retrieval power of the approaches.

2.3 Experimental results

Section 2.3.1 describes the experimental set-up, and Section 2.3.2 presents the real datasets and the parameter settings of the artificial dataset. In Section 2.3.3, the experimental results for both combining strategies are presented. Finally, further work on individual gene ranking is discussed in Section 2.3.4.

2.3.1 Experimental set up

As described in Figure 2.1, the first step in the training procedure is to estimate the informativeness of the genes individually. Several criteria may serve this purpose, such as Pearson correlation, Fisher criterion, or signal-to-noise ratio (SNR). Since for each feature both classes are normally distributed, we chose the SNR because it captures the difference between two normal distributions. Besides, the SNR is simple to compute and popular [Golu 99, Veer 02, Khan 01]. The SNR is defined as follows:

$$SNR = \frac{|m_1 - m_2|}{\sqrt{s_1^2 + s_2^2}}, \tag{2.6}$$

where m_1 and m_2 are the estimated means of the two classes and s_1 and s_2 are the estimates of the respective standard deviations. The higher the SNR the more informative the corresponding gene.

In the literature several signatures had appeared, proposing a limited number of genes to predict prognosis, e.g. the 70 and 76 gene signatures [Veer 02, Wang 05b]. The common opinion was that a small number of genes was not only sufficient to discriminate the classes of interest, but also beneficial to obtain an interpretable set of gene, and desirable to overcome the limitations of the small sample size. In line with this perspective, we fixed the maximum gene set size $k_{max}$to 100 in the inner loop of the cross validation.

The same classifier is used in each step of the methodologies described in Section 2.2.1. however, several types of classifiers are tested such as the nearest mean classifier (NMC), the Fisher Linear Discriminant (FLD), and the 5 Nearest Neighbour classifier (5-NN). For our artificial dataset, the nearest mean classifier is an optimal Bayes classifier, since the dataset is generated from independent features which have normal-based class conditional densities with equal variance (i.e. $cov(D_f^a/\omega_1) = cov(D_f^a/\omega_2) = I$). Thus we may expect that the low level of complexity of the classifier will not hamper the evaluation procedure. Additionally, the nearest mean classifier is a stable classifier that behaves favorably in a small sample size problems. The Fisher classifier is a more complex classifier than the NMC. It projects the data on a low dimensional space chosen by maximizing the ratio of the between-class and within-class scatter matrices of the dataset, and then classifies the samples in this space. The within-class scatter matrix is proportional to the pooled sample covariance matrix ($cov_{pooled} = \frac{1}{2}(cov(D_f^a/\omega_1) + cov(D_f^a/\omega_2))$), which is not singular if the number of samples is smaller than the number of features (dimensions) [Duda 01]. To avoid the inverse of an ill-conditioned covariance matrix a regularization parameter is needed [Skur 01]. Since in our experiments the size of the gene set may be larger than the number of samples, we use a regularized version in addition to a standard version of the Fisher classifier, i.e. $cov = cov + \lambda I$. Our preliminary work with the regularized FLD has shown little sensitivity over the regularization parameter λ, therefore we have here fixed the regularization parameter to 10. We use the 5-NN classifier as an example for a non parametric classifier. Since it is sensitive to local variation of the feature space, which is very noisy in our datasets, we do not expect improvements compared to the other classifiers. In addition to the Euclidean distance, the cosine distance is also used by the NMC and 5-NN classifiers.

As discussed in Section 2.2.1, cross-validation is a suitable procedure to estimate the classification error. For the sake of classification it is important to have a training set which is as large as possible. The number of cross-validation folds determines the sizes of the training and test set. However, we would like to have a test set large enough to be representative of the data. As a compromise we choose to use 10 fold cross-validation. This choice is also suggested by Ambroise *et al.* [Ambr 02] and Kohavi *et al.* [Koha 95].

As described in Figure 2.2, we propose to combine both gene sets and classifiers. In the *combining gene sets* strategy, applying 10-fold cross-validation, means that $F = 10$. The genes selected as relevant are the ones present in at least 1, 5, and 8 folds respectively, i.e $G_1^\star, G_5^\star$, or $G_8^\star$. Selecting $G_{10}^\star$ appeared to be too restrictive, i.e. the gene set was often composed of only a few genes, if not totally empty. Concerning the *combining classifier* method, the majority vote in the 10 fold setting may lead to ties if the sample to be labeled receives an equal number of vote for each class. To resolve this uncertainty, first the classifier with the higher error in the inner cross-validation loop is removed, and then the label is assigned according to the remaining nine classifiers.

In order to avoid the possible biases caused by a single realization of the artificial dataset, we repeat the experiment 10 times, using as datasets 10 different realizations from the same model.

2.3.2 Datasets used

Concerning the real datasets, the *Breast cancer*, and *Colon* datasets are used. The *Breast cancer* dataset consists of 145 lymph node negative breast carcinomas, 99 from patients that didn't had a metastasis within five years and 46 from patients that had developed metastasis within five years. The size of the feature (gene) space is 4919 [Veer 02]. Since it is public and widely used in the literature, the *Colon* dataset [Alon 99] is used for purpose of comparison. This dataset is composed of 40 normal healthy samples and 22 tumor samples in a 1908 dimensional feature space.

Concerning the artificial dataset, the first choice that has to be made concerns the dimensionality. We want to simulate real conditions, where the number of samples is much smaller than the number of features. Therefore, the number of the samples M is set to 100 which is comparable to the real datasets. The number of features is set to 1000 (N) mainly for computational reasons. The gene expression datasets are often unbalanced, due to the different availability of the samples within each of the classes. In the above mentioned real datasets, one class is roughly 30% of the number of samples (the other 70%). Therefore, we choose to preserve this imbalance also in the artificial dataset. The starting value μ_0, i.e. the class separability of the best features, is set to 0.25. For the index I, that limits the number of the informative features in the data, several values are tested, i.e. 100, 250, 500, and 1000. Additional experiments which motivate these settings are provided in Lai *et al.* [Lai 04]. Additionally large datasets, with the same values of I, are generated. They will be used to estimate the true error of the built classifiers. The dimensions of the test dataset are set to 1000 samples ×1000 features. To compare the retrieval power of the different approaches we assign the score s_g (see

Equation (2.5)), to each gene of all best sets. The value s_0 for the most informative gene is set to 10.

2.3.3 Results on real and artificial datasets

Table 2.1 summarizes the average score and the standard deviation of the score value S for the gene sets of the different methods: combining gene sets, i.e. using $G_8^\star, G_5^\star$, and $G_1^\star$ as gene sets, combining classifiers, and the *average individual ranking.* The columns presents the score, and its standard deviation, obtained with datasets containing different number of informative genes I. Results are normalized by dividing the mean and the standard deviation for the corresponding number of informative genes I present in the dataset. The gene sets were selected using the NMC classifier with Euclidean distance. Concerning the combining gene sets strategy, the highest score is reached while using the gene set $G_8^\star$, since only few and informative genes are selected. The score decreases while using the gene set $G_5^\star$, since the larger set increases the probability of having uninformative genes. The worst score is obtained when collecting the genes present in all folds, i.e. $G_1^\star$, since even more irrelevant genes are selected. The score in the combining classifier approach is computed by averaging the score of the gene sets used by all trained classifiers. Therefore, it is expected to be similar to the *average individual ranking* score. By looking at the score values, we would conclude that combining gene sets leads to a small and relevant gene set. The analysis of the classification errors, however, points out that this gene set doesn't lead to better results.

Table 2.1: Average and standard deviation of the score value S for the gene sets of the different combining strategies. The artificial datasets with different number of uninformative features are the input data, and the NMC is used as classifier. The results are normalized according to the number of informative genes I present in the artificial datasets.

Method	I = 100		I = 250		I = 500	
	Mean	Std	Mean	Std	Mean	Std
combining gene sets: $G_8^\star$	0.15	0.09	0.384	0.156	1.508	0.770
combining gene sets: $G_5^\star$	0.09	0.02	0.136	0.044	1.056	0.178
combining gene sets: $G_1^\star$	0.04	0.06	0.036	0.004	0.070	0.014
Combining classifiers	0.06	0.01	0.072	0.012	0.166	0.044
Average univariate ranking	0.06	0.02	0.064	0.012	0.092	0.032

Figure 2.3 (a) shows the average classification performance and the standard deviation of the combining strategies applied to different artificial datasets. The NMC with Euclidean distance is used as a classifier. The numerical value on the x-axis refers to the value of I used to generate the artificial dataset type, namely 100, 250, 500, and 1000. As discussed in Section 2.3.2, the classification performance is evaluated both in the external loop of the cross-validation and on the large dataset. The T on the x-axis refers to the error computed on the large validation set generated with the corresponding numerical value of the parameter I. Note that the standard deviation is much smaller than in the cross validation case. As expected, the more information in the dataset, i.e. the higher

the I, the higher the classification performances. Our main interest is in the combining strategies $G_8^\star$ and $G_5^\star$. Both of them reach, generally, a poor performance compared with the *average individual ranking*. The difference increases with the informativeness of the datasets. Surprisingly, the best results are obtained while using the larger gene set $G_1^\star$. Figure 2.3 (b) illustrate the average number of genes selected with the different strategies. The largest set is obtained with the $G_1^\star$ set. Interestingly, the higher the number of genes the higher the performance of the classifier.

The same experiments were carried out using the standard Fisher classifier and the 5-NN with cosine distance. For brevity the results are not shown, but the same behavior reached using the NMC, is observed. However, the classification performances are lower. This was expected, since the NMC is the classifier that better fits the artificial data. The same experiments were carried out with other values of the class separability μ_0, i.e. 0.15 and 0.35. The behavior of the score and of the classification performances are consistent with the findings presented here.

The same combining strategies were also applied to the *Breast cancer* dataset. To reach a more accurate estimate of the classification performances, we repeat the 10-fold cross-validation 10 times using different resamplings of the dataset. Figure 2.4 shows the average classification performances and the standard deviation. The results are grouped according to the classifier used to estimate the performances. As shown in the x-axis these classifiers are: Fisher classifier without and with regularization, NMC and 5-NN both with Euclidean and cosine distance measures. With the only exception of the Fisher with no regularization, the behavior follows the same pattern observed in the artificial dataset. The combining strategies do not improve on the performance reached using the *average individual ranking*. Also in this case, the best performance among the three set $\{G_8^\star, G_5^\star, G_1^\star\}$ is often reached while using the larger gene set $G_1^\star$. However, a high standard deviation is observed, and no single method is significantly better then the others. Only, the performances reached with the $G_8^\star$ set are generally significantly worse than the ones obtained with the other approaches.

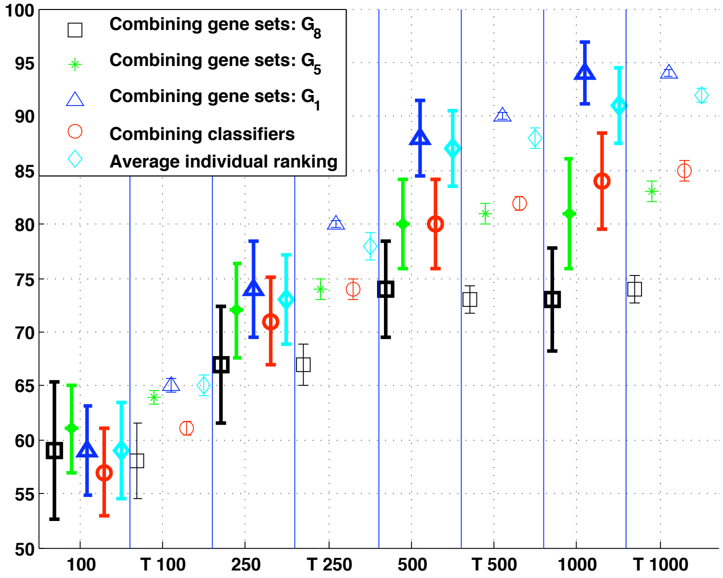
2.3.4 Discussion on individual ranking

The poor results of the combining strategy of either gene sets or classifiers, lead us to further investigation of the initial step, i.e. the feature ranking according to a criterion. If this step was unreliable, it would lead to a large number of uninformative genes in the selected set. This could explain the poor classification performances.

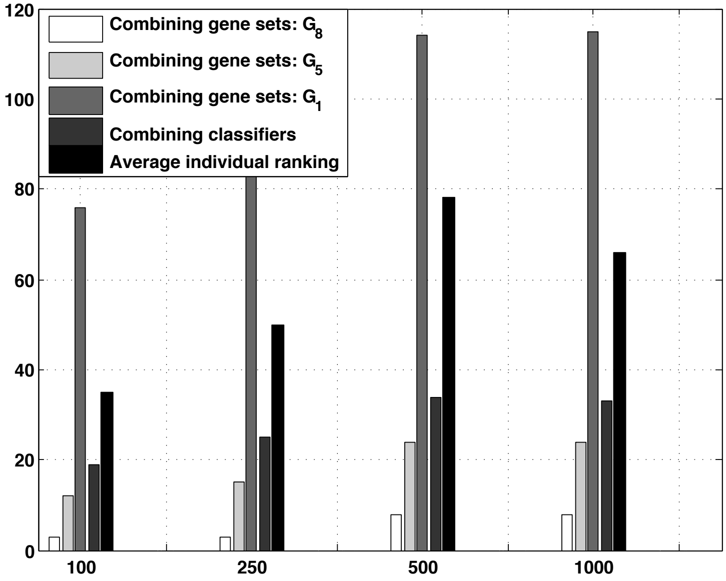
In order to investigate the feature ranking step we focus on the inner loop in Figure 2.1, which we consider now as the only loop of the cross-validation. We discuss the effects of the estimate of individual gene relevance on the gene selection procedure. More details can be found in [Lai 04].

The results are summarized in Figure 2.5. The classification error is now calculated in two ways. On the one hand the classification error, averaged across the 10 folds and the 10 artificially generated datasets, is plotted as a function of the number of features used to train the classifier. On the other hand, the classification error is calculated by testing the classifier on the large independent test set of 1000 samples. Due to the larger sample size, this test set allows the estimation of the true error of the classifier.

A first goal of this experiment is the evaluation of the methodology. In the artificial dataset the original index of the features corresponds to their amount of informativeness.



(a) Average classification performance and standard deviation



(b) Average number of genes used in the optimal gene set

Figure 2.3: Artificial dataset: results of 10 iterations of 10 fold cross-validation for several classifiers. The NMC with Euclidean distance is used as classifier. The numerical value on the x-axis refers to the value of I used to generate the artificial dataset type, namely 100, 250, 500, and 1000. The T in the x-axis refers to the error computed on the large test set generated with the corresponding numerical value of the parameter I.

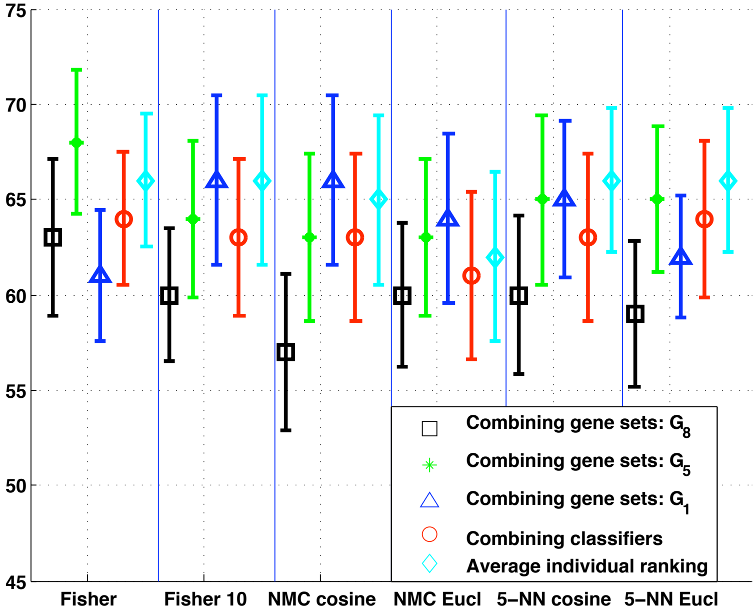


Figure 2.4: Breast cancer dataset: average classification performance and standard deviation of 10 iterations of 10 fold cross-validation for several classifiers. The different symbols represent the methodologies studied.

We can test the efficiency of the proposed methodology, directly using the original feature order, thus excluding the ranking step in Figure 2.1. The results while testing on the cross-validation test set and on the large test set, are plotted with the lines with × and □ respectively. Since the cross-validation error estimate of the method is closed to the *true* error, we can conclude that the cross-validation methodology is a good evaluation tool. The 10 fold cross-validation, however, has a larger variation on the error, as can be observed by the presence of local minima.

The necessity of a correct test procedure is emphasized. As pointed out by Ambroise *et al.* [Ambr 02], a bias is introduced if the estimation of feature relevance is made using all the data, since the test set in not independent anymore. The line called *rank bias* (see line with triangles in Figure 2.5) shows the cross-validation error while the features of the complete dataset were first ordered according to SNR and then the cross-validation procedure was run. This error is apparently very low, while the true error computed on the larger dataset (line with plus) is much higher. From this we conclude that all the steps taken to derive a classifier, i.e. gene ranking, selection and classifier training, must be performed only on the training set, keeping an independent test set aside. Otherwise we would have an error estimate that is positively biased and we would make wrong choices. In this example we would choose 200 as the optimal number of informative features, while the true error shows that better results are achieved with larger number of features.

The main goal of the experiments presented in Figure 2.5 is to evaluate the rank-

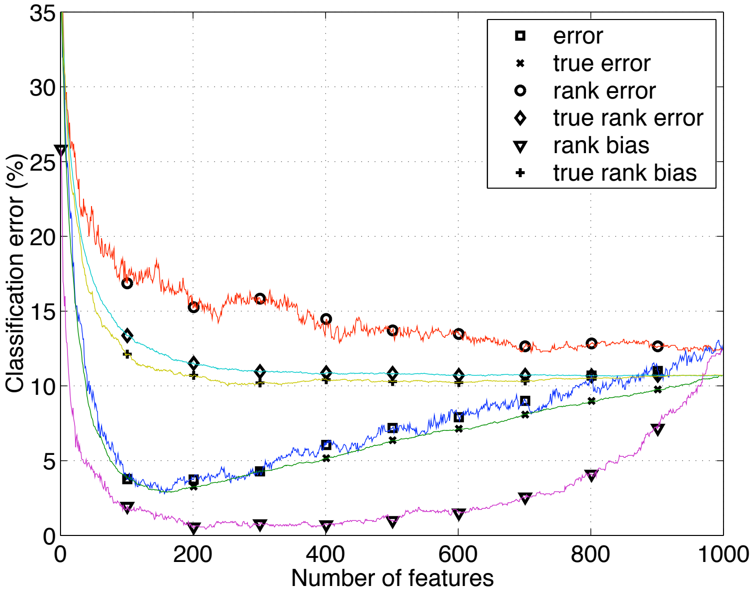


Figure 2.5: Average classification error as a function of the number of features (genes) used to train the classifier.

ing approach, i.e. first the features are ranked according to a criterion, and then the performance of a classifier, trained on an increasing number of features, is evaluated on an independent test set. The line with circles in Figure 2.5, which we call the *rank error*, represents the cross-validation error while applying this method. When comparing this error curve with the true error (i.e. features in original order, line with × in Figure 2.5) one can conclude that the ranking according to SNR is not able to identify the relevant features. Due to the small sample size, uninformative features have high SNR and consequently a high rank. The size of the feature set should increase to include the necessary informative features, but including more features also degrades the classifier. As a results no minimum is detected anymore. Clearly this does not fulfill the original target of deriving a small good signature. We can conclude that the estimate of the gene informativeness (the SNR) is very poor. This is due to the small sample effect since if we apply the same methodology to the large dataset of 10 000 samples (experiment not shown), the rank and true errors overlap.

2.4 Conclusions

We have investigated the univariate gene selection procedure, which is a popular approach for the selection of genes for classification purposes in expression data. We have studied possible improvements to the *average individual ranking* method with the aim of obtaining a more informative gene set. Several observations can be made. First, the combining rules proposed do not improve the classification results obtained with the base method (*average individual ranking*). The hypothesis of the existence of a small informative gene set does not seem to hold, as also suggested by the work of Ein-Dor *et al.* [Ein 05]. Using the *Breast cancer* dataset, they pointed out that the same classification performance can be achieved with many consecutive sets of 70 genes, not just with the top ranked 70 genes. This suggests that there is redundancy in the information carried by the gene expression, and, therefore, the larger the set the better. This is supported by our results, since $G_1^\star$ performs better than $G_5^\star$, which in turn outperforms $G_8^\star$.

In order to apply the combining strategies, it would probably be better to allow sets with larger sizes then the one used by us, i.e. setting $k_{max} > 100$. The fact that in our experimental results the gene sets were generally smaller that 100 can be explained by the presence of local minima. Combining larger sets would lead to larger best sets $G^\star$, and possibly to an increase in the classification performance.

A second observation concerns the ranking according to SNR. The experiments on the artificial dataset suggest that the small sample size hampers the ability of the univariate selection to precisely identify the informative features. In the real datasets the number of samples is limited, therefore any conclusion about the biological informativeness of a selected gene set should be taken with caution. Additional experiments are necessary to assess the biological relevancy of a gene set. An increase in the number of samples will also improve the identification of informative genes. In the future it is expected that larger cohorts will be available, thus allowing a re-evaluation of the selection procedures.

3

Random Subspace Method for multivariate feature selection

This chapter presents the Random Subspace Method (RSM) that we developed to perform feature selection in a multivariate manner. While univariate selection assesses the relevance of a feature on an individual basis, multivariate feature selection aims at identifying a number of features that, taken together, capture relevant information. The benefits of the RSM algorithm are illustrated on an artificial dataset, which provides ground truth information, and on a real dataset, that consists of autofluorescence spectra measured in the oral cavity of healthy and diseased patients. [1]

---

[1] This chapter was published in *Pattern Recognition Letters* [Lai 06b].

3.1 Introduction

In order to solve a classification task, the more features the better, since more information is present. However, addition of features beyond a certain point leads to a higher probability of error, as indicated in [Duda 01]. This behavior is known in pattern recognition as the curse of dimensionality [Duda 01, Jain 97, Trun 79, Raud 91], and it is caused by the finite number of samples.

Nowadays there are a growing number of domains that produce data with a large number of features, while the number of samples is limited. For example, the acquisition of spectral data, which give for a single sample the information across a large range of wavelengths. Other examples are the microarray datasets, that measure the gene activity of thousands of genes while the number of samples is limited to several hundreds, due to the high cost associated with the procedure and the sample availability. Assumptions often made in the literature are that many features are uninformative or noisy [Bo 02, Ambr 02, Xion 01b] and that features are likely to be correlated [Bo 02, Chow 01, Dudo 02].

Therefore, a feature selection strategy is needed to reduce the dimensionality of the feature space and to identify the relevant features to be used for a successful classification task. Feature selection algorithms can be divided in two categories: *filters* and *wrappers* [Koha 97]. Filter approaches evaluate the relevance of features based on a criterion indicative of the capacity of a feature to separate the classes, while wrapper approaches employ the classification algorithm that will be used to build the final classifier to judge feature quality. Both approaches involve combinatorial searches through the space of possible feature subsets. Several greedy procedures can be applied, such as forward or backward elimination, or less greedy approaches such as the more computationally demanding floating searches and genetic algorithms [Duda 01, Koha 97, Pudi 94, Li 01].

The relevance of a feature can be evaluated either individually (univariate approaches), or in a multivariate manner. Univariate approaches are simple and fast, therefore appealing and popular [Golu 99, Ben 00, Tibs 02, Veer 02, Khan 01, Xing 01]. However, they assume that the features are independent. Multivariate approaches, on the contrary, evaluate the relevance of the features considering how they function as a group, taking into account their dependencies [Xion 01b, Bo 02, Guyo 02, Bhat 03]. Several limitations however restrict the use of multivariate approaches. Firstly, they are prone to overtraining, especially in $p \gg n$ (many features and few samples) settings. Secondly, they may be computationally expensive, which prevents them from being applied to a large feature space.

The large number of features compared to the number of samples causes over-training when proper measures are not taken. In order to overcome this problem, we introduce a new multivariate approach for feature selection based on the Random Subspace Method (RSM) proposed by Ho [Ho 95, Ho 98] and studied further by Skurichina *et al.* [Skur 02]. Ho introduced the RSM to avoid overfitting on the training set while preserving the maximum accuracy when training decision tree classifiers. Skurichina *et al.*used the RSM in order to obtain weak classifiers to be combined in a second step of the classification process. We propose to use the RSM in order to better evaluate the informativeness of the features and, therefore, select a relevant feature subset on which to train a single classifier.

In this study, we apply a multivariate search technique on a subspace randomly selected from the original feature space. In this reduced feature space the multivariate feature selection may better handle the noise in the data and will consequently be able to retrieve the informative features. In order to take into account all the measured features of the dataset, the procedure is repeated many times. As a result several feature subsets are selected. These are combined into a final list of selected features, by ordering the features based on their relevance derived from their accuracy in the individuals runs. The final classifier can then be trained by using the final list of features. Our method can be applied in combination with existing classifiers and feature selection approaches, and is computationally feasible.

We compare our algorithm with other multivariate approaches, such as forward selection [Duda 01] and base-pair selection [Bo 02], as well as univariate techniques [Golu 99, Ben 00, Xing 01]. The comparison is performed on both a real dataset and on an artificial dataset which provides a controlled environment, and models the mentioned assumptions of correlation between features and the presence of a large number of uninformative features. The results show the importance of multivariate search techniques and the robustness and reliability of our new method.

The paper is organized as follows. Section 7.2 describes the feature selection algorithms and gives a detailed description of our Random Subspace Method for multivariate feature selection. The datasets used are presented in Section 4.2.3. Section 7.3 illustrates the experimental results of several multivariate and univariate feature selection techniques. Finally, the conclusions follow in Section 7.4.

3.2 Feature selection techniques

First, in Section 3.2.1, we briefly describe the univariate and multivariate techniques employed in comparison experiments. Then, in Section 3.2.2, the Random Subspace Method is introduced. Although the techniques are applied to a two class problem, they can be extended to a multi-class problem. A solution for multiclass problems could be to apply the complete technique as described here for all pairs of classes, or for one class against the others, and then use multiclass combiner strategies to create the final classifier. Examples can be found e.g. in [Allw 00, Diet 95, Tax 02]. Similarly the evaluation of a feature on the basis of the SNR for multiclass problem can be approached by computing the SNR criterion for all pairs of classes, and assign the minimum/maximum value to the features. Another alternative could be the comparison of the distribution of one class against the overall distribution (as been adopted by Tibshirani *et al.* [Tibs 02]), and again assign to the feature the minimum/maximum value.

3.2.1 Existing feature selection techniques

Univariate search technique

In the univariate approach the informativeness of each feature is evaluated individually, according to a criterion, such as the signal-to-noise ratio (SNR) [Golu 99, Chow 01] for a two class problem. The signal to noise ratio is defined as follows:

$$SNR = \frac{|m_1 - m_2|}{\sqrt{s_1^2 + s_2^2}}, \tag{3.1}$$

where m_1 and m_2 are the estimated means of the two classes and s_1 and s_2 are the estimates of the respective standard deviations. The higher the SNR the more informative the corresponding features, which are ranked accordingly, i.e. from the most to the least informative. This provides an ordered feature list L, and a cross-validation procedure is employed to judge the number of features from the top of this list to use.

Base-pair selection

The base-pair selection algorithm was proposed for microarray datasets by Bo *et al.* [Bo 02]. The relevance of features is judged by evaluating pairs of features. For each pair the data is first projected by the diagonal linear discriminant (DLD) onto a one-dimensional space. The score can then be computed by the t-statistic in this space. In our implementation we have used the Fisher discriminant and the SNR instead of the DLD and the t-statistic, respectively. This enables a better comparison with the other studied techniques. Both a full search and a less computationally demanding greedy search are investigated. The complete search evaluates all pairs and rank them in a list without repetition according to the scores. The computational complexity is a serious limitation of the method, therefore a faster greedy search is also employed. The features are first ranked according to the individual SNR. The best one is taken and then the method searches for a feature among all the remaining features, which together with the individual best one, obtains the highest score. This provides the first two features of the ordered list. From the remaining $2 - p$ features the best individual one is again taken and matched with the feature with which it achieves the highest score. This provides the second pair of features. By iterating the process the features are added, two at a time, until all of them are ordered.

Forward selection

Forward feature selection starts with the single most informative feature and iteratively adds the next most informative feature in a greedy fashion. Here, we select the features based on the criterion proposed by Bo *et al.* [Bo 02]. The first two features are obtained as the best pair described in the base-pair approach. For each of the $p - 2$ features, a third one is added to the best two features. The obtained three-dimensional feature space is projected onto the one dimensional space using again the Fisher discriminant, and the SNR criterion is computed. The best feature triplet will be the one that achieves the highest value of the SNR. By iterating the procedure, the features are added one by one, providing an ordered list of features of length L. Now the length of the list is limited to n. This upper limit stems from the fact that the Fisher classifier cannot be solved (without taking additional measures) if $L > n$.

Recursive Feature Elimination (RFE)

RFE is an iterative backward selection proposed by Guyon *et al.* [Guyo 02]. Initially a Support Vector Machine (SVM) classifier is trained with the full feature set. The quality of a feature is characterized by the weight that the SVM optimization assigns to that feature. A portion of the features with the smallest weights is removed at each iteration of the selection process. In order to build the ordered list of features length (L), the features that are removed are added at the bottom of the list. By iterating the procedure this list grows from the least to the most relevant feature. Note that the features are not considered individually, since their assigned weights are dependent on all the features considered during a given iteration.

Liknon

Recently Bhattacharyya *et al.* [Bhat 03, Grat 02] proposed a classifier called *Liknon* that simultaneously performs classification and relevant feature identification. Liknon is trained by optimizing a linear discriminant function with a penalty constraint via linear programming. This yields a sparse hyper-plane that is parameterized by a limited set of features (that are assigned non-zero weights by Liknon). By varying the influence of the penalty term the size of the selected features set can be varied.

3.2.2 Random Subspace Method

In case of a high dimensional feature space, it may be difficult for a multivariate search technique to identify the relevant features. In order to lower this risk we propose a new multivariate approach for feature selection based on the Random Subspace Method (RSM) introduced by Ho [Ho 95, Ho 98]. A multivariate search technique is applied on a subspace randomly selected from the original feature space. In this reduced space, the search technique can better handle the problem of dimensionality, and thus retrieve the informative features, since the number of samples per feature increases. In order to cover a large portion of the features in the dataset, we repeat the selection t times. As a result, t feature subsets are evaluated resulting in a weight associated with each feature for each of the selections. The weight is proportional to the relevance of the feature. We combine the results of all iterations in a final list of L features, ordered according to their relevance. Since this list is built upon the results of the more reliable feature evaluations that were performed in subspaces, the combined list is of better quality than a list constructed in the original complete feature space.

Our method can be applied together with different existing feature selection techniques. In the following, two algorithms are proposed using respectively Liknon and RFE as basal feature selection methods. Algorithm 1 applies in each of the t selections the Liknon classifier to a random feature set of size s. Only some of the s features will have a non-zero weight, that is proportional to the relevance of the feature in the subspace considered. The average of the weights across the number of times the feature was selected in the t subspaces is computed for each feature. Finally the features are sorted according to their computed average weights, then the top L features are selected, where L is optimized according to a cross-validation procedure. Again Liknon is employed as classifier in the greedily selected set of features.

Algorithm 2 applies RFE instead of Liknon on each of the t randomly selected subspaces. Also here all features of this subspace will be assigned a weight. In contrast to RSM-Liknon, the weights, however, are not comparable, since each weight is computed in a different feature space. Remember, RFE subsequently eliminates features on the basis of their weights. To establish the relevance of a feature we, therefore, have used the order in which the features are removed, i.e. the features that survive the RFE pruning the longest are the best. Instead of using the rank position itself as relevance indicator, we choose to quantize this. Only when the feature j is eliminated during the last l iterations of the RFE scheme it is indeed relevant and its score is incremented by

Algorithm 1 Random Subspace Method with Liknon (RSM-Liknon)

```
Input: training set $X$, label set $y$, number of selections $t$, size of the subspace $s$, matrix of zeros with $t$ rows and $p$ columns $Z = \mathbf{0}_{t \times p}$ to store the feature scores, vector $c = \mathbf{0}_{1 \times p}$ to count the number of times each feature is selected across the $t$ selections.
Repeat for $i = 1 : t$
    · Generate the random permutation index vector $p^i = perm(\{1, \ldots, p\})$.
    · Generate the index vector $v^i = \{p_1^i, p_2^i, \ldots, p_s^i\}$.
    · Extract the features indicated by $v^i$: $\tilde{X}^{v^i} \subset X$.
    · Train the Liknon classifier on the labeled dataset to obtain the weights:
      $w^i = \mathcal{C}(\tilde{X}^{v^i}, y)$, with $\mathcal{C}$ the Liknon classification rule applied on the subspace $v^i$ of the training set $X$.
    · Save the weights in the score matrix $Z$: $Z_{(i,v_j^i)} = w_j^i \quad \forall j = 1, \ldots, s$.
    · Update the counter $c : c_j = c_j + 1 \quad \forall j \in v^i$.
Compute the score vector $z$:
    $z_j = \begin{cases} \frac{\sum_{i=1}^{t} Z_{ij}}{c_j}, & \text{if } c_j \neq 0 \\ 0, & \text{otherwise.} \end{cases}$
Sort the list of features according to the score vector $z$.
Output: Ordered list of top $L$ features.
```

1. Finally, after the t selections, the score of each feature is determined, and the features are ordered accordingly. The top L features are selected, where L is optimized according to a cross-validation procedure using the Fisher classifier.

Algorithm 2 Random Subspace Method with RFE (RSM-RFE)

```
Input: training set $X$, label set $y$, number of selections $t$, size of the subspace $s$, threshold $l$ (if a feature is eliminated during the last $l$ iterations of RFE it is judged relevant), vector $z = \mathbf{0}_{1 \times p}$ to store the feature scores.
Repeat for $i = 1 : t$
    · Generate the random permutation index vector $p^i = perm(\{1, \ldots, p\})$.
    · Generate the index vector $v^i = \{p_1^i, p_2^i, \ldots, p_s^i\}$.
    · Extract the features indicated by $v^i$: $\tilde{X}^{v^i} \subset X$.
    · By applying the RFE procedure we obtain the order in which the
      features are removed: $L^i = \Phi(\tilde{X}^{v^i}, y)$. Here $L_k^i$ is the $k^{th}$ to last feature and $\Phi$ is the RFE procedure applied to the subspace $v^i$ of the training set $X$.
      to be removed.
    · Update the score vector $z : z_j = z_j + 1 \quad \forall j \in \{L_1^i, \ldots, L_l^i\}$.
Sort the list of features according to the score vector $z$.
Output: Ordered list of top $L$ features.
```

Two parameters need to be set in both RSM-RFE and RSM-Liknon: the subspace size s and the number of selections t. The smaller the subspace size s the faster the algorithm, but at the same time, the larger the chance of missing informative features or missing dependences between many features. Similarly the smaller the number of selections, t, the faster the algorithm, but the smaller the amount of data available for the evaluation of the feature occurrences when building the final feature set. In the case of RSM-RFE an extra choice regards the number l of features judged relevant. The smaller l the smaller the number of features judged relevant, but if the subset is too small good features may be missed. Large l may include irrelevant features, adding noisy dimensions to the subspace. The parameters are optimized empirically. For each parameter a set of possible values is chosen. Any available knowledge of the specific dataset adopted can guide the choice, e.g. if the number of informative features is expected to be low, a small value of the threshold l should be considered. The final choice of the best parameter combination is based on the cross-validation error on the training set, as described in the Section 3.4.1.

3.3 Datasets

3.3.1 The artificial dataset

In order to investigate the multivariate selection algorithms we generated an artificial dataset in which there is correlation between pairs of features. In this way, a pair of features is informative if considered together, and a multivariate selection strategy is necessary to find the truly informative features.

The informative features are generated in pairs: for each pair the samples are sampled from a Gaussian distribution with mean $\overline{\mu}_1 = \begin{bmatrix} 0 & 0 \end{bmatrix}^T$ for the first class and $\overline{\mu}_2 = \frac{2}{\sqrt{2}} \begin{bmatrix} d & 0 \end{bmatrix}^T$ for the second class. The covariance matrix, equal for both classes, is $\Sigma_1 = \Sigma_2 = \frac{1}{2} \begin{bmatrix} v+1 & v-1 \\ v-1 & v+1 \end{bmatrix}$. Pairs of correlated features are added until q informative features are generated. The remaining $p-q$ features are uninformative, i.e. the two classes are drawn from a spherical Gaussian distribution $N(\mathbf{0}, \frac{v}{\sqrt{2}} I)$, where I is the identity matrix.

The artificial dataset allows us to investigate the behavior of the methods in a controlled environment. The ground truth knowledge of the information present makes it possible to investigate the retrieval ability of the different techniques. In the following experiments we have set the number of the samples $n = 100$ and the number of features $p = 300$ to simulate a small sample size problem. In order to have a class overlap we set $d = 3$ and $v = \sqrt{40}$. Furthermore, we investigated the role of the informativeness in the dataset by varying q as follows: $q = [20, 50, 100, 150]$.

3.3.2 The spectra dataset

The algorithms are also compared on a small sample size version of the spectra dataset. The original dataset consists of autofluorescence spectra measured in multiple locations of the oral cavity [Veld 04]. There are two classes: 96 healthy volunteers and 155 patients with lesions in the oral cavity. After preprocessing, each spectrum consist of 199 bins (wavelengths). Although for the same person multiple spectra were acquired, we adopted one location only, in order to reduce the redundancy in the data. Therefore, the number of features is 199. We study the role of the sample size by using a dataset with $50, 100$ and 200 patients, balanced per class.

3.4 Experimental results

First the set up of the experiments are described. Then the experimental results are presented, followed by a discussion on the effect of a change in informativeness and number of samples in the datasets. The experiments are implemented in a Matlab environment using the PRTools [Duin 04] and PRExp [Pacl 05] toolboxes.

3.4.1 Experimental set up

As a criterion to judge the feature relevance in the one-dimensional space the SNR is adopted, due to its simplicity and popularity [Golu 99, Veer 02, Khan 01]. We use the Fisher criterion to project the data from multiple dimensions to a single dimension, as is required in both the forward and base-pair selection methods. We have used the Fisher classifier since it can exploit feature correlation. Also the Nearest Mean Classifier (NMC) with the cosine correlation as distance measure is employed in order to compare the results with other published approaches [Ein 05, Veer 02].

The Liknon classifier requires the optimization of the strength of the penalty term. The optimization of the strength of the penalty term is done beforehand in a 10 fold cross-validation procedure for a range of values $[10^{-1}, \ldots, 10^{3}]$. Since the average error was constant across this range, we chose 0.1 as a value for the strength of the penalty term because it selects a small feature subset.

The RSM-RFE approach was applied in combination with the Fisher classifier to allow for a better comparison with the other studied approaches which also employ the Fisher classifier. We chose to remove one feature per iteration. Concerning the optimization of the parameters, we varied the subspace size across $s = [40, 70, 100]$ and $s = [40, 70, 100, 130]$ for the artificial and real datasets, respectively. We selected as the best subset size on both datasets the value $s^* = 70$ that reached the lowest cross-validation error on the training set. The number of selections is set to $t^* = 130$. Smaller values, i.e. $t = 100$ and $t = 50$, were also tested. Although the performances are not always sensitive to this parameter, the larger t the more data is available for the evaluation of the feature relevance. For the RSM-RFE, several settings of the threshold l where experimentally investigated: $l = [3, 5, 8, 10, 15, 20]$ for the artificial dataset and $l = [3, 5, 10, 15]$ for the real dataset. The settings $l = 3$ and $l = 5$ respectively reached the lowest cross-validation error on the training set and are further presented in the experimental results.

When designing a classification system, two steps need to be taken. The first is the classifier training, and the second is the estimation of the classifier performance. Due to the small number of samples, a cross-validation procedure is a preferable approach to estimate the classification error. In order to have an unbiased error estimate the two steps should be performed independently [Ambr 02]. Therefore, we employ a double loop cross-validation procedure [Wess 05]. In the inner loop for each fold the feature selection is performed, giving the feature list L. The classifier is trained starting with the first 2 features of L till all features are used. The subset of L that shows the smallest classification error is selected. The selected feature sets from each fold are merged in a final subset $L^{\star}$, i.e. a list without duplication of all features present. In the external loop the performance of the classifier is estimated. This procedure ensures that the training and evaluation of the classifier are completely decoupled, as to prevent any bias in the

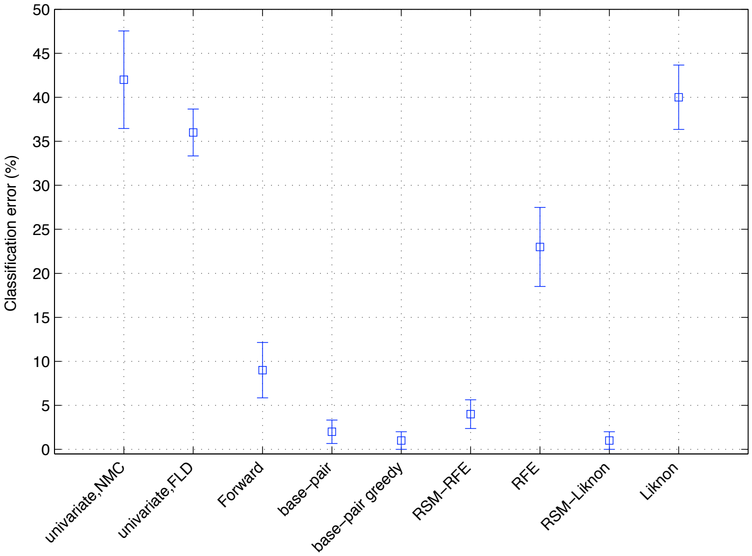


Figure 3.1: Artificial dataset with $n = 100, p = 300$ and informative features $q = 20$. Average classification error of the 10 fold cross-validation procedure for the different approaches.

performance evaluation. We chose to use 10 fold cross-validation for both the inner and external loop. All errors depicted in the figures are errors computed on the independent test set in the outer loop. This choice is also suggested by Kohavi [Koha 95].

3.4.2 Results

Figure 3.1 shows the behavior of the different methodologies on the artificial dataset with $n = 100$, $p = 300$, $q = 20$. The average classification error (computed in the outer loop), and its standard deviation are given. It is visible that the univariate approaches (univariate NMC, univariate FLD) perform the worst, while the base-pair approaches (both full and greedy search) reach the best performance. This is expected due to the fact that pairs of features are strongly correlated. Classical Liknon doesn't perform well. In Section 7.3.3 we will further explore this topic. The use of RSM with both Liknon and RFE improves the results dramatically, such that the results are comparable with the base-pair approaches. The same classification methodologies were also applied to different settings of the artificial dataset, i.e. a smaller number of informative features ($q = 10$) and sample sizes $n = [250, 50]$ (data not shown). The larger the number of informative features and/or samples the higher the performances of all classifiers. However, the different methodologies show the same behavior: univariate selection and classical Liknon perform poorly, while base-pair selection and RSM approaches give good results.

The knowledge of which features are informative in the artificial dataset allows us to study the retrieval capability of the different feature selection strategies. Figure 3.2 shows the number of informative features retrieved by the different methodologies as a function

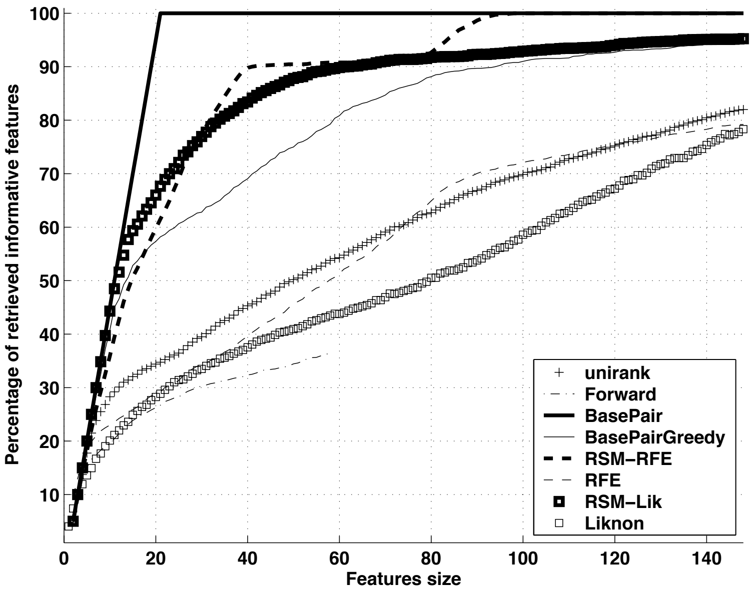


Figure 3.2: Relevant features of the artificial dataset retrieved by the different selection techniques.

of the subset size. The results are the average of the 100 folds of the inner 10 fold cross-validation loop. Surprisingly, both forward search and RFE retrieve more uninformative features than the univariate approach, since these selection methods should be capable of detecting the multivariate informative features. Apparently the small sample size hampers these methods severely. This will be tested in Section 7.3.3. The full search base-pair approach recovers all 20 informative features perfectly. RSM improves the number of informative features retrieved for both RFE and Liknon.

Figure 3.3 shows the average classification error on the spectra dataset with 100 samples and 199 features. The best performing approach is the RSM-liknon, which is statistically significantly better than all other methods but Liknon. RSM-RFE does not perform as well as expected, and others methods, such as the base-pair approach give surprisingly high error rate. To further investigate these aspects we looked at the error obtained in the inner cross-validation step for the RSM-Liknon, RSM-RFE and base-pair approaches.

Figure 3.4 displays the average error for both the artificial and the spectra datasets as a function of the first 150 features. In the techniques that employ the Fisher classifier, namely RSM-RFE and the base-pair approach, the 'peaking behavior' is visible, which occurs when the number of samples is comparable to the number of features. This phenomenon has been studied in [Skur 01, Frie 89, Dai 03]. The peaking behavior is not affecting the artificial dataset. This is because the Fisher classifier is applied to the selected subset of $L^\star$ genes. In all mentioned methods the size of $L^\star$ is not in the range of the peaking behavior, e.g. for the base-pair approach, the median value of $L^\star$ in the 10 folds is 10 features. Therefore, no further actions need to be taken. However, in the case of the spectra dataset the peaking effect is not negligible, since now the median

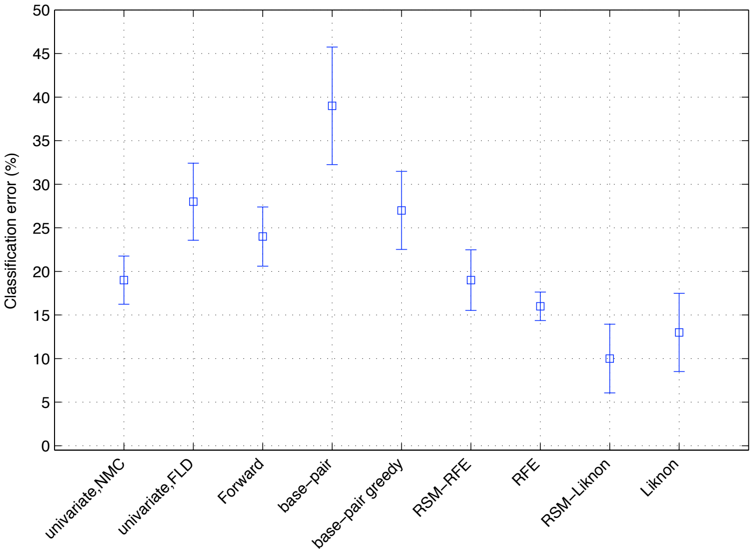


Figure 3.3: Spectra dataset with $n = 199$ and $p = 100$. Average classification error of the 10 fold cross-validation procedure for the different approaches.

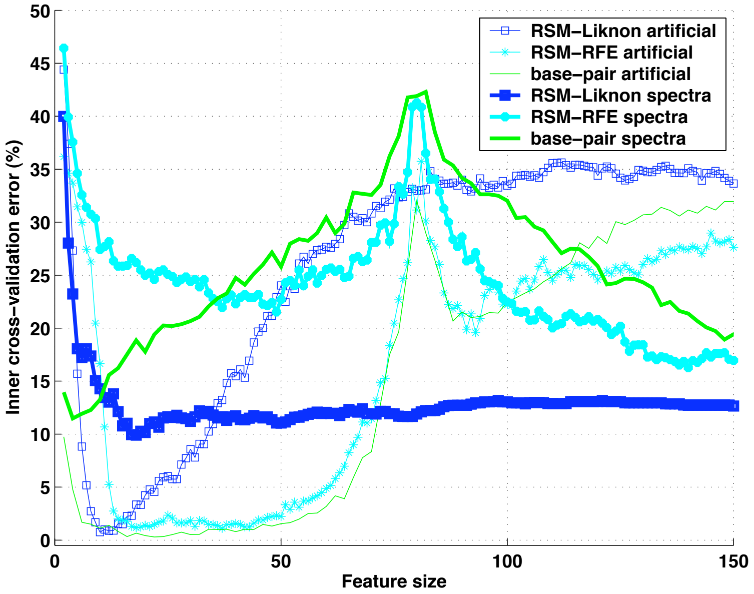


Figure 3.4: Average error of the inner cross-validation procedure as a function of the number of features for the artificial and the spectra datasets and three different approaches: RSM-Liknon, RSM-RFE and base-pair.

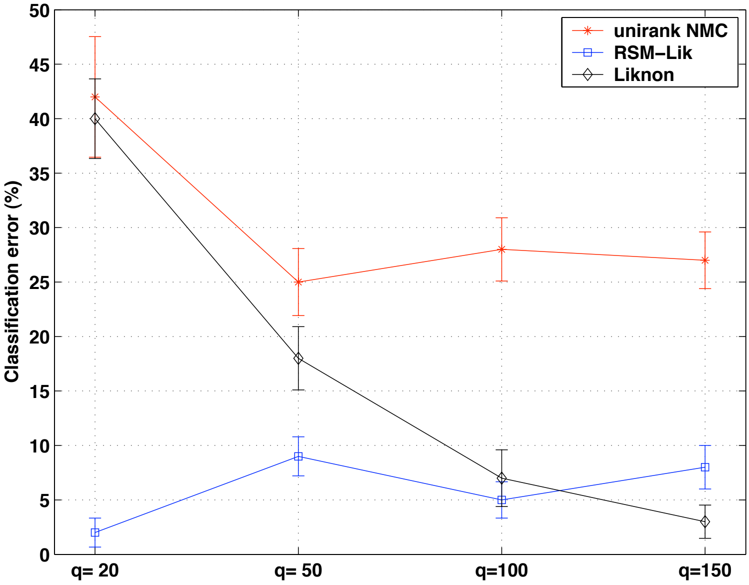


Figure 3.5: Classification error of three methodologies, namely univariate selection with NMC, RSM-Liknon, classical Liknon for different artificial datasets of the same size ($n = 100, p = 300$) but a varying number of informative features, i.e. $q = [20, 50, 100, 150]$.

value of $L^\star$ is in the peaking region, e.g. in the base-pair case the median $L^\star$ value in the 10 folds is 96 features. This explains the poor performances of the methods that use the Fisher classifier. Possible solutions to this problem are extensively described by Skurichina [Skur 01] and involve regularization by noise injection or by addition of redundant features. Unfortunately, these solutions are beyond the scope of this paper.

3.4.3 Discussion on the effect of informative features and samples

The Liknon classifier does not perform well on the artificial set, as shown in Figure 3.1. Our hypothesis is that this is due to the presence of too few informative features (20) relative to the total number of features. Therefore, we investigated how the number of informative features influences the classification performance of the three approaches, namely univariate with NMC, RSM-Liknon and classical Liknon.

Figure 3.5 shows the classification error of the mentioned methodologies when applied to four artificial datasets with $n = 100$, $p = 300$, and $q = [20, 50, 100, 150]$. The univariate approach with the NMC classifier benefits only when the number of informative features increased from 20 to 50. Thereafter it is hampered by the fact that it does not exploit the correlation between the features. RSM-Liknon is a stable methodology that proves to perform well also under difficult conditions, i.e. even when little information is present. Classical Liknon clearly decreases the error with an increase in the number of features, even up to the point where it outperforms RSM-Liknon significantly ($q = 150$). This behavior supports our hypothesis and exemplifies the need for the RSM technique.

The good performance achieved by Liknon on the spectra dataset would suggests

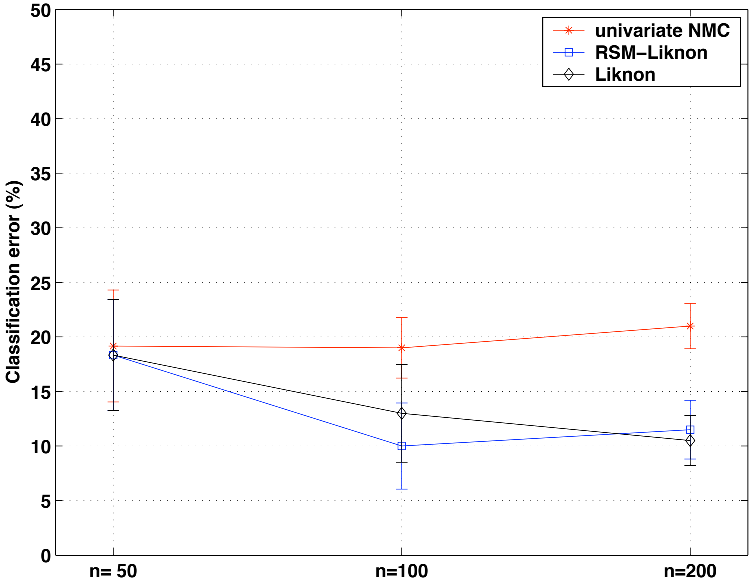


Figure 3.6: Classification error of three methodologies, namely univariate selection with NMC, RSM-Liknon, classical Liknon for the spectra dataset with different number of samples ($n = [50, 100, 200]$).

that this dataset contains many informative features. In order to evaluate the role of the sample size on the RSM approach, we consider the spectra dataset with different number of samples, i.e. $n = [50, 100, 200]$. Figure 3.6 shows the classification error of the following approaches: univariate with NMC, RSM-Liknon and classical Liknon. When the sample size is too small the multivariate search techniques are not able to retrieve any additional information, and the performance is comparable with the univariate approach. For an increased sample size the multivariate approaches are beneficial, and the RSM approach obtains the best performances. A further increase in the number of samples lower the need for the resampling in a subspace. In this case the classical method is not improved by the RSM.

3.5 Conclusions

In small sample size problems an important step is feature selection. This should lead to an informative feature space in which the classification task can be successfully performed.

In order to perform the selection, the informativeness of the features must be evaluated. We studied several approaches both univariate, where each feature is considered individually, and multivariate, where the criterion is dependent on multiple dimensions. A limitation of the multivariate approaches is the sensitivity to the dimensionality of the feature space, which often causes over-training. In order to overcome this difficulty we proposed a new method based on random subspace selection (RSM). A multivariate search technique is applied on a subspace randomly selected from the original feature space. In this reduced space the technique may better handle the noise in the data and retrieve the informative features.

We introduced an artificial dataset in order to have ground truth information. The artificial dataset models a small sample size dataset with both large number of uninformative features and a correlation between the informative ones, since both conditions are assumed to be present in real datasets. We tested our algorithm on a spectra dataset and illustrated the sensitivity to the sample size of the different studied approaches.

The results point out the importance of multivariate search techniques and the robustness and reliability of our new method. The univariate approach is outperformed by the multivariate methodologies. The RSM-RFE and the other methods that use the Fisher classifier are hampered by the sensitivity of the Fisher linear discriminant to the dimensionality of the feature space. Future study will be done on better overcoming these limitations. The RSM is a robust and a powerful approach for feature selection and classification especially in the small sample size conditions.

4

A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets

In this chapter we have performed an extensive comparison of several gene selection techniques, both univariate and multivariate. While many studies claimed good performance, the procedural errors made their results inconclusive. This motivated us to study in an unbiased protocol several state of the art techniques in order to understand the benefits and limitations of those techniques. [1]

---

[1] This chapter was published in *BMC Bioinformatics* [Lai 06a].

4.1 Background

Gene expression microarrays enable the measurement of the activity levels of thousands of genes on a single glass slide. The number of genes (features) is in the order of thousands while the number of arrays is usually limited to several hundreds, due to the high cost associated with the procedure and the sample availability. In classification tasks a reduction of the feature space is usually performed [Koha 97, Tsam 03]. On the one hand it decreases the complexity of the classification task and thus improves the classification performance [Ein 05, Ben 00, Blan 04, Chow 01, Stat 05a]. This is especially true when the classifiers employed are sensitive to noise. On the other hand it identifies relevant genes that can be potential biomarkers for the problem under study, and can be used in the clinic or for further studies, e.g. as targets for new types of therapies.

A widely used search strategy employs a criterion to evaluate the informativeness of each gene individually. We refer to this approach as univariate gene selection. Several criteria have been proposed in the literature, e.g. Golub *et al.* [Golu 99] introduced the signal-to-noise-ratio (SNR), also employed in [Jaeg 03, Bhat 03]. Bendor *et al.* [Ben 00] proposed the threshold number of misclassification (TNoM) score. Cho et al. [Cho 03] compared several criteria: Pearson and Spearman correlation, Euclidean and cosine distances, SNR, mutual information and information gain. The latter was also employed by [Xing 01]. Chow et al. [Chow 01] employed the median vote relevance (MVR), Naïve Bayes global relevance (NBGR), and the SNR, which they referred to as mean aggregate relevance (MAR). Dudoit *et al.* [Dudo 03] employed the t-statistic and the Wilcoxon statistic. In all cases, the genes are ranked individually according to the chosen criterion, from the most to the least informative. The ranking of the genes defines the collection of gene subsets that will be evaluated to find the most informative subset. More specifically, the first set to be evaluated consists of the most informative gene, the second set to be evaluated consists of the two most informative genes and the last set consists of the complete set of genes. The set with the highest score (classification performance or multivariate criterion) is then judged to be the most informative. For a set of p genes, this univariate search requires the evaluation of at most p gene sets.

Several multivariate search strategies have been proposed in the literature, all involving combinatorial searches through the space of possible feature subsets [Duda 01, Koha 97]. In contrast to the univariate approaches, which define the search path through the space of gene sets based on the univariate evaluation of genes, multivariate approaches define the search path based on the informativeness of a *group* of genes. Due to computational limitations, relatively simple approaches, such as greedy forward search strategies are often employed [Xion 01b, Blan 04]. More complex procedures such as floating searches [Pudi 94] and genetic algorithms have also been applied [Blan 04, Silv 05, Xion 01a, Li 01]. Guyon *et al.* [Guyo 02] employed an iterative, multivariate backward search called Recursive Feature Elimination (RFE). RFE employs a classifier (typically the Support Vector Machine (SVM)) to attach a weight to every gene in the starting set. Based on the assumption that the genes with the smallest weights are the least informative in the set, a predefined number of these genes are removed during each iteration, until no genes are left. The performance of the SVM determines the informativeness of the evaluated geneset. Bo *et al.* [Bo 02] introduced a multivariate search approach that performs a forward (greedy) search by adding genes judged to be informative when evaluated as a pair. Recently, Geman *et al.* [Gema 04, Xu 05] introduced the top-scoring pair, TSP method, which identifies a single pair of predictive genes. *Liknon* [Bhat 03, Grat 02] was proposed as an algorithm that simultaneously performs relevant gene identification and classification in a multivariate fashion.

The above mentioned univariate and multivariate search techniques have been presented as successfully performing the gene selection and classification tasks. The goal of this study is to validate this claim because a fair comparison of the published results is problematic due to several limitations. The most important limitation stems from the fact that the training and validation phases are not strictly separated, causing an 'information leak' from the training phase to the validation phase resulting in optimistically biased performances. This bias manifests itself in two forms. First, there is the most severe form identified by Ambroise *et al.* [Ambr 02]. (See also the erratum by Guyon [Guyo 03]). This bias results from determining the search path through gene subset space on the *complete* dataset (i.e. also on the validation set) and then performing a cross validation at each point on the search path to select the best subset. Although this bias is a well known phenomenon at this stage, a fairly large number of publications still carry this bias in their results [Cho 03, Chow 01, Khan 01, Xing 01, Jaeg 03, Ding 03, Bhat 03, Guyo 02, Silv 05]. The second form of bias is less severe, and was elaborately described in Wessels *et al.* [Wess 05]. See [Bo 02, Dudo 03, Ben 00] for instances of results where this form of bias is present. Typically, the training set is employed to generate a search path consisting of candidate gene sets, while the classification performance of a classifier trained on the training set and tested on the validation set is employed to evaluate the informativeness of each gene set. The results are presented as a set of (cross)validation performances - one for every geneset. The bias stems from the fact that the validation set is employed to pick the best performing gene subset from the series of evaluated sets. Since optimization of the gene subset is part of the training process, selection of the best gene subset should also be performed on the training set only. An unbiased protocol has been recently proposed by Statnikov *et al.* [Stat 05a] to perform model selection. Here, a nested cross-validation has been used to achieve both the optimization of the diagnostic model, such as the choice of the kernel type and the optimization parameter c of the SVM for example, and the performance estimate of the model. The protocol has been implemented in a system called GEMS [Stat 05b].

In addition to the raised concerns, the comparison between the results in available studies is difficult since the conclusions are frequently based on a small number of datasets, often the *Colon* [Alon 99] and *Leukemia* [Golu 99] datasets. See, for example [Blan 04, Bo 02, Ding 03, Guan 05, Guyo 02, Xing 01]. Sometimes even the datasets employed are judged by the authors themselves to be simple and linearly separable [Abul 05, Bhat 03, Silv 05, Xion 01a]. Therefore, no generally applicable conclusions can be drawn.

We perform a *fair* comparison of several frequently used search techniques, both multivariate and univariate, using an unbiased protocol described in [Wess 05]. Our conclusions are based on seven datasets, across different cancer types, platforms and diagnostic tasks. Surprisingly, the results show that the univariate selection of genes performs very well. It appears that the multivariate effects, which also influence classification performance, can not be easily detected given the limited sizes of the datasets.

4.2 Methods

4.2.1 Gene selection techniques

In this section we elaborate on the different univariate and multivariate selection strategies employed in this study. The approaches are cast in a general framework which highlights the choices made by the user, and facilitates direct qualitative comparison of these approaches.

Gene selection approaches are, in fact, optimization strategies, which input

1. D, a dataset consisting of n object-label pairs,

2. θ_Ω, a set of user-defined parameters which specify which type of classifier to use, and possible algorithm dependent choices such as the ranking criterion and

3. θ_Φ, another user-defined parameter defining the evaluation procedure (if cross-validation is employed, would specify the number of folds)

and which return the optimal value of a tunable parameter, ϕ, such that the gene set associated with ϕ^* (the optimal value of the tunable parameter) corresponds to the most informative gene set. During this optimization process, each gene selection approach is characterized by its own unique way to traverse and evaluate various gene sets. If we denote the mapping associated with selection approach A by Φ_A, this can be formally expressed in the following way:

$$\phi_A = \Phi_A(D, \theta_\Omega, \theta_\Phi). \tag{4.1}$$

For all the gene selection techniques described in this paper, the gene selection technique employs a classifier to evaluate the informativeness of the gene set associated with a given setting of ϕ. Given a dataset, D, and a setting of ϕ, the process which results in this classifier involves both a gene selection and classifier training step which could be separate or integrated. (This will be elaborated upon in the detailed descriptions of each technique). Formally, this process can be described as follows:

$$\omega_A = \Omega_A(D, \theta_\Omega, \phi_A), \tag{4.2}$$

where ω_A is the classifier trained on the geneset resulting from ϕ_A, θ_Ω represents the previously define parameters, and $\Omega_A(\cdot)$ is a mapping representing the training and selection process. During the optimization process, $\Phi_A(\cdot)$ repeatedly calls $\Omega_A(\cdot)$ with different settings for ϕ and employs the performance of ω_A as quality measure to guide the process. Upon completion of the optimization, the optimal classifier associated with the optimal gene set is given by:

$$\omega_A^* = \Omega_A(D, \theta_\Omega, \phi_A^*). \tag{4.3}$$

Univariate gene selection

In the univariate approach (U) the informativeness of each gene is evaluated individually, according to a criterion, such as the Pearson correlation, t-statistic or signal-to-noise ratio (SNR) [Cho 03, Dudo 03, Chow 01, Ben 00]. The genes are ranked accordingly, i.e. from the most to the least informative. This ranking defines a series of gene sets as well as the order in which they are subsequently evaluated. The first gene set is the best ranked gene, the second gene set the best two ranked genes, etc. The informativeness of each gene set is evaluated by estimating its cross-validation performance in combination with a particular classifier. As ranking criterion we adopt the SNR and the t-statistic. The former, due to its simplicity and popularity [Golu 99,Chow 01,Veer 02,Khan 01,Guyo 02], and the latter in order to enable a better comparison with [Bo 02]. For the evaluation of every gene set, we employ the Nearest Mean Classifier (NMC) with cosine correlation as distance measure and the Fisher classifier (FLD). The Fisher classifier [Fish 36,Duda 01] is a linear discriminant, it projects the data in a low dimensional space chosen by maximizing the ratio of the between-class and within-class scatter matrices of the dataset, and in this space classifies the samples. The within-class matrix is proportional to the pooled sample covariance matrix. In case of singularity of the matrix, which arises if the number of samples is smaller than the number of dimensions, the pseudo-inverse is used. In terms of the formal framework, θ_Ω represents the choice of univariate criterion (SNR or t-statistic) and classifier, while ϕ represents the desired number of genes selected. For $\phi = k$, this would correspond to the top k ranked genes. θ_Φ represents the type of cross validation to employ during the training process.

Multivariate gene selection

Base-pair selection (BP). The base-pair selection algorithm was proposed for microarray datasets by Bo *et al.* [Bo 02]. The informativeness of genes is judged by evaluating pairs of genes. For each pair the data is first projected by the diagonal linear discriminant (DLD) onto a one-dimensional space. The t-statistic is then employed to score the informativeness of the gene pair in this space. A complete search evaluates all pairs of genes and ranks them in a list – without repetition – according to the scores. The computational complexity of this method is a serious limitation, therefore, a faster greedy search is also proposed. The genes are first ranked according to the individual t-statistic - as in univariate selection. The best gene is selected and the method searches for a gene amongst the remaining genes which, together with the individual best gene, maximizes the t-statistic in the projected space. This provides the first two genes of the ordered list. From the remaining $p - 2$ genes the best individual gene is selected and matched with a gene from the remaining $p - 3$ genes which maximizes the score in the projected space. This provides the second pair of genes. By iterating the process, pairs of genes are added, until all the genes have been selected. Similar to the univariate selection approach, we have now established a series of gene sets as well as the order in which they are subsequently evaluated, once again by starting with the first pair in the ranking, and then creating new sets by expanding the previous set with the next pair of genes in the ranking. Following [Bo 02], the Fisher classifier is employed to evaluate each gene set. Formally, θ_Ω represents the choice of DLD as mapping function, the t-statistic as univariate criterion in the mapped space and the choice of the Fisher classifier to evaluate the extracted gene sets. ϕ represents the desired number of genes to be extracted and θ_Φ represents the type of cross validation to employ during gene set evaluation.

Forward selection (FS). Forward gene selection starts with the single most informative gene and iteratively adds the next most informative genes in a greedy fashion. Here, we adopt the forward search proposed by Bo *et al.* [Bo 02]. The best individual gene is found according to the t-statistic. The second gene to be added is the one that, together with the first gene, has the highest t-statistic computed in the one-dimensional DLD projected space. This set is expanded with the gene which, in combination with the first two genes, maximizes the score in the projected space – now a three-dimensional space projected to a single dimension. By iterating this process an ordered list of genes is generated, once again defining a collection of gene sets, as well as the order in which these are evaluated. Now the length of the list is limited to n genes. In [Bo 02] this upper limit stems from the fact that the Fisher classifier cannot be solved (without taking additional measures) when the number of genes exceed n. Although elsewhere we employ the pseudo-inverse to overcome this problem associated with the Fisher classifier, we chose to maintain this upper limit in order to remain compatible with the set-up of [Bo 02]. Moreover, it keeps the selection technique computationally feasible. The formal definition of parameters corresponds exactly to the base-pair approach, except that a greedy search strategy (instead of the approach proposed by [Bo 02]) is employed in the optimization phase.

Recursive Feature Elimination (RFE). RFE is an iterative backward selection technique proposed by Guyon et al. [Guyo 02]. Initially a Support Vector Machine (SVM) classifier is trained with the full gene set. The quality of a gene is characterized by the weight that the SVM optimization assigns to that gene. A portion (a parameter determined by the user) of the genes with the smallest weights is removed at each iteration of the selection process. In order to construct a ranking of all the genes, the genes that are removed, are added at the bottom of the list, such that the gene with the smallest weight is at the bottom. By iterating the procedure this list grows from the least informative gene at the bottom, to the most informative gene at the top. Note that the genes are not evaluated individually, since their assigned weights are dependent on all the genes involved in the SVM optimization during a given iteration. As was the case in all previous approaches, a ranked gene list is produced, which defines a series of gene sets, as well as the order in which these sets should be evaluated when searching for the optimal set. In our implementation we adopt both the Fisher classifier and the SVM, with the optimization parameter set to $c = 100$ and a linear kernel. Both setups where proposed by [Guyo 02]. While the Fisher classifier suffers from the dimensionality problem when $p \approx n$ (for $p > n$ regularization occurs due to the pseudo-inverse [Skur 01]), it has the advantage over the SVM that no parameters need to be optimized. Moreover, it allows for a comparison with the other studied approaches which also employ the Fisher classifier. We chose to remove one gene per iteration. Formally, θ_Ω represents the choice of SVM (or Fisher) as classifier to generate the evaluation weights for the genes, the regularization parameter of the SVM, as well as the number of genes to be removed during every iteration. ϕ represents the number of genes selected, while θ_Φ represents the type of cross validation to employ during gene set evaluation.

Liknon. Bhattacharyya *et al.* [Bhat 03, Grat 02] proposed a classifier called *Liknon* that simultaneously performs classification and relevant gene identification. Liknon is trained by optimizing a linear discriminant function with a penalty constraint via linear programming. This yields a hyper-plane that is parameterized by a limited set of genes: the genes assigned non-zero weights by Liknon. By varying the influence of the penalty one can put more emphasis on either reducing the prediction error and allowing more non-zero weights or increasing the sparsity of the hyperplane parameterization while decreasing the apparent accuracy of the classifier. The penalty term therefore directly influences the size of the selected gene set. Although [Bhat 03] fixed the penalty term ($C = 1$), we chose its value in a more systematic way, via cross-validation. The penalty term was allowed to vary in the range $C \in [0.1, \ldots, 100]$. Formally, θ_Ω is obsolete, ϕ represents the penalty parameter and θ_Φ the choice of cross validation type.

Top-scoring pair. A recent classifier called *Top-scoring pair* (TSP) has been proposed by [Gema 04, Xu 05]. The TSP classifier performs a full pairwise search. Let $\mathbf{X} = \{X_1, X_2, \ldots X_p\}$ be the gene expression profile of a patient, with X_i the gene expression of gene i. The top-scoring pair (i, j) is the one for which there is the highest difference in the probability of $X_i < X_j$ from Class A to Class B. A new patient $\mathbf{X}^d$ is classified as Class A if $X_i^d < X_j^d$ and as Class B otherwise. Advantages of the TSP classifier are the fact that no parameters need to be estimated (no inner cross-validation is needed), and that the classifier does not suffer from monotonic transformation of the datasets, e.g. data normalization techniques. Formally, θ_Ω and θ_Φ are obsolete, ϕ represents the best pair of genes.

4.2.2 Training and evaluation framework

In order to avoid any bias, the selection of the genes and training of the final classifier on the one hand and the evaluation of the classification performance on the other, must be carried out on two independent datasets. To this end, the framework formalized in [Wess 05], is adopted here. The framework is graphically depicted in Figure 4.1. The whole procedure is wrapped in an outer cross-validation loop. (The inner loop will be defined shortly). For N_o-fold outer cross validation, the dataset, D, is split in N_o equally sized and stratified parts. During each of the outer cross validation folds, indexed by j, the training set, $D_{(-j)}$ consists of all but the j^{th} part, while the j^{th} part constitutes the validation set, denoted by $D_{(j)}$. During the training phase, two steps are performed. First, gene selection is performed by optimizing the associated parameter (Equation 4.1). This process also employs an N_i-fold cross-validation loop (the inner loop) to generate and evaluate gene sets. Each inner fold provides the error curve of the classifier as a function of the number of genes. We compute the average of the curves across the folds. The number of genes that minimizes the average error is considered to define the optimal gene size. Subsequently the classifier is trained on the training set with the optimal parameter setting as input (Equation 4.3), e.g. the optimal gene size for the given classifier. The performance of this classifier is only then evaluated on the validation set:

$$p^*_{A,j} = \Psi_A(D_{(j)}, \omega^*_A), \tag{4.4}$$

where $p^*_{A,j}$ represents the performance of the optimal classifier on the outer loop validation set of fold j, and $\Psi_A(\cdot)$ the function mapping the dataset and classifier to a performance. Averaging the validation performance across the N_o folds yields the N_o-fold outer cross validation performance of the gene selection technique with the specific user-defined choices. We adopted 10-fold cross-validation for both the inner and outer loops. This choice is suggested by Kohavi [Koha 95], and was also applied to gene expression data by Statnikov *et al.* [Stat 05a]. The latter obtained similar results using a 10-fold

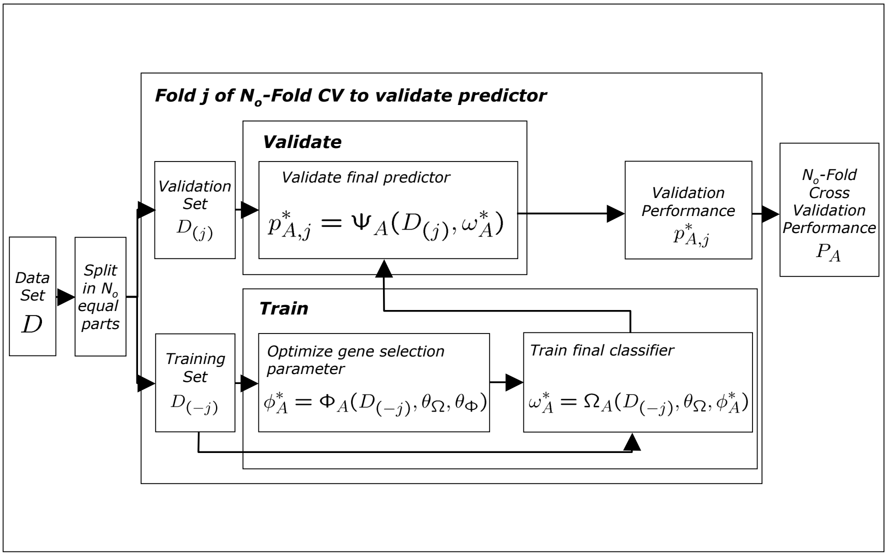

Fold j of N_o-Fold CV to validate predictor
Validate
Validate final predictor
$p^*_{A,j} = \Psi_A(D_{(j)}, \omega^*_A)$
Validation Set
$D_{(j)}$
Validation Performance
$p^*_{A,j}$
N_o-Fold Cross Validation Performance
P_A
Data Set
D
Split in N_o equal parts
Train
Training Set
$D_{(-j)}$
Optimize gene selection parameter
$\phi^*_A = \Phi_A(D_{(-j)}, \theta_\Omega, \theta_\Phi)$
Train final classifier
$\omega^*_A = \Omega_A(D_{(-j)}, \theta_\Omega, \phi^*_A)$

Figure 4.1: The training-validation protocol employed to evaluate various gene selection and classification approaches in simplified schematic format. The input is a labeled dataset, D, and the output is an estimate of the validation performance of algorithm A, denoted by P_A. The most important steps in the protocol are the training step (Block labeled 'Train') and the validation step (Block labeled 'Validate'). The training step, in turn, consists of two steps, namely 1) the optimization of the gene selection parameter, ϕ, employing a N_i – fold cross validation loop and 2) training the final classifier given the optimal setting of the selection parameter. The validation step estimates the performance of the optimal trained classifier (ω^*_A) on the completely independent validation set.

or leave-one-out cross-validation. The former is preferable due to lower computational requirements and lower variance.

To estimate the performance of a classification system we use the balanced average classification error which applies a correction for the class prior probabilities, if these are unbalanced. In this way the results are not dependent on unbalanced classes, and the results on different classifiers can be better compared.

The algorithms were implemented in Matlab employing the PRTools [Duin 04] and PRExp [Pacl 05] toolboxes.

4.2.3 Datasets

In total we employed seven microarray gene expression datasets. Four datasets, *Central Nervous System* (CNS) [Pome 02], *Colon* [Alon 99], *Leukemia* [Golu 99] and *Prostate* [Sing 02], were measured on high-density oligonucleotide Affymetrix arrays. Three datasets, *Breast Cancer* [Veer 02, Vijv 02], Diffuse Large B-cell Lymphoma (*DLBCL*) [Aliz 00] and *Head and Neck Squamous Cell Carcinomas* (*HNSCC*) [Roep 05] were hybridized on two-color cDNA platforms. The datasets represent a wide range of cancer types. The tasks are (sub)type prediction (*Colon*, *Leukemia*, *DLBCL* and *Prostate*) while for the remaining problems the goal is to predict the future development of the disease: patient survival (*CNS*), probability of future metastasis (*Breast Cancer*) and lymph node metastasis (*HNSCC*).

The *Breast Cancer* dataset consists of 145 lymph node negative breast carcinomas, 99 from patients that did not have a metastasis within five years and 46 from patients that had metastasis within five years. The number of genes is 4919. The *CNS*; dataset is a subset of a larger study. It considers the outcome (survival) after embryonic treatment of the central nervous system. The number of genes is 4458, while the number of samples is 60, divided into 21 patients that survived and 39 that died. The *Colon* dataset is composed of 40 normal healthy samples and 22 tumor samples in a 1908 dimensional feature space. The *DLBCL* dataset is a subset of a larger study which contains measurements of two distinct types of diffuse large B-cell lymphoma. The number of genes is 4026. The total number of samples is 47, 24 belong to the 'germinal center B-like' group while 23 are labeled as 'activated B-like' group. The *Leukemia* dataset contains 72 samples from two types of leukemia where 3571 genes are measured for each sample. The dataset contains 25 samples labeled as acute myeloid (AML) and 47 samples labeled as acute lymphoblastic leukemia (ALL). The *Prostate* cancer dataset is composed of 52 samples from patients with prostate cancer and samples from 50 normal tissue. The number of genes is 5962. For the *HNSCC* dataset, the goal is to predict, based on the gene expression in a primary *HNSCC* tumor, whether a lymph node metastasis will occur. This dataset consists of 66 samples (39 which did metastasize, and 27 that remained disease-free) and the expression of 2340 genes.

The datasets present a variety of the tissue types, technologies and diagnostic tasks. In addition, the panel of sets contains relatively simple, clinically less relevant tasks, such as distinguishing between normal and tumor tissue, as well as more difficult tasks, such as predicting future events based on current samples. We therefore consider the datasets suitable to perform a comparative investigation between univariate and multivariate gene selection techniques.

4.3 Results

The focus of our work is on gene selection techniques. We adopted several univariate and multivariate selection approaches. For each dataset, the average classification error across the folds of the 10-fold outer cross-validation and its standard deviation are reported in Tables 4.1 and 4.2. The best result for each dataset is emphasized in bold characters. For comparison the performance of three classifiers, namely Nearest Mean Classifier (NMC), Fisher (FLD) and the Support Vector Machine (SVM), is evaluated without any gene selection being performed, i.e. when the classifiers are trained with all the genes. We judge that method A with mean and standard deviation of the error rate μ_A and σ_A is significantly better than method B with mean and standard deviation of the error rate μ_B and σ_B when $\mu_B \geq \mu_A + \sigma_A$. The stars in Tables 4.1 and 4.2 indicate results that are similar when employing this rule-of-thumb.

As can be observed from Tables 4.1 and 4.2, the univariate approaches are significantly better than both the multivariate approaches and cases where no gene selection was performed in two cases: *DLBCL* and *HNSCC*. In addition, univariate approaches are the best but not significantly better for the *Breast Cancer* and *CNS* datasets, and comparable to the best approach in the remaining two cases (*Leukemia* and *Prostate*). Only for the *Colon* dataset, the univariate approaches perform significantly worse than the multivariate TSP.

Method	CNS	Colon	Leukemia	Prostate
gene selection	mean ± std	mean ± std	mean ± std	mean ± std
U, SNR, NMC	**30.4 ± 6.5** ⋆	12.9 ± 4.2	4.8 ± 2.7 ⋆	9.7 ± 4.2
U, SNR, FLD	42.5 ± 7.3	19.2 ± 5.9	8.0 ± 3.2	10.0 ± 3.0
U, t-test, NMC	32.5 ± 4.9 ⋆	12.5 ± 4.2	4.8 ± 2.7 ⋆	10.8 ± 3.4
U, t-test, FLD	35.8 ± 6.5 ⋆	11.7 ± 3.5	12.0 ± 4.2	8.0 ± 2.5 ⋆
BP greedy, FLD	43.8 ± 6.2	12.9 ± 3.8	11.6 ± 3.6	9.8 ± 3.3
FS, FLD	47.9 ± 5.1	15.4 ± 4.1	10.2 ± 4.2	14.0 ± 3.4
RFE, FLD	34.2 ± 5.0 ⋆	22.9 ± 4.4	**3.5 ± 2.6** ⋆	10.0 ± 2.6
RFE, SVM	35.4 ± 5.0 ⋆	22.1 ± 3.5	4.5 ± 2.6 ⋆	8.0 ± 2.9 ⋆
Liknon	32.9 ± 6.1 ⋆	13.3 ± 4.2	11.8 ± 4.0	10.8 ± 3.7
TSP	47.0 ±5.6	**5.4 ±2.9**	10.6 ± 3.8	**7.0 ±2.6** ⋆
no gene selection	mean ± std	mean ± std	mean ± std	mean ± std
NMC	42.1 ± 5.5	17.9 ± 3.3	**3.5 ± 2.6** ⋆	33.7 ± 3.9
FLD	32.9 ± 6.3 ⋆	21.7 ± 3.7	4.5 ± 2.6 ⋆	8.0 ± 2.5 ⋆
SVM	35.4 ± 7.0 ⋆	22.1 ± 3.5	**3.5 ± 2.6** ⋆	8.0 ± 2.9 ⋆

Table 4.1: The mean and the standard deviation of the 10-fold cross-validation error (in percentage) for the different approaches and the Affymetrix platform datasets employed in the study.

Employing the t-test or SNR in the univariate approaches has no effect on the error rate when employed in combination with the NMC. However, it has a significant effect in combination with the Fisher classifier. This is mainly due to the sensitivity of the Fisher classifier when the number of training objects approaches the number of selected genes during training [Skur 01]. This stems from the fact that the size of the selected gene-sets changes considerably across the folds of the gene optimization procedure, and may lead to sub-optimal gene set optimization.

Concerning the studied multivariate techniques, the base pair (BP) and forward search (FS) approaches of Bo *et al.* [Bo 02] are significantly worse in the majority of the datasets, with the exception of the base pair approach in the case of the *Colon* dataset. The *Liknon* classifier reaches error rates comparable to univariate results on the *CNS* and *Colon* datasets. The Recursive Feature Elimination [Guyo 02] performs slightly better than the other multivariate approaches achieving performances that are not significantly worse than the best approach on four datasets. However, in three of these cases, the performance is similar to the results achieved without any gene selection. As was observed by [Guyo 02], our results also indicate that there is no significant difference between RFE employing the Fisher or SVM classifiers. Although the TSP method is the best performing approach for the *Colon* and *Prostate* datasets, its performance is not stable across the remaining datasets, in fact, it is worse than the best performing method in all the remaining datasets.

Summarizing, in six of the seven adopted datasets there is no detectable improvement when employing multivariate approaches, since better or comparable performances are

Method	DLBCL	HNSCC	Breast
gene selection	mean ± std	mean ± std	mean ± std
U, SNR, NMC	**2.5 ± 2.5** ⋆	**21.2 ± 7.1** ⋆	33.0 ± 3.4 ⋆
U, SNR, FLD	15.8 ± 6.4	33.3 ± 6.6	**29.9 ± 3.6** ⋆
U, t-test, NMC	**2.5 ± 2.5** ⋆	**21.2 ± 7.3** ⋆	33.5 ± 3.8 ⋆
U, t-test, FLD	15.8 ± 6.4	36.2 ± 6.2	32.6 ± 3.0 ⋆
BP greedy, FLD	10.0 ± 4.3	36.2 ± 7.0	35.8 ± 2.3
FS, FLD	10.8 ± 3.7	45.4 ± 8.5	35.4 ± 4.2
RFE, FLD	16.7 ± 5.3	35.0 ± 6.3	33.8 ± 3.5
RFE, SVM	15.8 ± 5.2	35.4 ± 7.2	32.6 ± 3.2 ⋆
Liknon	13.3 ± 5.3	37.5 ± 7.4	34.5 ± 5.2
TSP	27.5 ±2.8	37.6 ± 6.0	49.9 ± 4.6
no gene selection	mean ± std	mean ± std	mean ± std
NMC	6.7 ± 3.5	29.2 ± 7.2	36.7 ± 3.2
FLD	14.2 ± 5.4	32.5 ± 6.6	35.8 ± 4.1
SVM	9.2 ± 3.8	29.6 ± 5.7	34.3 ± 4.2

Table 4.2: The mean and the standard deviation of the 10-fold cross-validation error (in percentage) for the different approaches and the cDNA platform datasets employed in the study.

obtained with univariate methods or without any gene selection. The classification performance alone cannot be regarded as an indication of biological relevance, since a good classification could be reached with different gene sets, and gene-set sizes, depending on the methodology employed. This is in agreement with the studies of Ein-Dor at al. [Ein 05] and Michiels *et al.* [Mich 05]. These studies pointed out that the selected gene sets are highly variable depending on the sampling of the dataset employed during training. However, different gene-sets perform equally well [Ein 05, Bhat 03, Chow 01, Golu 99], indicating that there is, in fact, a large collection of genes that report the same underlying biological processes, and that *the unique gene set* does not exist. The lack of performance improvement when applying multivariate gene selection techniques could also be caused by the small sample size problem. This implies that there are too few samples to detect the complex, multivariate gene correlations, if these were actually present. Only one multivariate approach, namely the TSP method, was able to extract a pair of genes that significantly improved the classification performance.

4.4 Conclusions

In gene expression analysis gene selection is undertaken in order to achieve a good classification performance and to identify a relevant group of genes that can be further studied in the quest for biological understanding of the cancer mechanisms. In the literature it is claimed that both multivariate and univariate approaches successfully achieve both purposes. However, these results are often biased since the training and validation phases of the classifiers are not strictly separated. Moreover, the results are often based on few and relatively simple datasets. Therefore no clear conclusions can be drawn.

Therefore, we have performed a comparison of frequently used multivariate and univariate gene selection algorithms across a wide range of cancer gene expression datasets within a framework which minimizes the performance biases mentioned above.

We have found that univariate gene selection leads to good and stable performances across many cancer types. Most multivariate selection approaches do not result in a performance improvement over univariate gene selection techniques. The only exception was a significant performance improvement on the *Colon* dataset employing the TSP classifier, the simplest of the investigated algorithms employing multivariate gene selection. However, the performances of the TSP method are not stable across different datasets. Therefore, we conclude that correlation structures, if present in the data, cannot be detected reliably due to sample size limitations. Further research and larger datasets are necessary in order to validate informative gene interactions.

Epilogue on Chapter 4

In Chapter 4 we have investigated several univariate and multivariate approaches for the selection of an informative representation (gene-set) for classification purposes. The conclusion of the study is that the univariate gene selection leads to better and more stable performances across many cancer types then the investigated multivariate approaches. The multivariate gene selection was not able to point out gene-sets which would improve the classification performance, proving to select less informative gene-sets than the univariate ones. Can we conclude that the univariate approaches are the best choice? And are they good enough?

We want to emphasize that the limited sample size is currently the major constraint limiting the complexity of the gene selection and classification algorithms. Multivariate aspects of the genes may not be detected simply due to the limited number of samples. Another aspect concerns the use of the univariate selection methods. Are the classification performances satisfactory for promoting the adoption of these procedures in the clinic? This depends on many aspects. First of all on the problem definition. We have observed that the classification performance is below 5% in some datasets (i.e. Colon, Leukemia and DLBCL) while is about 30% in others (e.g. the NKI dataset). This suggests that is easier to address questions such as discrimination of tumor/normal tissue (Colon dataset) or a well defined subtypes of cancer (Leukemia, DLBCL datasets), then questions such as the ability of a tumor to metastasize (NKI dataset). Whether these performances are satisfactory depends on a comparison with the classification performances of already available classical approaches. A second important aspect is the definition of the requirements. In the comparative analysis we have performed, the average error with equal class prior was used as a criterion to judge the methodologies. This may be not a valid criterion e.g. in a screening situation, where the number of healthy persons is expected to be much larger than the number of unhealthy ones. In this case a more specific analysis is required (e.g. the Receiving Operating Curve analysis). Therefore, the conclusions of this study cannot be extended without further investigation to different scenarios.

Part II

5

SIRAC: Supervised Identification of Regions of Aberration in aCGH datasets

Chapter 5, concentrates on copy number data. We have developed an algorithm (SIRAC) that exploits spatial dependencies in order to identify regions of chromosomal aberrations, which are correlated with the classes of interest. In particular, the focus has been on the characterization of copy number aberrations in the cancer subtypes identified by Sorlie and Perou [Pero 00, Sorl 01, Sorl 03]. [1]

---

[1] This chapter was published in *BMC Bioinformatics* [Lai 07].

5.1 Background

Genomic alterations in DNA copy number are important events in cancer development [Leng 98]. A tumor suppressor gene can be disabled by the physical loss of the gene, or similarly an oncogene may be over-expressed via the amplification of the region where it is located. The identification of chromosomal aberrations is, therefore, a powerful instrument in studies of cancer. It may suggest target genes for new drugs or shed light on the mechanisms which regulate the response to therapies [Soti 07, Pink 05, Bert 03].

The first approach to search for copy number alterations in CGH has been made by Kallioniemi *et al.* [Kall 92] using metaphase chromosomes. The extensions of this technique employ array technology to perform a high resolution scan of the genome. As reviewed by Pinkel *et al.* [Pink 05], several array CGH (aCGH) techniques have been developed. The spotting technology makes use of BAC clones ($100 - 200$ kb), cDNA clones ($\sim 100 - 1000$ bp) and lately oligonucleotides ($30 - 100$ bp). More recently, *in-situ* technologies synthesize small oligonucleotides directly onto the array. Since the oligos can be a few tens bp long, higher resolution are reached, if a good coverage of the genome is adopted.

An important challenge to analyze aCGH data is to find the aberrated chromosomal regions specific to the problem under study, e.g. to distinguish between subtypes of cancer. In order to reach this goal, three groups of approaches can be found in the literature. The first group of approaches uses only the aCGH data. First they identify the amplifications/deletions in each sample individually, and then search for the common aberrations between the samples. The identification per sample of chromosomal regions of aberration is a task in itself that has been approached in several ways. The simplest solution is the application of a threshold. The DNA-probes (BAC clones, cDNA clones or oligonucleotides) which exceed the threshold are considered amplified/deleted [Velt 03, Call 05, Nayl 05, Schw 04]. The choice of the threshold is a very critical parameter. Moreover, the threshold methods have the limitation that they do not take into account the spatial location of the DNA-probes. Since amplicons (i.e. regions that are amplified in a sample) are commonly assumed to involve more than a single DNA-probe, the spatial position is an important factor. Several more complex algorithms have been developed to identify, per sample, the aberrated regions in more robust ways. Lai *et al.* [Lai 05] reviewed eleven different methods available in the literature. Numerous segmentation methods have been proposed to divide the aCGH profile in piece-wise constant segments, and a likelihood function is used to estimate the model parameters from the data. For example, Picard *et al.* [Pica 05] modeled the aCGH profile with a random Gaussian process and introduced an adaptive penalized likelihood to estimate the segments and their locations. Jong *et al.* [Jong 03, Jong 04] proposed a genetic algorithm to maximize the likelihood function. A different approach was introduced by Wang *et al.* [Wang 05a]. They identified the regions of amplification/deletion via a hierarchical clustering along the chromosome.

The biologically relevant aberrations are not the ones that characterize a single sample, since these can be the consequence of the genomic instability of the particular tumor. The more interesting aberrations are the ones shared by many samples, ideally by all the samples in the same class. Previous studies combined the information of the per sample aberration by looking at the frequency of patients that carry the aberration [Frid 06, Wang 05a, Velt 03, Nayl 05, Hyma 02, Guo 02]. Again a threshold on the minimal frequency is chosen. For example, Fridlyand *et al.* [Frid 06] require the aberrations to be present in more than 50% of one class and less than 30% of the second class, whereas Hyman *et al.* [Hyma 02] demands that the aberration be present in at least two specimens.

These approaches have in common that the class information is taken into account only in the second stage of the analysis, i.e. when computing the aberration frequency across the samples. In the first phase also the aberrations common to more classes are considered, even if they are not of interest for the study. This introduces an extra parameter when evaluating the significance of the aberrations to distinguish the classes of interest. Recently, Diskin *et al.* [Disk 06] proposed a more complex and systematic way to evaluate the significance of aberrations across samples. However, they require the input data to be discretized per sample into amplifications and deletions. This step can be performed using one of the mentioned above methods, but makes the results dependent on the particular approach chosen for discretization.

A second group of approaches to detect aberrations across samples uses only the gene expression data together with the chromosomal location of the genes. The assumption is that an amplification directly affects the expression of the genes. Therefore, the genes in that region should have a detectable common over-expression. Similarly, the genes located in a deletion would have a detectable under-expression. Furge *et al.* [Furg 05] applied the binomial test per sample on the genes within a given window size. In order to cover the whole genome, the window is slid across the genome, performing a test at fixed intervals. The z-scores of the test for a particular location are averaged across several window sizes and a threshold is chosen. The locations above/below the threshold are identified as regions of chromosomal aberration. Levin *et al.* [Levi 05] applied a Poisson model to the expression data and incorporated the genomic location in their model-based scan statistic. These results are compared per sample with the aCGH data. Yi *et al.* [Yi 05] used a sliding window size of 5 genes to test the significance of the region according to two scores, which account for the homogeneity of behavior in the window and the power of the genes in discriminating the classes of interest. Dressman *et al.* [Dres 03] observed that the genes over-expressed shared the same location, hypothesized an amplification and validated their findings with PCR. These studies show interesting examples of aberrations identified using the transcriptome data only. However, the assumed strong correlation of aCGH and expression could not be detected by other studies [Mele 05, Guo 02, Mart 03, Sanc 03]. Since the alteration in expression may be due to diverse mechanisms, the potentially underlying chromosomal aberrations would need to be verified either by PCR or FISH, if the number of loci to be tested is tractable, otherwise by aCGH data. The advantage of the aCGH technology arises in the genome-wide coverage of the analysis.

The third group of approaches combines aCGH and expression data to detect regions of chromosomal aberration. The SLAM algorithm (Adler *et al.* [Adle 06]) is a prime example of this group. First the SAM analysis [Tush 01] is applied to the aCGH data in order to identify the DNA-probes which distinguish the two classes. Then the focus is on the DNA-probes that are correlated with the expression data. Based on the observation that many of them were on the same chromosome arm, the hyper-geometric distribution was used to test the significance of that arm.

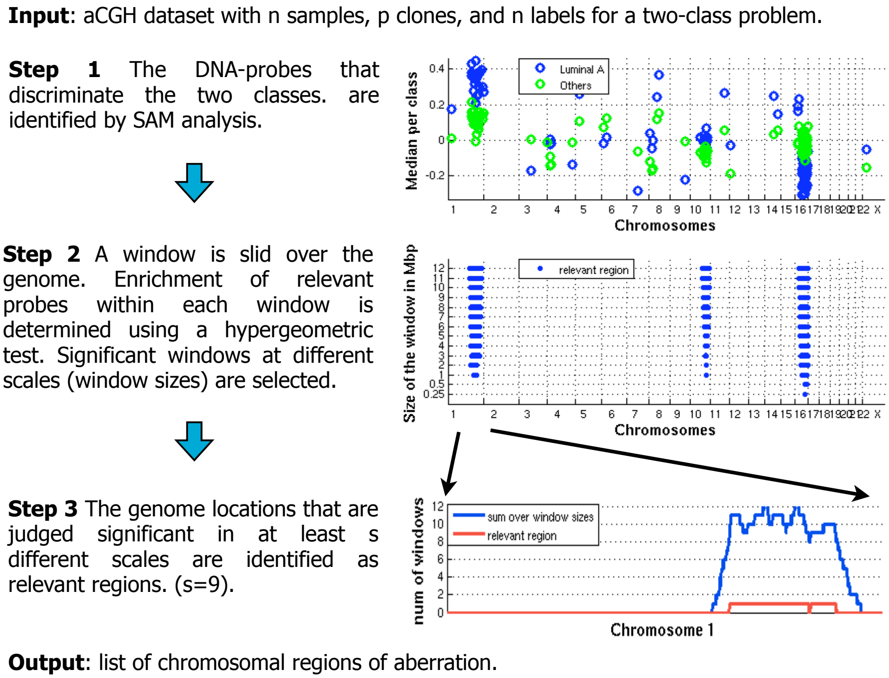
Inspired by the work of Adler *et al.* [Adle 06], we propose a supervised procedure to identify chromosomal regions of aberration using solely aCGH data. We use the SAM analysis to determine the "relevant" DNA-probes, i.e. the DNA-probes that distinguish the classes of interest. While Adler *et al.* [Adle 06] evaluated only a single location chosen in an ad hoc fashion, we build a systematic search to test the whole genome. We adopt a sliding window approach similar to the one proposed by Furge *et al.* [Furg 05]. More specifically, we apply a hyper-geometric test to window sizes of different length, and test the significance of the number of relevant DNA-probes in those windows. Our algorithm belongs to the first group of approaches, since it uses only aCGH data. However, it differs from the typical approaches in this group in the following ways. First of all it focuses only on the aberrations specific to the problem of interest, by exploiting the class labels in the first step (recognizing relevant DNA-probes). Importantly, no discretization, smoothing or segmentation algorithms are applied to the aCGH data. This leads to the advantage that the data is not altered based on the preconceived models that these algorithms presume. Moreover, we also avoid the optimization of the parameters that these models usually require (avoiding results sensitive to these choices). The use of the hyper-geometric test corrects for the non-uniform background distribution of the DNA-probes. This is particularly important since the DNA-probes are not equally spaced along the genome. In this way we build a robust algorithm to identify areas of interest specific to the problem under study. We illustrate the benefit of our procedure on an artificial dataset, and show the results on two breast cancer datasets.

5.2 Algorithm description

Figure 5.1 illustrates our algorithm SIRAC (Supervised Identification of Regions of Aberration in aCGH data). A detailed description is given in Algorithm 4. An aCGH dataset D and its label set y provide the starting point. The procedure consists of three steps.

STEP 1. We identify with the SAM analysis [Tush 01] the DNA-probes which discriminate between the classes of interest. We call these DNA-probes the "relevant" probes. In Figure 5.1 (Step 1) the relevant DNA-probes are depicted on the genomic location. Each probe is plotted with two circles of different color representing the median value of the samples in the two classes.

STEP 2. We test, in a systematic way, whether the number of relevant DNA-probes in a region is higher than expected by chance. For this purpose we use the hyper-geometric test for a genomic position, and test whether the fraction of relevant DNA-probes in the window of length $2w$ represents a significant enrichment. By sliding the window of observation along the genome, shifting it a single DNA-probe position at a time, we obtain the test results for all positions. This procedure can be done effectively since the genomic locations where the test presents uncertainty, and therefore, needs to be computed, are only a subset of all genome positions. The locations are dependent on the positions where the relevant DNA-probes are situated. More precisely, for a given window w, the test needs only to be performed for three positions: a window centered on the location l of the DNA-probe itself, and two windows centered at $l - w$ and $l + w$, i.e. centered at the end points of the first window. Consequently, tests are done for the three windows $[l - 2w, l]$, $[l - w, l + w]$ and $[l, l + 2w]$ around the relevant DNA-probe. In total $3k$ tests are performed, where k is the number of relevant probes. This solution is computationally fast and allows a feasible multiple testing correction while providing the coverage of all genome positions relevant to the test. A Bonferroni correction for multiple testing is applied by multiplying the p-value of each test by the number of tests



Input: aCGH dataset with n samples, p clones, and n labels for a two-class problem.

Step 1 The DNA-probes that discriminate the two classes. are identified by SAM analysis.

Step 2 A window is slid over the genome. Enrichment of relevant probes within each window is determined using a hypergeometric test. Significant windows at different scales (window sizes) are selected.

Step 3 The genome locations that are judged significant in at least s different scales are identified as relevant regions. (s=9).

Output: list of chromosomal regions of aberration.

Figure 5.1: Illustration of the algorithmic steps of SIRAC with the corresponding results for the label Luminal A subtype versus all others subtypes. In Step 1 the relevant DNA-probes are selected. Each DNA-probe is plotted on the genomic location with two circles of different color representing the median value of the samples in the two classes. In Step 2, the vertical axis represents the different window sizes, the lines along the genome (the horizontal axis) show the regions judged significant by the algorithm. In the final step, Step 3, the number of window sizes for which the location is judged significant by the hyper-geometric test are shown along the vertical axis. The relevant region selected when $s = 9$ is highlight by the lower curve.

performed ($3k$). Note that the Bonferroni correction is a rather conservative correction, since the windows of observation of different DNA-probe may not be independent.

In order to identify the regions of aberration, we interpolate the corrected p-values of the hypergeometric test using the maximum value; i.e. given two successive locations with corrected p-value a and b, the base-pairs positioned between those locations are assigned the maximum of a and b. The base-pairs of the genome where the corrected p-value is smaller than 0.05 are considered significantly enriched for genomic aberrations. This step is repeated for different window sizes in order to detect both small and large aberrations. An illustrative result is shown in Figure 5.1(Step 2). On the vertical axis are the different window sizes, the lines along the genome (the horizontal axis) show the regions judged significant by the algorithm.

STEP 3. The regions of aberration are identified based on a consensus between the results of the different window sizes. As illustrated in Figure 5.1 (Step 3), the number of window sizes for which a location is judged significant by the hyper-geometric tests are shown on the vertical axis. The "relevant" regions are the locations judged significant by at least s window sizes (the result for $s = 9$ is depicted by the lower curve in Figure 5.1 (Step 3). The researcher can decide to accept relevant regions as those in which any of the window sizes showed a significance, or may be more strict and demand the significance across several scales. The regions of chromosomal aberration are provided as output.

Algorithm 3 SIRAC: Supervised Identification of Relevant Aberration in aCGH datasets

```
Input: dataset $D$, label set $\mathbf{y}$, SAM parameters: $\delta$ for the desired false discovery rate and number of iterations $I$; vector $\mathbf{W}$ with half the sizes of observation windows; threshold $t$ for the hyper-geometric distribution; minimum number of windows sizes $s$ for which the location is judged significant.
Apply the SAM analysis with the given parameters $\delta$ and $I$ to the labeled dataset $D, \mathbf{y}$. A vector $\mathbf{J}$ stores the indexes of the relevant DNA-probes obtained.
Initialize variables: $P = ones(|\mathbf{W}|, 3|\mathbf{J}|)$, stores the p-value of the test; $POS = zeros(|\mathbf{W}|, 3|\mathbf{J}|)$ stores the location where the test is applied.
$\forall w \in \mathbf{W}$ (for all window sizes)
  · Initialize: bon=0; (count the number of tests performed)
  · $\forall j \in \mathbf{J}$ (for all relevant DNA-probes)
    · Determine position of the window centers $\mathbf{C} = [l^j - w, l^j, l^j + w]$ around the DNA-probe, with $l^j$ the position of the $j$th DNA-probe.
    · If $\mathcal{C}h(l^j - w) = \mathcal{C}h(l^j) = \mathcal{C}h(l^j + w)$, with $\mathcal{C}h$ a function that assigns the chromosome number of the corresponding base pair location
    · Then
      · Initialize: $\mathbf{H} = ones(1, 3)$, (stores the test value for the triplet position in $\mathbf{C}$)
      · $\forall c \in \mathbf{C}$ (for all window positions)
        · $h = \sum_{i=0}^{x} \mathcal{H}(i|M, k, N)$, with:
            $x$ = number of relevant DNA-probes in the window $[c - w, c + w]$,
            $M$ = number of DNA-probes in the dataset $D$,
            $k$ = number of relevant DNA-probes in the dataset $D$,
            $N$ = number of DNA-probes in the window $[c - w, c + w]$.
        · $H^c = 1 - h$;
        · bon=bon+1; (update the counter)
     · End
     · $P^{wj} = \mathbf{H}$; ($P^{wj}$ is the p-value on row $w$ and probe triplet $j$);
     · $POS^{wj} = \mathbf{C}$; ($POS^{wj}$ stores the triplet window location);
  · $P^w = P^w \times bon$; (Bonferroni correction)
$\forall\, l \in \mathbf{G}$ (all positions in the genome):
  · $F_l = \sum_w N_l^w$, ($F_l$ = number of window sized where the test is above the threshold t), with:
      $N_l^w = \begin{cases} 1, & \text{if} \, \exists \, j_a \, | \, l^{j_a} \leq l \leq l^{j_{a+1}} \& \, max(P^{wy_a}, P^{wy_{a+1}}) \leq t), \\ 0, & \text{otherwise.} \end{cases}$
Output: all locations with $F_l \leq s$.
```

Complexity and scalability issues

Our real datasets are BAC aCGH, with $\sim$ 3000 DNA-probes. The complexity of the SIRAC algorithm is 1) $\mathcal{O}(Nlog(N))$, where N is the total number of elements (DNA-probes) in the array (since the SAM analysis has to be performed for each single probe), and 2) $\mathcal{O}(k)$, where k is the number of relevant elements selected by SAM (since the hypergeometric test is applied three times per relevant probe). Therefore, SIRAC can be used also with higher resolution aCGH, such as the cDNA or the oligo arrays. To give an indication of the time demands we have evaluated the run time of SIRAC on our computer server (an Intel Xeon 2.33 GHz with 8 G of memory). The run time for the NKI dataset with 2952 DNA-probes, and 692 relevant DNA-probes was 50 seconds; while SIRAC took 401 seconds to run on a cDNA array dataset with 30601 DNA-probes, 2532 of which were judged relevant by the SAM analysis.

5.3 Experimental results

5.3.1 Set-up

We illustrate our algorithm on an artificial dataset, described in the following Section and apply our method to two breast cancer datasets. The first dataset (*NKI*) is composed by 67 patients and 3219 BAC clones (DNA-probes). The samples are a selected series of the 295 breast cancer samples described in [Vijv 02], and the BAC platform is discussed in [Beer 05]. The second dataset (*Fridlyand*) contains 67 samples and 2464 BAC clones, as described in [Frid 06].

In our proposed algorithm there are a few choices that the researcher has to make. A first important decision concerns the number of relevant DNA-probes. We choose to be conservative and require that the selected DNA-probes have a false discovery rate smaller than 0.005. This ensures that we include a very small fraction of false positive DNA-probes in further steps. Another parameter is the range of window sizes that are used to probe the genome. Since the average space between the clones is 1 Megabase (Mb), the minimum window of observation is set to 1 Mb. The maximum window size is fixed to 24 Mb because this is roughly half the length of the shortest chromosome. In this way, we enforce that the largest window does not always cover both the p and q arm of the chromosome.

5.3.2 Results

Artificial dataset

The artificial dataset is created using the clone distribution of the 207 clones of Chromosome 1 on the *NKI* array. The amplitude of the DNA-probes is drawn from a normal distribution with zero mean and unit variance $\mathcal{N}(0,1)$. We chose to have two classes with 35 samples each. In Class 1 we simulate an amplification of amplitude m spanning u DNA-probes situated between positions l_s and l_e. The remaining DNA-probes have an average amplitude of zero. For Class 2 all samples have an average amplitude of zero for all DNA-probes. Zero mean, unit variance Gaussian noise is added to all samples across all DNA-probes. More formally, for a DNA-probe p at position l in a sample of class 1

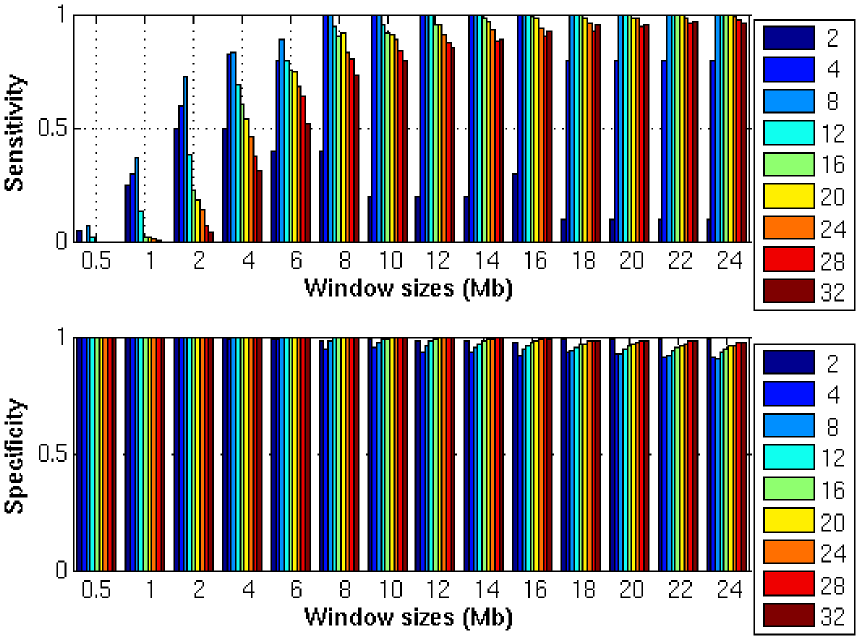


Figure 5.2: Sensitivity and specificity of the different window sizes for 10 instantiations of the artificial dataset with amplification amplitude $m = 0.8$. For each plot, on the horizontal axis are the distinct window sizes used to detect the amplification. The different color bars show the results for the individual amplification lengths u adopted.

the following holds:

$$p(l) = \begin{cases} \mathcal{N}(m, 1), & l_s < l < l_e \\ \mathcal{N}(0, 1), & \text{otherwise.} \end{cases} \tag{5.1}$$

The samples in the other class are all drawn from the normal distribution $\mathcal{N}(0, 1)$. The artificial dataset provides us with a ground truth which allows us to investigate the sensitivity and specificity of the algorithm and the effects of different window sizes. We applied our algorithm to amplifications with a range of amplitudes ($m \in \{0.2, 0.4, 0.6, 0.8, 1\}$) and widths ($u \in \{2, 4, 8, 12, 16, 20, 24, 28, 32\}$ megabases (Mb)).

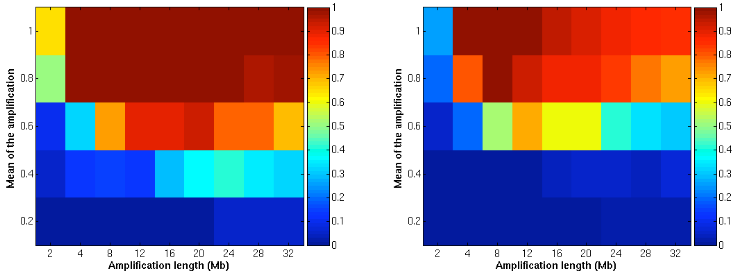
Given the region of amplification found by the algorithm, the DNA-probes located in this region that also belong to the interval between positions l_s and l_e are defined as true positive, while the DNA-probes outside the interval are denoted false positives. Similarly, for the DNA-probes outside the region of amplification found by the algorithm, true negatives are the DNA-probes outside the interval between positions l_s and l_e, while false negative are the DNA-probes included in this interval. In general, the same trend for specificity and sensitivity as a function of m is observed. Figure 5.2 shows the average sensitivity and specificity for 10 different instantiations of the artificial dataset with the amplitude of the amplification $m = 0.8$. On the horizontal axis are the different window sizes used to detect the amplification. The different color bars show the results for the different amplification lengths, u, adopted. In the upper plot the sensitivity is shown. Let us focus on the amplification of length 2 Mb (dark blue bar). It can be seen that the maximum sensitivity is reached for window sizes of length 2 and 4 while the sensitivity decreases for larger window sizes. Similarly the amplification of length 16 Mb (green bar) is detected with the maximum sensitivity of 1 by a window size 18 Mb. Consequently, smaller window sizes detect small amplifications better, while larger window sizes more accurately reveal the larger amplifications. This behavior highlights the benefits of using window sizes of different lengths, to detect both large and small chromosomal aberrations. As expected, the specificity is maximal for small window sizes and decreases when larger window sizes are used. This behavior is due to the fact that wider window sizes include a larger number of false positives DNA-probes than the smaller windows sizes.

In our algorithm, we combine the different window sizes in order to obtain a unique region of amplification, by setting the parameter s. A location is amplified if it is judged amplified in s window sizes. We also investigated the effect of the parameter s. The top four plots of Figure 5.3 illustrates the sensitivity and specificity for two values of the parameter s, i.e. $s \in \{2, 9\}$. We choose $s = 2$ as a loose constraint, while the more strict value of $s = 9$ requires the consensus of two-thirds of the window sizes. For each plot, the horizontal axis depicts the different amplification lengths, u used, and the vertical axis the amplitudes of the amplification, m. The colors code the value of the sensitivity and specificity from 0 to 1. The small amplification of $m = 0.2$ is very difficult to detect, therefore the sensitivity is very low regardless of the length of the amplification (bottom row of black squares in Figure 5.3 (a)). When the amplification amplitude increases, the sensitivity rises as well. If $s = 9$ fewer extremely large and small aberrations are not detected compared to $s = 2$, in other words, the sensitivity is lower when $s = 9$ compared to $s = 2$. However, at the same time, the specificity increases (Figure 5.3 (d)).

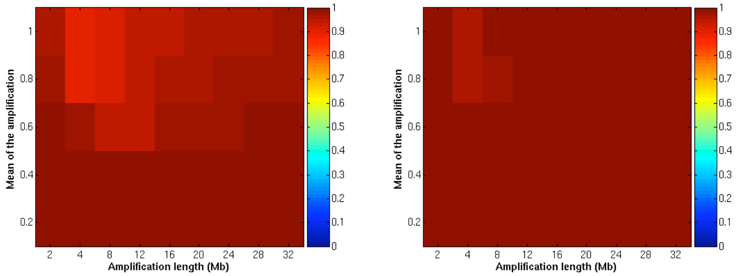
In order to evaluate the control of the error rate, we computed the False Positive Rate (FPR), which is defined as $\frac{FP}{FP+TP}$, with FP representing the number of False Positives and TP the number of True Positives. Figure 5.3 (e) and (f) shows the FPR for $s = 2$ and $s = 9$ discretized into 10 equal sized intervals of size 0.1. We can observe that when $s = 9$ the FPR is mostly below 0.1, while the control of the FPR is not so strict when $s = 2$. However, the improved control of the FPR is achieved at the cost of the sensitivity. In the following experiments with real data, we choose to use the less stringent constrain of $s = 2$ to maximize the sensitivity. A further prioritization of the DNA-probes in a region can take into account the “strength” of the amplification. For example, the list of DNA-probes may be prioritized according to the number of window sizes in which each DNA-probe is judged aberrated. In this way, the strong aberrations can be differentiated from the weak ones.

The *NKI* dataset

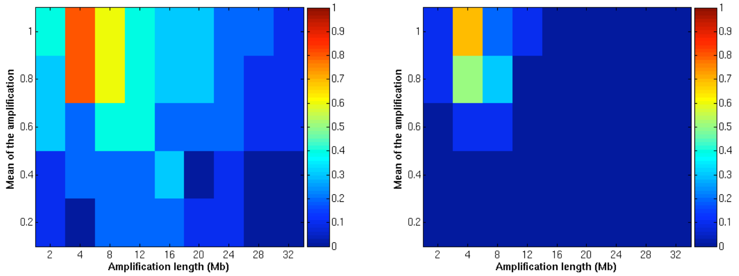
Sorlie and Perou [Pero 00, Sorl 01, Sorl 03] introduced the distinction of breast cancer into five different subtypes (Basal, ERBB2, Luminal A, Luminal B, Normal-like) based on the gene expression of the so called *intrinsic genes*. These genes were selected as the genes that had significantly greater variation in expression between different tumors than between paired samples of the same tumor. Using these genes, the profile of a centroid was obtained for each subtype. These centroids, in combination with the gene expression of 295 breast tumors [Vijv 02] were employed to assign each sample in the *NKI* set to



(a) Sensitivity, s = 2 (b) Sensitivity, s = 9



(c) Specificity, s = 2 (d) Specificity, s = 9

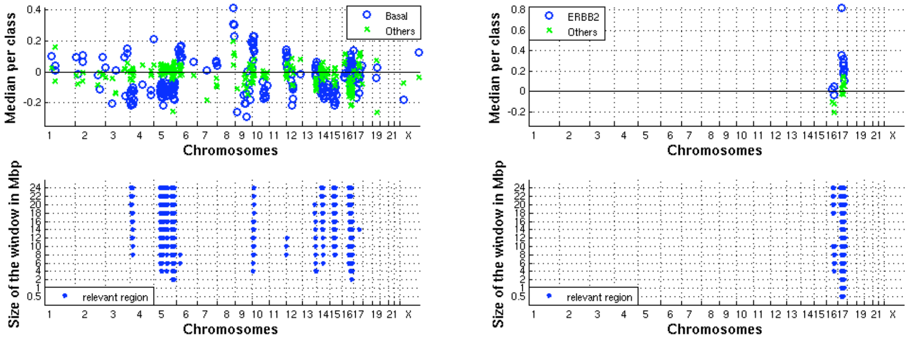


(e) False Positive rate, s = 2 (f) False Positive rate, s = 9

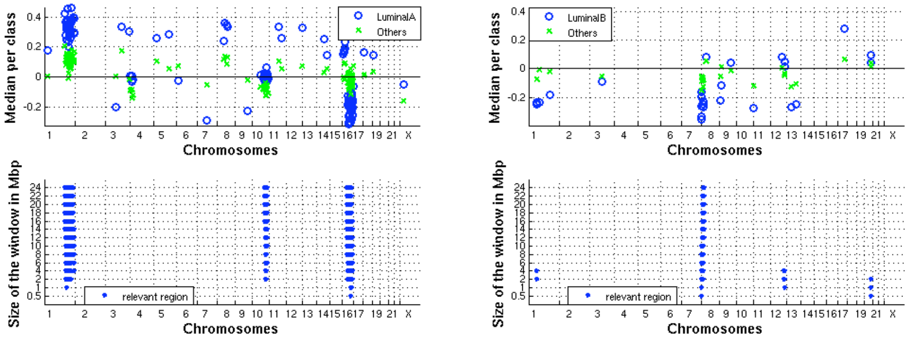
Figure 5.3: Sensitivity, specificity and False Positive Rate (FPR) for two values of the parameter s, i.e. $s \in \{2, 9\}$. For each plot, on the horizontal axis are the different amplification lengths u used, and on the vertical axis are the different amplitudes of the amplification m. The colors code the value of the sensitivity, specificity and FPR from 0 to 1.

one of the subtypes based on its correlation with the centroid profiles across the intrinsic genes. In the *NKI* data, 21 out of 67 samples were labeled as Basal, 10 as ERBB2, 21 as Luminal A, 12 as Luminal B and 3 as Normal-like.

Recently, Bergamaschi *et al.* [Berg 06] studied the genomic aberrations of the different subtypes on a aCGH dataset. We applied our method to the *NKI* dataset and compare our findings to the results of Bergamaschi *et al.* [Berg 06]. More specifically, we applied the SIRAC algorithm four times, each time analyzing one subtype against the rest. The Normal-like subtype was not considered in this analysis due to the small number of



(a) Basal (b) ERBB2



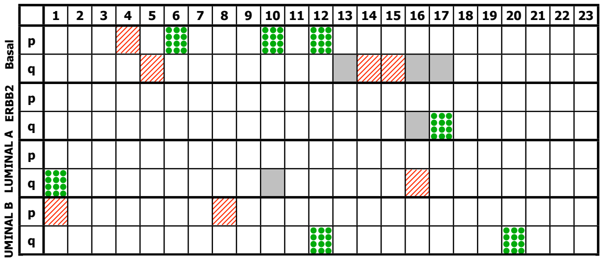
(c) Luminal A (d) Luminal B

Figure 5.4: Results of Step 1 and 2 of the SIRAC algorithm for the four different subtypes in the *NKI* dataset. In the top panel of each figure the relevant DNA-probes detected by the SAM analysis are displayed on their genomic location (horizontal axis). For each relevant DNA-probe a circle and an x mark are plotted at its location, representing the median of the class of interest (Basal, ERBB2, Luminal A or Luminal B) and the median of the remaining samples. In the bottom plot, the regions identified by the algorithm as significantly aberrated are shown for all genomic locations (horizontal axis) with a line for each window width used (vertical axis). The length of the line indicates the region on the genome that is significantly enriched with relevant DNA-probes.

samples.

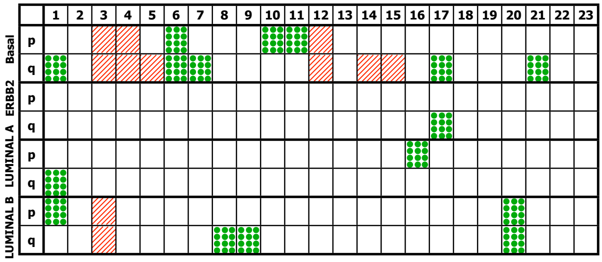
Figure 5.4 shows the results of Step 1 and 2 of the SIRAC algorithm for the four different subtypes in the *NKI* breast cancer dataset. In the top plot of each figure the relevant DNA-probes detected by the SAM analysis are displayed. For each relevant DNA-probe a circle and an x mark are plotted at its location on the genome, representing the median of the class of interest (Basal, ERBB2, Luminal A or Luminal B) and the median of the remaining samples. From these plots it is visible how some locations are significantly more densely populated by relevant DNA-probes than others. The lower plot of each subtype highlights the regions of aberration detected by SIRAC.

Figure 5.5 (a) summarizes the aberrations found on the p or q chromosomal arms of the different subtypes when $s = 2$. a box with inclined lines specifies a deletion,



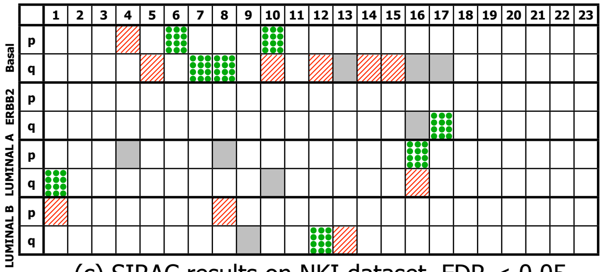
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Basal	p				▨		●				●		●											
	q					▨								■	▨	▨	■	■						
ERBB2	p																							
	q																■	●						
LUMINAL A	p																							
	q	●									■						▨							
LUMINAL B	p	▨							▨															
	q												●								●			

(a) SIRAC results on NKI dataset, FDR < 0.005



		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Basal	p			▨	▨		●				●	●	▨											
	q	●		▨	▨	▨	●	●					▨		▨	▨		●				●		
ERBB2	p																							
	q																	●						
LUMINAL A	p																●							
	q	●																						
LUMINAL B	p	●		▨																	●			
	q			▨					●	●											●			

(b) Bergamaschi results



		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Basal	p				▨		●				●													
	q					▨		●	●		▨		▨	■	▨	▨	■	■						
ERBB2	p																							
	q																■	●						
LUMINAL A	p				■				■								●							
	q	●									■						▨							
LUMINAL B	p	▨							▨															
	q									■			●	▨										

(c) SIRAC results on NKI dataset, FDR < 0.05

Figure 5.5: Summary of the aberrations per chromosome arm for the four different subtypes (Basal, ERBB2, Luminal A or Luminal B). The numbers in the top of the tables denotes the chromosomes. A arm is indicated with a inclined lines when a significant region is found on that arm that shows a deletion of the clones of interest. Similarly, a box filled with circles indicates amplification. The uniform gray boxes indicate that the aberration was not present in the class of interest but in the rest of the samples. The top and the bottom tables show the aberrations found with the SIRAC algorithm on the *NKI* dataset for two different values of the FDR, i.e. FDR < 0.005 and fdr < 0.05 respectively. The middle table presents the results of Bergamaschi *et al.* [Berg 06] on their breast cancer dataset.

a box filled with circles an amplification, and a uniform gray box indicates that the aberration was not in the class of interest. Note that the resolution of SIRAC is neither restricted to chromosome arms nor to cytobands. The representation per chromosomal arm given in Figure 5.5 is adopted only for the sake of conciseness. The Basal subtype is associated with the largest number of aberrations, with deletions on Chromosomes 4, 5, 14 and 15, and amplifications on Chromosomes 6, 10 and 12. The ERBB2 subtype has only an amplification on the q arm of Chromosome 17, covering the genomic position where the ERBB2 gene is located. This is a known aberration, and the results suggest that this is the only aberration that differentiates this subtype from the other samples. The fact that this known aberration is found, also serves as a positive control for the SIRAC algorithm. The Luminal A subtype is characterized by a strong amplification on Chromosome 1 and a deletion on Chromosome 16. The Luminal B has less pronounce aberrations on Chromosomes $1, 8, 12$ and 20.

We compared our findings with the conclusions of Bergamaschi *et al.* [Berg 06] that also searched for aberrations associated with subtypes on a *different* aCGH dataset. They first used the CLAC algorithm [Wang 05a] to determine per sample the chromosomal gains and losses, then discretized the information per cytoband. Finally they use the SAM analysis to identify the aberrations correlated with the class labels. The aberrations found by them are summarized in Figure 5.5 (b). In the Basal subtype, 6 of the 7 aberrations found by applying SIRAC to the *NKI* dataset are also in their list. The ERBB2 subtype only has the amplification on Chromosome 17, as in our findings. In the Luminal A subtype the strong amplification on Chromosome 1 is present while the one on the p arm of Chromosome 16 only reaches significance for an $\mathrm{FDR} = 0.05$. In fact, as it is visible from Figure 5.4 (c), on the q arm of this chromosome many relevant DNA-probes show a deletion, while fewer DNA-probes on the p arm, although present, are not significant. In the Luminal B subtype, one of the three regions found by us is also present in Bergamaschi *et al.* [Berg 06] results.

Some of the differences between our results obtained on the *NKI* dataset and Bergamaschi results can be explained by the fact that our algorithm targets only the aberrations specific for a given class when compared to the rest of the samples. Therefore, we don't have the same aberrations for two subtypes. This is, for example, the case for the amplification on Chromosome 17 that is present both in the Basal and ERBB2 subtype for Bergamaschi *et al.* [Berg 06] while it is only a feature of the ERBB2 subtype in our results. Similarly, the amplification on the q arm of Chromosome 1 is a strong aberration only in the Luminal A subtype in the *NKI* dataset, while Bergamaschi *et al.* [Berg 06] reported it for both the Luminal A and the Basal subtypes. Another aspect to take into account is that we choose an $\mathrm{FDR} < 0.005$ for the identification of the relevant DNA-probes by the SAM analysis. This rather strict value limits the number of false positives, and enables us to highlight the stronger aberrations. We repeated the experiments with a less strict constraint, i.e. using a FDR smaller than 0.05 or 0.1. The results for the $\mathrm{FDR} < 0.05$ are shown in Figure 5.5 (c). Four more aberrations were detected in the Basal, two of which are present in Bergamaschi *et al.* [Berg 06] (the amplification on Chromosome 7, and the deletion on the q arm of Chromosome 12). The ERBB2 still shows only the amplification on Chromosome 17. In the Luminal A subtype we detected one more amplification on the p arm of Chromosome 16, in agreement with the results of Bergamaschi *et al.* [Berg 06]. On the other hand, we find a few more aberrations for the Luminal B subtype, but these did not match the findings of Bergamaschi *et al.* [Berg 06].

Overall, given the differences in the datasets and in the methodology used, we can see striking similarities in the subtype characterization of the cancer. Especially the Basal, the ERBB2 and the Luminal A subtypes seem better defined, while the Luminal B type, seems rather weak, and we advocate that a better definition of this subtype needs to be established.

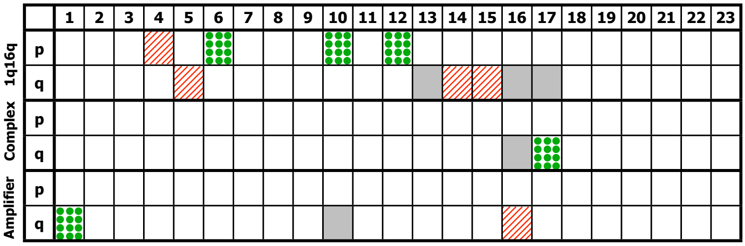
As stated earlier, we simply chose to represent the detected aberrations in terms of chromosome arms in order to ease the comparison with Bergamaschi *et al.* [Berg 06].

However, such a representation does not highlight a very useful feature of the SIRAC algorithm: the scale space. The scale space allows evaluation of aberrations at different genomic resolutions, and the number of scales across which an aberration remains significant can also be employed to judge the importance of a region, for a fixed SAM-FDR. By employing this feature, one can zoom in on potentially interesting regions, where the aberration has a larger average amplitude, and is of medium length (see Figure 5.1 (Step 3)). When increasing the number of scales (s) across which an aberration should be significant, the number of DNA-probes in significant regions across the genome is typically reduced strongly. More specifically if, for the *NKI* dataset, s is changed from s=2 to s=9, the number of DNA-probes in significant regions decrease from 174 to 56 for the Basal subtype (68% reduction), 76 to 31 for ERBB2 (59% reduction), 135 to 86 (36% reduction) for Luminal A and 33 to 7 (79% reduction) for Luminal B. Therefore, if only copy number is employed to identify putative regions (genes), the scale space analysis provides a powerful tool to reduce the list of genes putatively involved in the studied process.

The *Fridlyand* dataset

Recently, Fridlyand *et al.* [Frid 06] analyzed the aberrations of 67 breast cancer samples. First they smoothed each sample using Circular binary segmentation [Olsh 04], and defined chromosomal aberrations per sample. Based on the clustering of the smoothed data they identified three subtypes, i.e. the *1q16q*, the *Complex* and the *Mixed amplifier* subtypes. The *1q16q* subtype is named after the only copy number aberrations detected, i.e. a gain on 1*q* and a loss on 16*q*. The *Complex* subtype is characterized by many low level copy number alterations, mainly ER negative tumors, and worse outcome than the others subtypes. The *Mixed amplifier* subtype tumors were both ER positive and ER negative and did show several aberrations. They analyzed the aberration frequency in each subtype in order to find patterns of chromosomal changes across samples.

We applied our algorithm to their data, analyzing each subtype against the remaining samples. Figure 5.6 summarizes our findings. We identified a loss on the q arms of Chromosomes 16 and 4 for the *1q16q* and the *Complex* subtypes respectively, and the amplifications on Chromosomes $8, 16$ and 20 for the *Mixed amplifier* subtype. The comparison with the conclusions of Fridlyand *et al.* [Frid 06] is not straightforward, since their goal was not to identify aberrations specific for one class. Their results consist in a frequency plot for each subtype of the copy number changes more frequently associated with it. More specifically, they show the frequency of the clone aberrations present in more than 50% of the samples of one subtype and in less than 30% of the samples in the other subtypes. This illustration is not clearly pointing out the differences between subtypes, since often a percentage of the same aberration is present in two or more subtypes. However, our findings show correspondences with the results of Fridlyand *et al.* [Frid 06]. They define the class *1q16q* as exhibiting an amplification on Chromosome 1 and a deletion on Chromosome 16. We only detect the deletion on Chromosome 16. We think that the aberration on Chromosome 1, which is not detected by our algorithm, may be not specific for this class. From the data it is apparent that this amplification is present in all samples, i.e. not specific for the *1q16q* subtype. Other aberrations detected by our algorithm reflect a pattern in the frequency plot of Fridlyand *et al.* [Frid 06], such as for the deletion in 4*q* of the *Complex* subtype and the amplification in 8*q* of the *Mixed am-*



		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1q16q	p				▨		●				●		●											
	q					▨								■	▨	▨	■	■						
Complex	p																							
	q																■	●						
Amplifier	p																							
	q	●									■						▨							

Figure 5.6: Summary of the aberrations per chromosome arms for the *Fridlyand* dataset. The deletions are depicted with inclined lines, and the amplification with circles, the uniform gray boxes indicates that the aberration was significant not in the class of interest but in the rest of the samples.

plifier subtype. In other cases, such as the amplifications on Chromosomes 16 and 20 in the *Mixed amplifier* class, our findings are not reflected in the frequency plot of Fridlyand *et al.* [Frid 06]. In conclusion, the results of SIRAC and Fridlyand *et al.* [Frid 06] exhibit partial overlap. The advantage of our algorithm is that it better highlights the differences between subtypes and clearly points out the specific chromosomal aberrations.

5.4 Discussion and conclusions

We have presented a method to identify aberrant chromosomal regions that are specific for the problem under study. Our emphasis is not on the identification per sample of a chromosomal gain or loss, but we strive to evaluate what makes two classes different from each other, and what are the aberrations that distinguish them. We also want to limit the number of preprocessing steps, in order to reduce the set of inevitable parameters to be tuned. This motivated us to avoid the characterization per sample of the DNA-probes being amplified or deleted, which is instead the necessary input data for the STAC algorithm [Disk 06] and the approach followed by Fridlyand *et al.* [Frid 06]. We chose to use the raw data as input and assumed that a DNA-probe amplified/deleted in one class and not in the other is selected as significant by the SAM analysis. Of course the researcher has to choose the appropriate false discovery rate. This decision influences the number of DNA-probes preselected as relevant. This is an important starting point of our algorithm. We opted for a low false discovery rate for all the problems analyzed. The different number of relevant DNA-probes selected in the distinct cases already gave us an indication of the number and the strength of the chromosomal aberrations. For example in the *NKI* dataset the largest number of relevant DNA-probes was present in the Basal subtype, while the ERBB2 class was associated with only a few DNA-probes mainly on Chromosome 17.

Our algorithm is designed to identify the copy number alterations in the aCGH data. The core of the algorithm resides in the identification of the regions of chromosomal aberration. We assumed that an aberration involves more than a single DNA-probe. Therefore, we tested in a systematic manner the candidate regions, i.e. the locations in the vicinity of the DNA-probes identified by the SAM analysis. The use of different window sizes allows us to detect different lengths of copy number changes and not to miss aberrations in regions sparsely covered by the aCGH probes. Since for the samples in the *NKI* data also the expression is available, we tested if similar results could be obtained by applying our algorithm to the expression directly, as Furge *et al.* [Furg 05] did. However, the assumption that an over/under expression should involve more than a single gene here does not hold anymore. Even if a region is amplified, not all genes may be active and, therefore, differentially expressed with respect to the reference. Moreover, while in the aCGH data the only cause of aberration resides in the copy number variation, the variance in the expression is due to multiple factors. In general, we observed in our expression dataset that the relevant genes selected by the SAM analysis were scattered across the genome and, therefore, no clear regions of significance were identified. This result further indicates that the detection of genomic aberration using gene expression datasets should be performed with caution, and results should always be validated with other tests, such as FISH or PCR, if not with genomic copy number data itself.

Instead, the expression data can be used to perform a post-processing step on the algorithm applied to the aCGH data. Once the aberrated regions have been identified, the expression data allows for a further analysis of the genes present in these regions. For example, the genes can be prioritized according to the correlation between the expression and the aCGH data, or according to the ability of each gene to distinguish between the classes of interest. This is especially relevant since we expect that, for instance, not all genes in a region of aberration will be active, some may be silent and not contributing to the mechanism of cancer. A selection can be done based on this additional information source, resulting in a smaller list of potentially interesting genes to be further analyzed. The benefits of the use of the expression data are exemplified by the ERBB2 subtype in the *NKI* dataset. The genes present in the amplified region of Chromosome 17 were ranked according to the product of the p-value of the t-test (computed on the gene expression and class labels) and the p-value of the correlation between the expression of each gene and its closest DNA-probe. The top two genes are the ERBB2 gene itself and the GRB7, i.e. the growth factor receptor-bound protein 7. This is expected since the ERBB2 subtype is characterized by the amplification of the ERBB2 gene, and the GRB7 is found to be over-expressed and co-amplified with the ERBB2 gene [Kaur 01, Dres 03, Reya 05]. Therefore, a combined approach of SIRAC and the use of gene expression is a powerful additional tool in the search for marker genes.

In the SIRAC algorithm we first detect associations of single probes with the class label, and then search for regions that are enriched for class label associated probes. This is advantageous especially when working with tumor samples. The heterogeneity of the tumors may lead to signals for the aberrations smaller than the ones expected if the sample cells were homogeneous. Therefore, amplifications/deletions with small absolute values may be of interest as well, especially when they discriminate the classes of interest. Several authors (e.g. Saramaki *et al.* [Sara 06], Fridlyand and Chin *et al.* [Frid 06, Chin 06], and Nymark *et al.* [Nyma 06]) have recently pointed out that even low-level copy number aberrations may have significant effects on the gene-expression and, therefore, on the cell functioning and tumor development.

The error rate control of SIRAC is performed in two different steps. First the null-hypothesis being constructed during the permutation steps of the SAM procedure, second, the Bonferroni correction for multiple testing applied to the p-values of the hypergeometric test. The artificial experiment illustrates how the dependencies between these two steps may lead to an anti-conservative control of the error rate. The choice of the parameter s, which combines the outcomes of different window sizes, plays an important role. The artificial experiments suggests that the stricter the value, e.g. $s = 9$, the better the control of the error rate. However, this is achieved at the expenses of the sensitivity. Therefore, less conservative choices, e.g. $s = 2$, may be used. In this case, the p-values of the hypergeometric test need to be interpreted with caution. The SIRAC algorithm, however, provides useful details, such as the number of window sizes in which each DNA-probe was judge significant, that can be used to further prioritize the regions. Moreover, if the expression data is available, further validation of the aberrations may be performed by investigating the correlation with the expression of the genes in the identified region.

In conclusion, we focused on the identification of the chromosomal aberrations that discriminate between the classes of interest and proposed a robust algorithm for the evaluation of their significance. Our algorithm does not require preprocessing of the data such as discretization or smoothing, and uses a limited number of parameters. Our findings on the two breast cancer datasets are in agreement with previous studies, and better highlight the dissimilarities between the classes of interest.

6

Integration of DNA copy number alterations and prognostic gene signatures to predict prognosis of patients with breast cancer

This chapter extensively describes the implications of genome copy number alterations in breast cancer. Special attention is devoted to 68 *samples selected from the NKI cohort [Vijv 02], for which both copy number and expression data were available. The regions of aberrations identified with SIRAC, have been further investigated by analyzing the expression of the genes on the same genomic location. The objective has been to identify the genes that were affected by the copy number alterations and have major functional involvement in breast cancer development.* [1]

---

[1] This chapter will be submitted to Cancer Research [Horl ed]

6.1 Introduction

Invasive breast cancers are a diverse group of tumors whose clinical behavior is complicated to predict. Currently, clinical and pathological prognostic factors such as, age, lymph node status, tumor diameter, histological grade, HER2 gene amplification or protein over-expression and estrogen receptor status, are employed to identify patients at relatively high risk of developing metastases. It is of utmost importance to identify those patients because they benefit most from adjuvant systemic treatment. For this reason, several decision making tools have been developed [Blam 07, Gold 07, Ravd 01]. However they do not perfectly predict the exact clinical behavior of breast tumors and as a consequence patients may be over-treated or under-treated. Additional factors are therefore needed to guide decisions on adjuvant systemic treatment.

In recent years high-throughput technologies such as gene expression micro-arrays have offered new opportunities to improve the ability to determine individual prognosis in breast cancer. Studies of gene expression have identified expression profiles and gene sets that are prognostic for patients with breast cancer (e.g. 70-gene prognosis signature [Veer 02], molecular subtype signature [Hu 06, Pero 00, Sorl 03], wound signature [Chan 05a], chromosomal instability signature [Cart 06], genomic grade index [Soti 03]). These signatures are strong indipendent prognostic factors and may show agreement in the outcome predictions for individual patients [Fan 06]. Large-scale genomic analyses of breast tumors also suggest the relevance of molecular subtypes of breast cancer, (e.g. Basal-like, Luminal A, Luminal B, HER2+ and normal-like, may exist [Hu 06, Pero 00, Sorl 03]), although the underlying mechanisms that drive these expression patterns remain unknown.

Differences in gene expression between breast tumor samples indicate genetic [Poll 02, Jone 02] and epigenetic [Jone 02, Wids 02, Nova 06] changes, or reaction to changed activities of transcriptional regulators in cancer [Visv 03]. Genetic alterations (i.e. DNA copy number alterations) are key mechanisms in human breast cancer development. Array Comparative Genomic Hybridization (aCGH) has identified a number of recurrent regions of DNA copy number alterations in human breast tumors and cell lines. Some of the recurrent regions in breast tumors contain known or candidate oncogenes (FGFR1; 8p11), (MYC; 8q24), (CCND1; 11q13), (HER2; 17q12) and tumor suppressor genes (RB1; 13q14 and TP53; 17p13) and have been shown to be useful for risk stratification [Dell 02, Al K 04, Han 06, Hick 06, Lete 06]. Association between gene dosage and gene expression levels has been demonstrated, and a significant proportion of gene expression variation can be explained by DNA copy number [Poll 02]. Several studies have correlated DNA copy number changes with mRNA expression variation of individual genes [Hyma 02, Monn 01, Orse 04, Orse 05] and more recently with gene expression signatures [Berg 06, Chin 06, Chin 07b, Chin 07a, Frid 06]. For example, Adler *et al.* [Adle 06] showed that the wound response signature is induced by coordinated amplifications of the transcription factor MYC (8q24) and CSN5 (8q13) in breast cancer. Other prognostic signatures, including the Basal-like [Pero 00] and the 70-gene [Veer 02] signatures, were not activated by MYC and CSN5 [Adle 06]. In two other studies is shown that Basal-like tumors were associated with an amplification of 6p21-25 [Berg 06] and 10p14 [Adel 07].

Integration of DNA copy number alterations and their co-expression genes may provide even more opportunities to improve the ability to determine individual prognosis in breast cancer. We, therefore, performed integrative analysis of DNA copy number and gene expression (mRNA) within 68 primary breast carcinomas sub-classified based on well known clinicopathological features, the 70-gene prognosis signature and the breast cancer molecular subtypes [Hu 06, Pero 00, Sorl 03]. The genomic regions we have identified together with their co-expressing genes are putative drug targets, and could potentially include genes with major functional involvement in breast cancer development, or may act as prognostic markers in breast cancer patients.

6.2 Material and Methods

Patient selection

Breast tumor specimens were collected from a series of 295 consecutive women with breast cancer from the fresh-frozen tissue bank of the Netherlands Cancer Institute/Antoni van Leeuwenhoek Hospital (NKI/AVL), for which gene expression and clinical data were previously published [Vijv 02]. We selected 68 samples according to the following criteria. Previously, Hu *et al.* [Hu 06] created the Single Sample Predictor (SSP) to classify tumors according to molecular subtypes [Pero 00, Sorl 03]. For the SSP, a mean expression profile was created for each subtype (Luminal A, Luminal B, Basal-like, HER2+, Normal-like). All 295 samples were compared to each centroid and assigned by the SSP to the nearest centroid/subtype as determined by Spearman correlation. Next we ranked all 295 samples by correlation score for each of the 5 subtypes. 68 samples, all with a high correlation score with one of the subtypes, were selected. The samples consisted of 21 Basal-like, 10 HER2+, 21 Luminal A, 12 Luminal B and 4 Normal-like tumors. Of these, 44 were identified as having the 70-gene poor prognosis signature and 24 with the good prognosis signature. This study was approved by the Institutional Review Board of the Netherlands Cancer Institute.

Clinicopathological characteristics

All tumors were primary invasive breast carcinoma less than 5 cm in diameter at pathological examination (pT1 or pT2). The age at diagnosis was 52 years or younger. All patients had been treated by modified radical mastectomy or breast-conserving surgery, including dissection of the axillary lymph nodes, followed by radiotherapy if indicated.

All 68 patients had no prior malignancies, except adequately treated in situ carcinoma of the cervix (CIN) or non-melanoma skin cancer and did not receive any systemic therapy before surgery (33 had lymph-node negative disease and 35 had lymph-node positive disease, 1 patient with lymph-node negative disease received adjuvant systemic therapy consisting of chemotherapy, and 31 lymph-node positive patients received chemotherapy (22), hormonal therapy (5), or both (4). The median duration of follow-up was 10.5 years (range, 1.78 to 21.23) for the 47 patients without metastasis as the first event and 5.4 years (range, 0.71 to 14.37) for the 21 patients with metastasis as the first event. Follow-up information was done by individual chart review until January 1 2005. The median follow-up among all 68 patients was 8.9 years (range, 0.71 to 8.9).

Immunohistochemistry

All 68 breast carcinomas were included in tissue microarrays. We used a manual tissue arrayer (Beecher Instruments, Silver Spring, MD, USA) following previously described techniques. Eight Core tissue biopsies of 600-μm cores were taken from each individual paraffin-embedded tumor and arrayed in triplicate in a new paraffin block. Serial sections of 3 μm were cut from the tissue microarray blocks, deparaffinized in xylene, and hydrated in a graded series of alcohol. Staining was performed using the Lab Vision Immunohistochemical Autostainer (Lab Vision Corporation, Fremont, CA, USA) with primary antibodies towards estrogens receptor-α (ER; 1D5+6F11, dilution 1:50, Neomarkers, Lab Vision Corporation, Fremont, CA, USA), progesterone receptor (PgR; R-1, dilution 1:500, Klinipath, Duiven, Netherlands), HER2 (3B5, dilution 1:3000) (van de Vijver *et al.*, 1988) and Keratine 5/6 (D5/16 B4; dilution 1:100; Dako). Detection was performed using the antigen retrieval method (citrate pH 6.0).

PCR, Sequencing, and Mutational Analysis

TP53 mutations were identified by temporal temperature gradient gel electrophoresis (TTGE) followed by DNA sequencing as described [Geis 01]. PIK3CA mutations were identified on 10-100 ng of genomic DNA using a standard protocol. PCR products were purified over a QIAquick spin column (QIAGEN) and were sequenced using the BigDye Terminator Cycle Sequencing Kit (Applied Biosystems) and an ABI 3730 automated capillary sequencer. For all PCR products with sequence variants, both forward and reverse sequence reactions were repeated for confirmation.

Scoring of Immunohistochemistry

Staining for ER and PgR was interpreted as negative when no tumor cells were stained and positive when more than 10 % of tumor cells showed staining. A sample was considered to be HER2 positive when either a strong membrane staining (3+) could be observed by IHC or CISH revealed amplification of HER2 in samples with weak (1+ or 2+) membrane staining at IHC, and tumors were considered positive for Keratin 5/6 if at least 1% of tumor cells showed staining. For two patients ER expression level was estimated on the basis of the hybridization results from the microarray experiments, which is a reliable assay for ER status [Gong 07].

Genomic DNA isolation and labeling

Twenty-five sections each 30-μm thick were used for DNA isolation. Before and after cutting one slide was stained with hematoxylin and eosin to select samples with 50% or more tumor cells. Frozen tissue sections were digested in 15 ml TNE (100 mM sodium chloride, 10 mM Tris-HCl (pH = 8.0), 25 mM EDTA (pH = 8.0) and 1% sodium dodecyl sulphate (SDS)) and proteinase K (50 μg/ml) was added and incubated at 55°C for minimum of 24 hours. Subsequently phenol-chloroform extraction was performed followed by ethanol precipitation. DNA was diluted in 10 mM Tris-HCl (pH = 7.6)/ 0.1 mM EDTA. For all cases, aCGH was performed using 2 μg of genomic DNA. All labeling reactions were performed with the Cy3 and Cy5 conjugates from the Universal Linkage System (ULS, Kreatech Biotechnology, Amsterdam the Netherlands) [Raap 04]. Labeling efficiency for ULS-Cy3 and ULS-Cy5 were calculated from A260 (DNA), A280 (protein), A550 (Cy3) and A649 (Cy5) after removal of unbound ULS, on a NanoDropsND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA).

Microarray hybridization and data preprocessing

As described previously [Joos 07], hybridizations were done on microarrays containing 3.5k BAC/PAC derived DNA segments covering the whole genome with an average spacing of 1Mb. The whole library was spotted in triplicate on every slide (Code Link Activated Slides, Amersham Biosciences, Piscataway, NJ, USA, Prod. No.30001100) Data processing of the scanned microarray slide included signal intensity measurement using ImaGene Software (BioDiscovery, Inc., El Segundo, CA, USA) followed by median print tip normalization. Intensity ratios (Cy5/Cy3) were log2-transformed and triplicate spot measurements were averaged. This resulted in a 3277 x 68 data matrix used for further analysis.

Data analysis

Frequency of gains and losses Based on the copy number levels, the frequency for gain and loss for all BAC clones was calculated using fixed log2-ratio thresholds of 0.2 and -0.2 respectively. All data will be made publicly available at `http://research.nki.nl/vandevijverlab`.

Unsupervised analysis BRB Array Tools, (http://linus.nci.nih.gov/BRB-ArrayTools-.html), was employed to perform hierarchical clustering. Unsupervised analysis was carried out using the complete linkage-clustering algorithm based on a centered Pearson correlation similarity matrix of the raw DNA copy number levels of 68 primary tumors and 3277 BAC Clones. Chi-square tests were employed to study the relationship between gene-expression profiles, DNA copy number and clinicopathological characteristics, performed using SPSS, version 15.0 (SPSS Inc. Chicago, Illinois, USA). Results were considered statistically significant when the p-value was smaller than 0.01.

Supervised analysis We used SIRAC (Supervised Identification of Regions of Aberration in aCGH datasets) [Lai 07] to identify chromosomal regions which are associated with the classes defined by prognostic gene expression signatures or clinicopathological characteristics of the tumor. The SIRAC algorithm focuses on the aberrations specifically associated with the sample labeling being studied by incorporating the labels (e.g. 70 gene poor prognosis versus 70 gene good prognosis signature) in the analysis. More specifically, as a first step, a SAM analysis [Tush 01] is employed to select DNA-probes that discriminate between the classes of interest at a chosen false discovery rate (FDR). We required that the selected DNA-probes have an FDR smaller than 0.05. We call these significant DNA-probes the "relevant" probes. An illustrative result is shown in Figure 5.1, Step 1. Next we tested whether the number of relevant DNA-probes in a given genomic region is higher than expected by chance. For this we used the hypergeometric test applied at each given genomic region, and tested whether the fraction of relevant DNA-probes represents a significant enrichment. A Bonferroni correction for multiple testing was applied by multiplying the p-value of each test by the number of tests performed. Regions of aberration with a corrected p-value smaller than 0.05 were considered significantly enriched for genomic aberrations. This step was repeated for different window sizes in order to detect both small and large aberrations (Figure 5.1, Step 2). Finally, regions of aberrations were identified based on a consensus between the results of the different window sizes (Figure 5.1, Step 3). Importantly, no discretization, smoothing or segmentation algorithms are applied to the aCGH data prior to the SIRAC analysis. This avoids the parameter optimization step these models usually require, rendering the results independent of these choices.

Integration of aCGH and gene expression data We used the expression data to perform a post-processing step on the output of SIRAC. For all 68 breast tumors, gene expression profiling was previously published [Vijv 02]. Both expression probes (oligonucleotides) and DNA copy number probes (BAC clones) were ordered by position as assigned by NCBI-Build32 http://genome.ucsc.edu/cgi-bin/hgGateway) on the genome. Only probes for which a genome location was found were used, resulting in an aCGH dataset of 2952 BAC clones and a gene expression dataset with 10986 genes. Association between DNA copy number and mRNA gene expression levels was calculated using the Pearson correlation between genes and the relevant probes in the DNA copy number data, revealing DNA dosage sensitive genes. The correlation p-value is the p-value associated with the correlation between the DNA copy number and gene expression data (mRNA). The t-test p-value is the p-value of the t-test for each gene, which quantifies the ability of that gene to distinguish between the classes of interest (for example good versus poor outcome classes). We prioritized relevant genomic regions that (1) were found in at least 10% or more of the class of interest; (2) had a correlation p-value smaller than 0.01 with the corresponding genes; and (3) had a t-test p-value for each gene smaller than 0.01. Finally, we evaluated if the DNA dosage sensitive genes were: (A) putative drug targets; (B) potential biomarkers; (C) transcription factors as defined in the TRANSFAC database [Maty 03]; (D) belong to the 70-gene prognosis signature [Veer 02]; (E) belong to intrinsic gene list of the molecular subtypes [Hu 06]. We compared and validated our results on aCGH and mRNA expression data from independent breast cancer samples published in earlier reports [Berg 06, Chin 06, Chin 07a, Adel 07].

6.3 Results

Here we report the detected chromosomal aberrations, their locations and their relationship with clinicopathological characteristics, prognostic gene expression profiles, potential candidates of drug targets and prognostic markers in 68 primary breast cancers.

Frequent aberrations in breast carcinomas

Figure 6.1 represents a frequency plot summarizing the distribution of aberrations in all 68 tumors. A fixed threshold of $\lg_2 < 0.2$ for gain (circles) and $\lg_2 < -0.2$ (xs) for loss was used. Based on these thresholds, we observed frequent DNA copy number alterations gains at 1q (34% of tumors), 8q (24%), 17q (10%) and 16p (7%), and losses at 13q (13%), 16q (13%), 23p (11%) Four regional (> 10 MB) gains (1q41, 1q31-32, 8q21-8q24, 8q12, 17q24) and five regional losses (13q21, 13q13-14, 16q21, and 16q12) were observed in

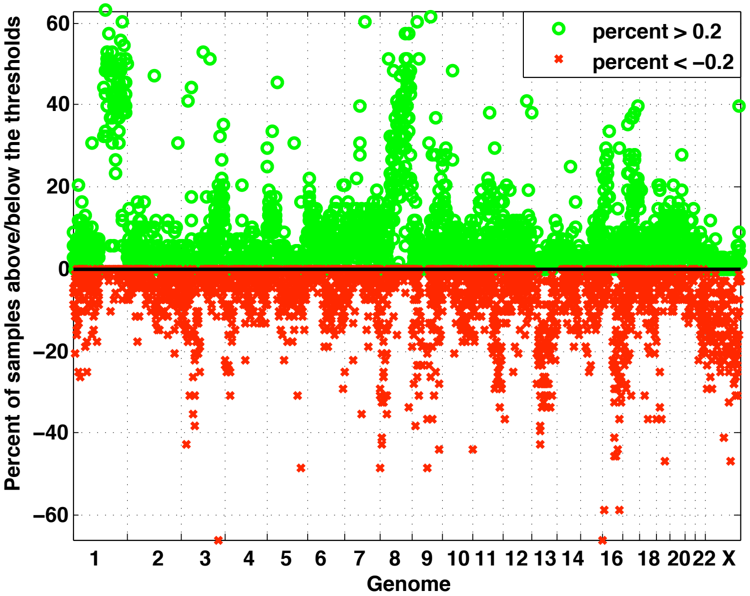


Figure 6.1: Frequency plot summarizing the distribution of aberrations in all 68 tumors. A fixed threshold of $\lg_2 > 0.2$ for gain (circles) and $\lg_2 < -0.2$ (x mark) was used.

$\geq$ 15% (10 out of 68) tumors. These observations are consistent with earlier reported aCGH based breast cancer studies [Berg 06, Chin 06, Chin 07b, Chin 07a, Frid 06].

Unsupervised hierarchical clustering of aCGH data

Figure 6.2 represents an unsupervised hierarchical clustering of the 68 breast tumors according to their DNA copy number profile. Columns represent individual tumor samples while the assignment to the different prognostic gene signatures and clinicopathological characteristics are shown in the rows. Hierarchical clustering of the samples revealed a subdivision into four clusters, as indicated in Figure 6.2. We found a strong association between the samples in clusters 1, 2 and 4 and characteristics of poor prognosis (chi-square test, p-value <0.01). Interestingly, patients in cluster 4 have frequently more genomic losses (chi-square test p-value 0.014) compared to patients in the other clusters. Patients with good prognosis characteristics were found in cluster 3 (chi-square test p-value <0.01) Figure 6.2.

Supervised analysis of aCGH data

As we observed distinct DNA copy number profiles for tumors with good and poor prognosis, we wanted to discover the exact associations between DNA copy number alterations and clinicopathological characteristics and prognostic gene expression profiles. To this end we employed the SIRAC algorithm to perform a supervised analysis for each of the clinicopathological characteristics and prognostic gene signatures. The results identified with SIRAC for the 70 gene prognosis signature and molecular subtypes are

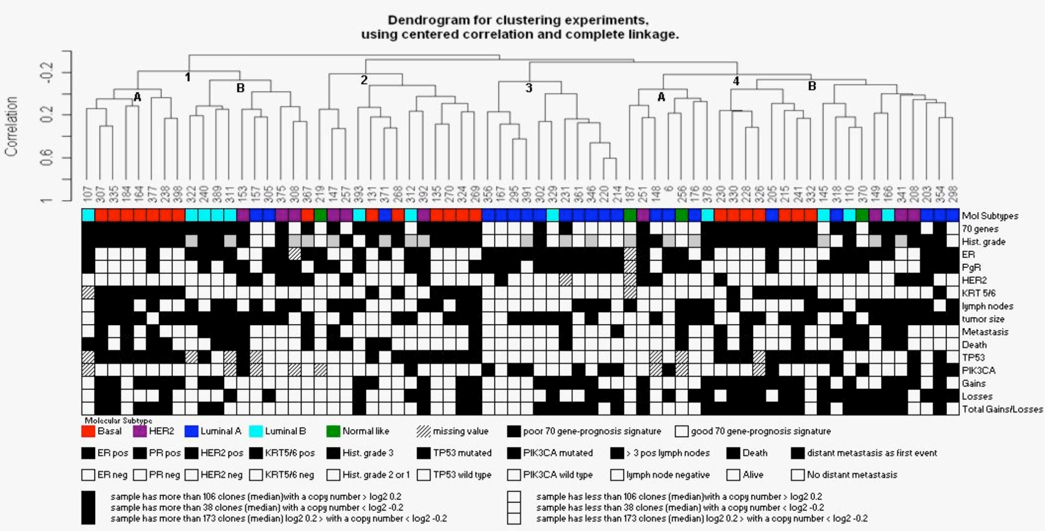
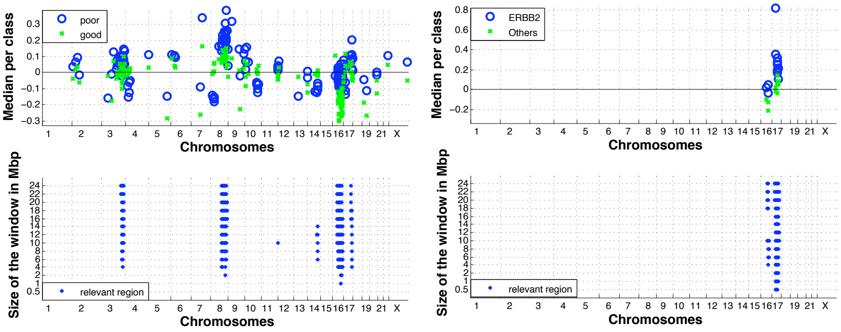


Figure 6.2: Hierarchical clustering of aCGH data measured in 68 breast tumors. The columns contain the samples and the rows the various clinicopathological parameters and the gene expression signature assignments. Fore clusters were identified, and each cluster can be subdivided in 2 sub-clusters, indicated as A and B.

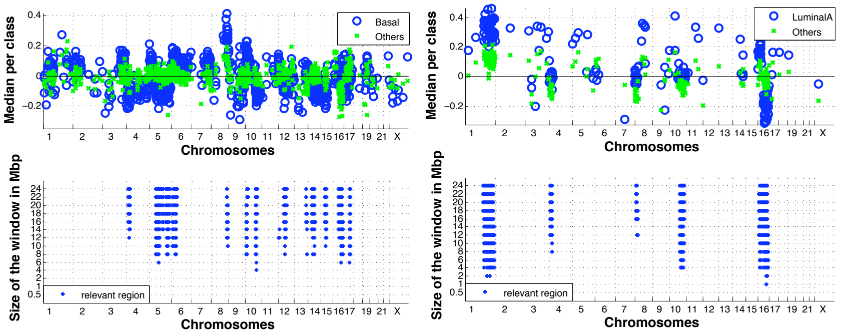
given in Figure 6.3. A graphical overview of the aberrations per chromosomal arm per prognostic gene signature or clinicopathological characteristic is given in Figures 6.5 and 6.4 the representation per chromosomal arm was adopted only for the sake of conciseness.

DNA copy numbers alterations and clinicopathological characteristics We identified several gains and losses associated with clinicopathological characteristics, as illustrated in Figure 6.4. ER negative and high grade tumors showed a higher frequency of gains and losses than ER positive and low grade tumors. ER negative and grade three tumors showed both gains at 6p12 and 6p21, and losses at 4p15, 10q23 and 14q12. Additionally gains at 10p14-15 and losses at 4q23-24, 5q21-22, 10q22-24 and 12q13 were identified in ER negative tumors, while high grade tumors showed additional gains at 8q22-24, 12p13 and 17q24-25 and losses at 8p21-22 and 15q14-21. Low grade tumors (histological grade 1 and 2) and ER positive tumors showed only losses including 16q11-13, 16q21-24 and 13q31-34. We identified a loss at 4p15-16, in TP53 mutant tumors and losses at 16q11-13 and 16q22-24 in tumors with a PIK3CA mutation, but no specific additional alterations Losses at 16q11-13 and 16q22-24 without additional aberrations were found in tumors with favorable prognosis of breast cancer, including the following characteristics: ER positive, histological grade 1, PIK3CA mutated.

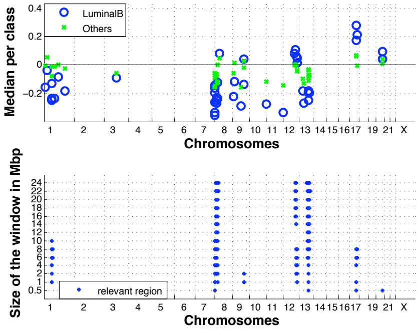
DNA copy numbers alterations and prognostic gene signatures We found that the tumors with a 70-gene poor prognosis gene signature were associated with gains at 3q26-27, 8q22-24 and 17q24-25 and losses at 14q31. Tumors with a 70-gene good prog-



(a) 70 genes (b) ERBB2

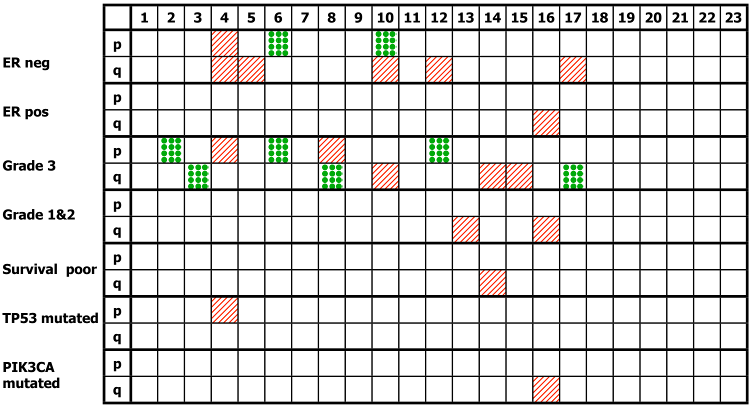


(c) Basal (d) LuminalA



(e) LuminalB

Figure 6.3: The results of the SIRAC algorithm on the 68 breast tumors for each of the gene expression signatures. We show the aberrations per chromosome with respect to the class of interest versus the other classes. The class of interest are: (a) the poor prognosis group derived from the 70 gene signature, (b) the HER2 positive cases (n=10), (c) the Basal-like cases (n=21), (d) the Luminal A (n=21) and (e) the Luminal B (n=12). In the top plot of each panel, the relevant DNA-probes detected by the SAM analysis are displayed. For each relevant DNA probe a circle and an x mark are plotted at its location on the genome, representing the median of the class of interest and the median of the remaining samples, respectively. The lower plot depicts the regions of aberration detected across different window sizes, with the window size on the vertical axis.



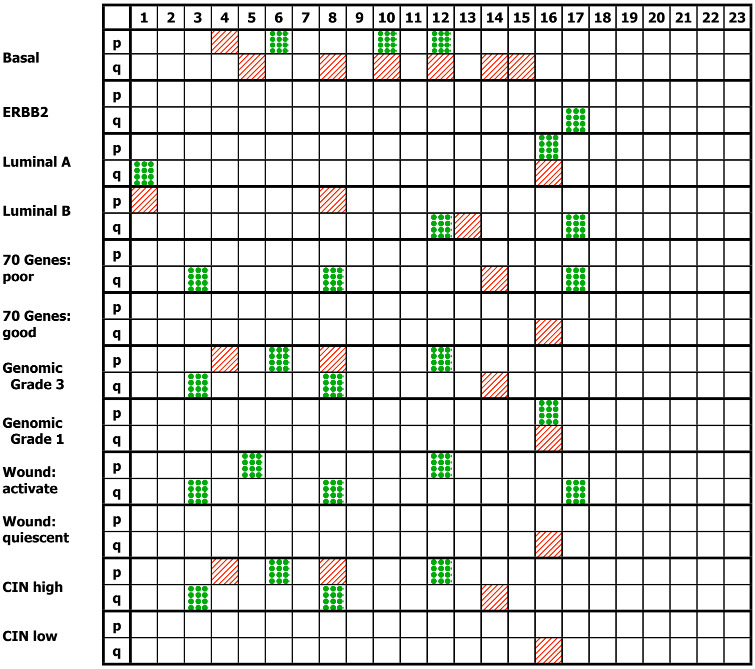
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
ER neg	**p**				▨		●				●													
	q				▨	▨					▨		▨					▨						
ER pos	**p**																							
	q																▨							
Grade 3	**p**		●		▨		●		▨				●											
	q			●					●		▨				▨	▨		●						
Grade 1&2	**p**																							
	q													▨			▨							
Survival poor	**p**																							
	q														▨									
TP53 mutated	**p**				▨																			
	q																							
PIK3CA mutated	**p**																							
	q																▨							

Figure 6.4: A summary of the results of the SIRAC algorithm on the 68 breast tumors for the clinicolpathological labels. We show the aberrations per chromosomal arm with respect to the class of interest. The regions of aberration detected by SIRAC have been coded based on the direction of the aberration. Circles was used for gain, inclined lines for losses. The representation per chromosomal arm was adopted only for the sake of conciseness. For example, for TP53 mutated, the samples with a P53 mutation represent the class of interest and these tumors are characterized by a loss on 4p, hence the block filled with inclined lines in cell 4p. For Er negative, the ER negative samples are the class of interest and these tumors are characterized by a gain on 6p and 10p, hence the block filled with circles in the corresponding cells.

nosis gene signature were associated with losses at 16q11-13 and 16q22-24 (Figures 6.3 and 6.5). The Basal-like tumors were associated with the largest number of aberrations. Copy number gains were found in chromosomal regions 6p12, 6p21-24, 8q24, 10p12-14, 12p13 and losses including 4p15, 5q11-14, 5q21-34, 10q23-24, 12q13-15, 14q12-23 and 15q14-21. For the Luminal A subtype we identified gains at 1q21-41, 16p11-13, and losses at 16q11-13 and 16q22-24. The Luminal B had less pronounced aberrations with gains at 12q23-24 and 17q23 and losses at 1p31, 8p21-23, 13q22 and 13q31. The HER2+ subtype showed only gain on the q arm of Chromosome 17 (17q12 and 17q21-23), covering the genomic position where the HER2 gene is located. This is a known aberration, and the results suggest that this is the only aberration that differentiates this subtype from the other samples.

Finding DNA dosage sensitive genes in prognostic gene expression signatures

We used gene expression data to perform a post-processing step on the output of SIRAC. The power of this combined approach (SIRAC followed by an expression analysis) is illustrated by the HER2+ subtype. A total of 80 genes showed gain correlated with up-regulated mRNA expression Using our filtering criteria to search for potential drug targets or prognostic biomarkers we found 20 potential candidates in two cytogenetic regions, 17q11.1-17q12 and 17q21.32-23.2, which were associated with HER2+ breast tumors.



		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Basal	p				▨		●				●		●											
	q					▨			▨		▨		▨		▨	▨								
ERBB2	p																							
	q																	●						
Luminal A	p																●							
	q	●															▨							
Luminal B	p	▨							▨															
	q												●	▨				●						
70 Genes: poor	p																							
	q			●					●						▨			●						
70 Genes: good	p																							
	q																▨							
Genomic Grade 3	p				▨		●		▨				●											
	q			●					●						▨									
Genomic Grade 1	p																●							
	q																▨							
Wound: activate	p					●							●											
	q			●					●									●						
Wound: quiescent	p																							
	q																▨							
CIN high	p				▨		●		▨				●											
	q			●					●						▨									
CIN low	p																							
	q																▨							

Figure 6.5: A summary of the results of the SIRAC algorithm on the 68 breast tumors for the prognostic gene signatures. Again, we show the aberrations per chromosomal arm coded based on the direction of the aberation: inclined lines is used for deletion, and circles filling is used for the amplifications.

Only 17q12 showed high level amplification with a median copy number of lg$_2$ = 0.82. The other regions showed copy number gain. The top two ranked genes were the HER2 gene itself and GRB7, i.e. growth factor receptor-bound protein 7. This was expected since the HER2+ subtype is characterized by the amplification of the HER2 gene, and GRB7 is found to be over-expressed and co-amplified with HER2. Both genes were found in 80% of the samples that were classified as HER2+. Therefore, a combined approach involving SIRAC followed by a gene expression-based post-processing is a potentially powerful tool to search for putative drug targets, candidate regulators, or biomarkers for a specific molecular breast cancer subtype or prognostic gene signature. In addition, DNA copy number gain of 84 genes showed correlation with down regulated mRNA expression. This result suggests that other mechanisms, such as epigenetics, may play a role regulating mRNA expression levels in HER2+ breast tumors.

In the case of Basal-like tumors, significant correlation between copy number gain and upregulated gene expression was found for 199 genes at 6p21.1-21.33, 6p22.1-22.2, 6p23, 8q24.21-24.33 and 10p12.33-10.14. The top 10 ranked genes were KCNK5, KIFC1, PIM1, BYSL, (6p21), C10orf38/LOC221061, C10orf7/CDC123, SEPHS1, MCM10 (10p13) and NDRG1 (8q24). A total of 43 genes showed losses correlated with downregulated mRNA expression, including: CPEB3 (10q23.33), MYG1, SMUG1 (12q13.13) HER3, PYM, RNF41, MBC2, KIAA1002, (12q13.3), ZFYVE19 (15q15.1), EID-1, MYEF2, SORD1 (15q21.1), SLIT2 (4p15.31) KIAA0303 (5q12.3), MRPS27 (5q13.2), RBM22, DCTN4 (5q33.1) Seven genes including C10orf38, C10orf7/CDC123, SUV39H2, KIN, HSPA14, C6orf108 and CPEB3 belonged to the original 1300-gene signature of Hu *et al.* [Hu 06]. Recently, many studies have illustrated that Basal-like tumors are associated with worse prognosis. As we showed, the region of 6p21 is commonly amplified in Basal-like tumors and harbors several candidate oncogenes, including DEK, E2F3, NOTCH4, PIM1, and CCND3. A gene identified as DNA dosage sensitive gene in Basal-like breast cancer was VIM (vimentin) at 10p13, which is expressed in the myoepithelial layer of the glandular breast cells, and moreover can be used to distinguish Basal-like from Luminal breast tumor cancers. In the study of Korsching *et al.* [Kors 05] vimentin-expressing carcinomas revealed a significantly higher average absolute number of cytogenetic alterations per case, but a significantly lower frequency of chromosome 16q losses compared to vimentin-negative cases. This is consistent with our findings that Basal-like breast carcinomas exhibit more gains and losses and a lower frequency of loss at Chromosome 16q.

Among 277 genes associated with the Luminal A type subtype, 40 genes (1q21.3-1qter, 16p13.12-13) showed gain correlated with upregulated gene expression. Thirty-eight of which were located at 1q21-44. The top ranked gene (KCTD3) and three others (RAB13, MUC1, PPP2R5A) belonged to the original signature of the molecular subtypes [Hu 06, Pero 00, Sorl 03]. Losses correlated with downregulated mRNA expression were found in 127 genes of which 28 potential candidates. All 28 genes were located at 16q11.2-24.1 including NOC4, DC13, BM039, CKLF which belonged to the original molecular subtype signature [Hu 06, Pero 00, Sorl 03]. Moreover it was suggested by Naderi *et al.* [Nade 07] that BM039 (CENPN) is part of a 'core' prognostic signature, because it was also identified as one of the 231 genes that correlated with prognostic categories in the original van 't Veer paper [Veer 02]. Luminal A defined tumors showed frequent high amplification at 1q and loss at 16p12-13. Two interesting genes in these regions are MUC1 and DICER1. MUC1 showed gain and overexpression in Luminal A tumors and belonged to the original signature of Hu *et al.* [Hu 06]. MUC1 is expressed at the Luminal surface of the mammary gland and was associated with many indicators (i.e. low tumor grade, ER+, PgR+ and absence of distant metastasis) of good prognosis. DICER 1, located at 1q41, was recently discovered to be involved in human breast cancer; as an RNase III endonuclease, it is an essential component of the microRNA machinery. We found significant changes in the expression of DICER1 (t-test p-value= 5.04E-04) between Luminal A and other molecular subtypes. This result is consistent with earlier findings of Blenkiron *et al.* [Blen 07] which showed that DICER1 and AGO2 were higher expressed in the less aggressive Luminal A type tumors than in the more aggressive Basal-like, HER2+ and Luminal B subtypes. In line with our findings, Zhang *et al.* [Zhan 06a] recently discovered high frequency copy number abnormalities of Dicer1, AGO2, and other miRNA associated genes in breast cancer. Interestingly, a recent study showed that conditional deletion of Dicer1 enhanced tumor development in a K-Ras induced mouse model of lung cancer [Kuma 07]. Together these data suggest that DICER1 deregulation might be involved in the etiology of human breast cancer.

In case of gain correlated with up-regulated mRNA expression associated with Luminal B type tumors only 4 out of 41 genes were potential candidates. Strikingly, they are all located at 17q23.2, including TLK2, PSMC5, CCDC44 and SMARCD2. In case of loss and down-regulated Luminal B tumors showed a distinct loss at 8p21.3-8p23.1, 12 out of 13 filtered genes where located on this region, including; MGC29816, DPYSL2, FLJ10569, XPO7, SH2D4A, LZTS1, PDLIM2, PDGFRL, TUSC3, C8orf16, LONRF1, PRAGMIN, of which only PRAGMIN belonged to the original 1300-gene signature of Hu *et al.* [Hu 06].

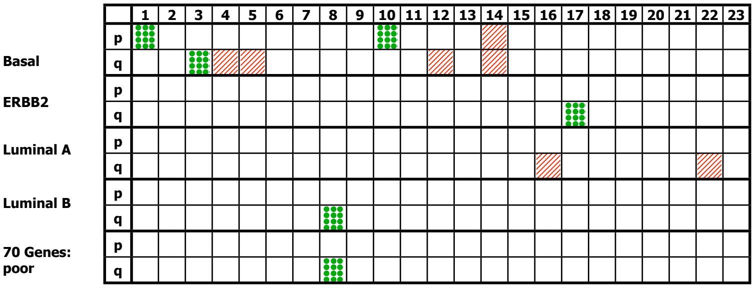
Twenty-four genes that were associated with 70-gene poor prognosis signature showed DNA copy number gain correlated with upregulated mRNA expression. Three out of the 24 genes (CCNE2, LYRIC both at 8q22.1, EXT1 at 8q24.11 appeared in the original 70-gene prognostic signature and 11 (BIRC5, TK1, EVER1 (17q25.1), EIF4G1, POLR2H (3q27.1), LAPTM4B (8q22.1), ZNF706, WDSOF1 (both 8q22.3), CML66 (8q23.1), SQLE, ATAD2 (8q24.13)) belonged to the original 1300-gene signature of Hu *et al.* [Hu 06] indicating their importance in breast cancer. Loss correlated with down-regulation mRNA expression in association with a 70-gene poor prognosis signature was found in three genes; nevertheless none of them passed our filtering criteria of being potential candidate of prognostic marker or putative drug target.

On the other hand loss correlated with down-regulation was found in 123 genes associated with a 70-gene good prognosis signature, of which 14 could act as possible candidates. All 14 genes were located at 16q and were included in the 28 genes that were associated with the Luminal A subtype. The presence of common candidate genes in different prognostic gene signatures suggest the existence of common candidate genes of breast tumorigenesis.

Data comparison with DNA copy number studies of breast cancer

First we compared the molecular subtypes specific regions in the aCGH data with two recent DNA copy number studies of molecular subtypes of breast cancer [Berg 06,Adel 07]. Worth noticing is that we have found striking similarities of aberrated regions in the Basal-like, HER2+ and the Luminal A subtype. They seem better defined than the Luminal B subtype. The association between loss of 4p15, 5q11-35, 14q23 and gain at 6p12, 6p21, 10p12-13, 12p13 with the previous identified Basal-like molecular subtype by gene expression profiling, is now found in three independent breast cancer aCGH datasets. However, there are numerous discordant regions, which may be due to different selection criteria to include patients in the studies or differences in methodology employed to measure DNA copy number levels. We could only compare the Luminal A and B subtype with the data of Bergamaschi *et al.* [Berg 06], as Adelaide *et al.* [Adel 07] did not subdivide the Luminal type tumors into A and B. Interestingly, the Luminal A specific gain on Chromosome 1q12-41 and 16p12-13 was also presented by Bergamaschi *et al./*,and Adelaide also found strong association between these regions and Luminal type tumors. In the Luminal B subtype, only the region of 8p12-13 was also found by Bergamaschi *et al.* [Berg 06].

Next we compared altered regions that were associated with clinicopathological characteristics with 6 previous published aCGH datasets [Berg 06,Chin 07b,Frid 06,Adel 07, Loo 04,Ness 05]. Most commonly estrogene-negative breast tumors showed loss of 5q21-35 and gain of 6p21-22. The HER2+ subtype with a know amplification at 17q12-21 was found in all independent datasets. Association between loss of 5q21-25 and TP3 mutated tumors was not supported in our dataset, as we found only loss at 4p15-16.



		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Basal	p	●									●				▨									
	q			●	▨	▨							▨		▨									
ERBB2	p																							
	q																	●						
Luminal A	p																							
	q																▨						▨	
Luminal B	p																							
	q								●															
70 Genes: poor	p																							
	q								●															

Figure 6.6: Results of the Sirac algorithm applied to the Chin dataset. Summary of the aberrations per chromosome arm for the four different subtypes (Basal-like, HER2+, Luminal A and Luminal B) and the 70-gene prognosis. The numbers in the top of the tables denotes the chromosomes. Again, an arm is indicated with an inclined lines when a significant region is found on that arm that shows a deletion of the clones of interest. Similarly, a circles filled block indicates amplification.

Validation of DNA dosage sensitive genes with integrated genomic studies

aCGH data were used as acquired from the Supplementary Data of Chin K. *et al.* [Chin 06]. Preprocessing and normalization was performed and described by the authors. Because no exact mapping information was available for all clones in the Chin dataset, we gave the clones a 3-bp length centered on the mapping position as supplied by Chin *et al.* [Chin 06]. We removed all clones with more than 50% missing values. We imputated the remaining missing values using the averaged values of their two positional neighbors. Probes mapped to the same area were averaged and represented as a single clone. This resulted in 2149 unique clones. Gene expression data were also acquired from Chin *et al.* [Chin 06] (Arrayexpress accession number: E-TABM-158). Probes not mapping to a single ENSEMBL ID were removed; probes mapping to Y chromosome genes were also removed. This resulted in 21339 unique Affymetrix probe measurements. First, we applied our SIRAC algorithm to this data, analyzing 70 gene good prognosis versus poor prognosis and each of the molecular subtype against the remaining samples. A summary of the results are given in Figure 6.6. Second, we computed the common regions between the two datasets, and then select only the genes in those regions on the two datasets. Together with matched array expression data we identified common genomic regions in both datasets showing strong coordinate expression changes and associated with: Basal-like tumors with losses on 5q12.3-5q33.3, 12q13.11-12q15, and 14q13.12-14q23.1; Luminal A with loss on 16q21-16q24.1; HER2+ with gain on 17q11.2 and 17q22; and a 70 gene poor prognosis with gain on 8q22-24.23.

Recently, Chin S.F. *et al.* [Chin 07a] profiled DNA copy number alterations in 171 breast tumors using high-resolution genome-wide profiling. Together with matched array expression data they identified genomic regions showing strong coordinate expression changes ("hotspots") and frequently amplified hotspots on 8q22-8q24,3, on 8q22.3 (EDD1, WDSOF1) and on 8q24.11-13 (THRAP6, DCC1, SQLE, SPG8). The regions 8q22.3 and 8q24.13 were also identified in our data and previously identified in prognostic gene signatures (70-gene signature [Veer 02], wound signature [Chan 05a], and molecular subtypes [Hu 06]). Also interesting, CCNE2 (cycline E2) at 8q22.1 was one of three overlapping genes between the 76-gene prediction signature from Wang *et al.* [Wang 05b] and the 70-gene signature from van't Veer *et al.* [Veer 02], both signatures being predictive of distant metastasis-free survival (DMFS) in lymph node negative patients. These candidate genes were also previously identified in wound signature [Chan 05a] and molecular subtypes [Hu 06]. These results suggest that DNA copy number information at 8q22-24.23 could act as potential candidates of prognostic markers in breast cancer.

6.4 Conclusions

In this report we established a link between genetic changes and prognostic gene expression profiles and clinicopathological parameters that determine tumor cell behavior in breast cancer. Our goal was to identify DNA copy number alterations that may harbor putative drug targets or target genes with major functional involvement in breast cancer development by comparing genome wide DNA copy number with gene expression data. Using SIRAC, we identified associations between DNA copy number alterations, clinicopathological parameters and prognostic gene signatures. We found that gains and losses varied between prognostic gene signatures and clinicopathological features. As expected, gains and losses were more frequently found in tumors with unfavorable prognostic features (i.e. histological grade 3, ER negative, HER2 positive, 70-gene poor prognosis, Basal molecular subtype). In particular, gains at 3q22-27, 6p12-22, 8q21-24, 17q12-25 and losses at 4p15-16 and 14q21-31 have been associated with an unfavorable prognosis of breast cancer. Loss at 16q11-13 and 16q22-24, without additional aberrations, was found in tumors with the following characteristics: ER positive, histological grade 1, PIK3CA mutated, and are associated with a favorable prognosis. We compared the identified associations between DNA copy number alterations and clinicopathological parameters and prognostic gene expression signatures in independent aCGH datasets where gene expression data was also available for the same tumors [Berg 06, Chin 07b]. These results suggested that DNA copy number information of 8q22-24.23 could act as potential candidates of prognostic markers in breast cancer.

Since gene expression profiling is based on mRNA, an instable molecule, and because of the stability of DNA, DNA copy number alterations can be more useful as prognostic and/or predictive markers in breast cancer patients. Integration of DNA copy number information and gene expression signature help in establishing a link between genetic changes and gene expression signatures that determine tumor behavior in breast cancer.

7

A genome-wide search for spatial organization of copy-number and expression dependencies

In this chapter our interest has been on the detection of causal spatial dependencies and interactions between copy number and expression alterations across the whole genome. An unsupervised extension of the SIRAC algorithm has been developed to highlight the patterns of correlation between the two data types. The new algorithm (IGDam) extends the search for spatial dependencies from the one dimensional space of the copy number data to the two dimensions of the combined copy number and expression data.

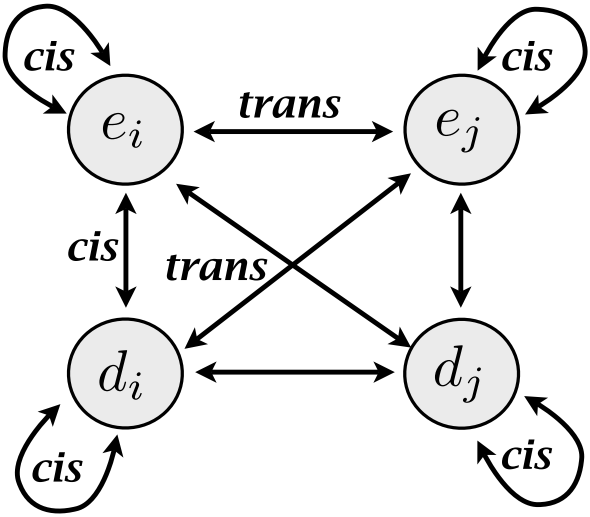


Figure 7.1: A graph diagram of the hypothesized interactions between copy numbers (d) and expression measurements (e) in two generic genomic locations i and j. The cis-effects occur within probes in the same region i, while the trans-effects involve the probes on different regions (i and j).

7.1 Introduction

In cancer, gene expression arrays have been widely used for prognosis prediction [Veer 02, Vijv 02, Chan 05a, Rama 03, Wang 05b, Mill 05, Cart 06], better cancer stratification in classes which share the same phenotype [Pero 00, Sorl 01, Sorl 03], and response to therapy prediction [Ayer 04, Chan 05b, Hann 05, Ma 04]. In addition, array comparative genome hybridization (aCGH) allows the analysis of copy number alteration of thousands genomic regions, represented by copy number measurements, on a single slide simultaneously. Since genomic alterations in DNA copy number are important events in cancer development [Leng 98], the identification of chromosomal aberrations is a powerful instrument in studies of cancer [Bert 03, Pink 05].

Expression and aCGH data are known to be dependent [Hyma 02, Furg 05, Levi 05, Yi 05]. The spatial organization of the probes provides essential information to investigate the dependencies between different parts of the genome and the different data types. Two main types of spatial interactions can be hypothesized in aCGH and expression data. A dependency between probes that are closely located (so called cis-effect), and an interrelation between probes that are located in different genomic regions, e.g. on different chromosomes, i.e. trans-effects. A schematic diagram of correlations is given in Figure 7.1. The nodes of the graph represent the copy number data d and the expression data e in the regions around the genomic locations i and j. The arrows represent the dependencies (cis and trans effects).

Various aspects of cis-effect have been previously investigated in the literature. The relationship between copy number measurements within the same region (cis-effects on d's in Figure 7.1) is used to identify regions of copy number aberration within a sample [Lai 05]. The copy number measurements closely located that exceed a certain threshold are judged to be aberrated [Poll 02, Velt 03, Call 05, Nayl 05, Schw 04]. Alternatively, segmentation and clustering approaches are proposed to divide the aCGH profile in piecewise constant segments [Pica 05, Jong 03, Jong 04, Wang 05a].

Examples of cis-effects on the expression measurements (cis-effects on e's in Figure 7.1) use the genomic location of the genes to estimate regions of chromosomal instability [Call 06, Levi 05, Dres 03, Yi 05], or epigenetic effects [Reya 05, Stra 06, Furg 05]. For example, Reyal *et al.* [Reya 05, Stra 06] have developed the TCM method to determine a region containing neighboring genes, which showed correlated expression profiles. They have pointed out that these regions of correlation are not only due to genomic aberrations, but also to epigenetic mechanisms, such as chromatin regulation, and co-regulation by the same transcription factor. They have focused on the epigenetic effects and identified a region where the loss of expression was due to histone methylation of the genes in the region.

Several studies investigate the dependencies between DNA and expression probes (cis-effect between d_i and e_i in Figure 7.1) [Adle 06, Chin 06, Hyma 02, Frid 06, Poll 02]. Adler *et al.* [Adle 06] have applied a SAM analysis [Tush 01] to the aCGH data in order to identify the copy number measurements which distinguish the two classes of interest. Then, they have focused on the copy number measurements that are correlated with the expression data. Hyman *et al.* [Hyma 02] have determined the copy number alteration of each copy number measurement based on a threshold, and then have quantified the Signal to Noise Ratio (SNR) of the expression of the corresponding gene and the copy number measurement. Fridlyand *et al.* [Frid 06] have identified aberrations in the aCGH data, and have investigated the correlation with the expression of a different dataset. Pollack *et al.* [Poll 02] have compared the expression measurements with the aCGH data discretized into five levels (deletion, normal, low,medium and high amplification).

To the best of our knowledge, only cis-effects between aCGH and expression data have been investigated so far. Here, we explore both cis and trans-effect dependencies between DNA copy number and expression datasets in a systematic way. We aim at identifying the effect that the copy number alterations have on the genome activity. Figure 7.2 illustrates our view. The big square represents a matrix storing a measure of dependencies, such as Spearman Correlation, between each copy number and expression measurement. In the rows of the matrix are the copy number measurements along the genome, while the column shows the genes sorted on their location. If there is a one-to-one correspondence between DNA and expression probes, then the matrix is a square, as in Figure 7.2. The diagonal represents the dependencies between copy number and expression measurements of the same genes. The cis-effects happen between neighboring probes. Box 1 along the diagonal illustrates the positional enrichment between the DNA and expression probes around location i. In contrast, the trans-effects involve probes that are not within each other neighborhood. In Figure 7.2, Box 2 represents the position of such an effect between the copy number measurements located around position i and the expression of the genes at location j.

In this work we chose the correlation across samples as a measure of dependency between aCGH and expression data, and propose a method to detect the regions on the genome where the positional concentration of high correlation between DNA and expression probes is larger than expected by change. We apply our IGDam algorithm (Identification of Genome-wide Dependencies between aCGH and mRNA datasets) to a

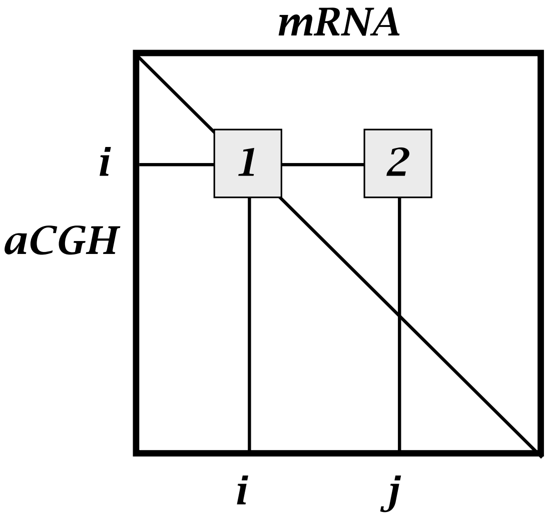


Figure 7.2: Illustration of the cis and trans effects between copy numbers and expression measurements. The cis-effects occur between probes in the same region i (Box 1), while the trans-effects involve the copy number measurements around position i and the expression of genes in location j (Box 2).

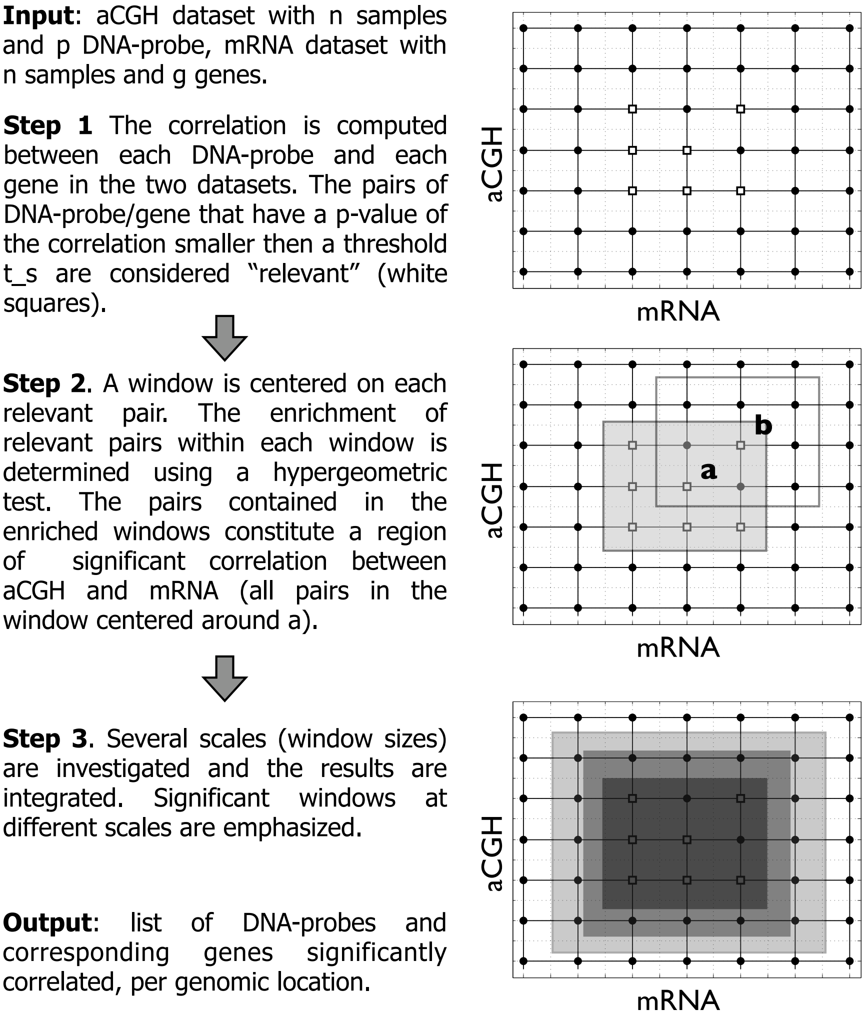
selection of 68 patients from the Netherlands Cancer Institute (*NKI*) cohort [Vijv 02], and discuss the results and possible interpretations.

7.2 Method description

Our approach investigates the correlation between aCGH and expression data. Therefore, these two data types measured on the same samples are required as input. Figure 7.3 illustrates our IGDam method, while a more detailed description is given in Algorithm 4. For the same n samples both an aCGH dataset with p copy number measurements and an expression dataset with g expression measurements are given. We refer to the g measurements of the expression dataset as *genes*.

STEP 1. The correlation is computed between the profile of each copy number measurement and gene over the samples. The value of the correlation test and its p-value are stored in matrices of size $p \times g$. The higher the correlation the smaller the corresponding p-value. We select for further analysis the pairs with a p-value smaller than threshold t_s. These pairs are called *relevant*. An illustration of the thresholded matrix of p-values M is given in Figure 7.3 (Step 1). On the vertical axis is the genomic location of the p copy number measurements, while on the horizontal axis the g genes are ordered according to the genomic position of the accompanying probes. The p-value of each pair copy number measurement/gene is represented with a square or a circle in the grid. The filled back circles represent all the pairs, while the *relevant* ones are depicted with a white square.

STEP 2. In order to investigate the spatial dependencies, we test in a systematic way whether the density of *relevant* neighboring pairs is higher than expected by chance. We



Input: aCGH dataset with n samples and p DNA-probe, mRNA dataset with n samples and g genes.

Step 1 The correlation is computed between each DNA-probe and each gene in the two datasets. The pairs of DNA-probe/gene that have a p-value of the correlation smaller then a threshold t_s are considered "relevant" (white squares).

Step 2. A window is centered on each relevant pair. The enrichment of relevant pairs within each window is determined using a hypergeometric test. The pairs contained in the enriched windows constitute a region of significant correlation between aCGH and mRNA (all pairs in the window centered around a).

Step 3. Several scales (window sizes) are investigated and the results are integrated. Significant windows at different scales are emphasized.

Output: list of DNA-probes and corresponding genes significantly correlated, per genomic location.

Figure 7.3: Illustration of the IGDam algorithm.

center to each *relevant* pair a window of length w. The window length is in the kilobase range and, therefore, has a much higher resolution than the copy number/gene grid depicted in Figure 7.3. Figure 7.3 (Step 2) illustrates two windows centered in the pairs a and b respectively. We evaluate with the hypergeometric test the enrichment of *relevant* pairs, i.e. we test if the number of *relevant* pairs in the window is higher than expected by chance. In order to correct for multiple testing, the Bonferroni correction is applied by multiplying the p-value of the hypergeometric test for the number of tests performed. All pairs in the windows with corrected p-value smaller than the threshold t_h are considered enriched (e.g the pairs contained in the grey filled square in Figure 7.3 (Step 2), while the test applied to the square centered in the pair b does not reach significance.). Note that, although the window of observation is a square in the kilobase pair grid, the numbers of copy number and genes measurements captured, may be different. This is dependent on the resolution of the aCGH and expression arrays. In Figure 7.3 the same resolution is considered, with a one to one correspondence of copy number measurements and genes.

STEP 3. The genomic regions where there is a correlation between copy number and expression measurements may have different lengths. The use of different lengths of the window w increases the ability of detecting different sizes of local dependencies. In our previous work [Lai 07], we have employed a similar scale search. There, we have illustrated with an artificial dataset the benefits of the scale search in detecting chromosomal aberrations of different lengths. Since similar concerns apply here, i.e. the dependencies may have different local characteristics, we also adopt the scale search in this work. Therefore, the procedure described in Step 2 is repeated for a set of scales W. We combine the different scales by counting, for each probe, the number of significantly enriched window scales in which that pair was contained. An illustration is provided in Figure 7.3 (Step 3). The color codes for the number of scales in which the windows W are enriched, from 0 (white), i.e. the locations were not enriched in any scale, to all scales (dark grey), i.e. there was a significant enrichment in all window scales.

The results of the test indicates the genomic locations where the aCGH and the expression data are correlated. This helps us to highlight genomic regions of dependencies between aCGH and mRNA data. Those regions can be studied further in order to improve our understanding of cancer development. The output of the IGDam algorithm is the enriched locations, represented by the correlated copy number and genes measurements.

7.3 Results and Discussion

First the datasets used and the set up of the experiments are described, then the results of the IGDam algorithm are presented. A discussion further investigates several open questions.

7.3.1 Experimental setup

We use a cohort composed by 68 patients selected from the 295 breast cancer samples described in van de Vijver *et al.* [Vijv 02]. The aCGH data is obtained from a BAC array platform [Beer 05], and it consist of 3219 copy number measurements with a value for each of the 68 patients. The expression dataset for the same samples is comprised of 10986 genes, obtained with an oligo array platform [Vijv 02].

As described in Section 7.2, the IGDam algorithm requires a few parameters, i.e. a measure of the correlation, the thresholds t_s and t_h, and the window scales W. As a measure of the correlation we have adopted the Spearman correlation because it is not as sensitive to outliers as the Pearson correlation. We chose a strict threshold $t_s = 10^{-4}$. Therefore, the relevant pairs are the ones that have a p-value of the Spearman correlation smaller than 10^{-4}. The window search approach applies to several window scales the search for significant enrichment of the correlation between aCGH and expression data. These scales should be selected in relation to the resolution of the arrays, because the windows of observation should include more than a single probe. In our data the limiting factor is given by the BAC array, which has a resolution of $\sim$ 1 megabase (Mb). Therefore, the minimum window of observation is set to 1 Mb. The maximum window scale is fixed to 24 Mb because this is roughly half the length of the shortest chromosome. In

Algorithm 4 Description of the IGDam algorithm.

```
Input: aCGH and mRNA datasets $D_{p \times n}$ and $G_{g \times n}$; vector $\mathbf{W}$ with the sizes of the observation windows; thresholds $t_s$ and $t_h$ for the p-values of the Spearman correlation and the hyper-geometric distribution respectively; minimum number of windows scales $s$ for which the pair copy number/genes is judged significant.
Compute the Spearman correlation between each pair copy number/gene in the datasets $D$ and $G$. The p-values are stored in a matrix $M_{p \times g}$.
Evaluate which pairs are *relevant*, i.e. have p-value $<= t_s$. The index are stored in the vectors $\mathbf{X}$ and $\mathbf{Y}$ for the aCGH and mRNA respectively.
Initialize: $C = \{c_{ij} = 0 \, \forall i, j\}$ with $i = 1...p, j = 1...g$, stores the number of scales in which the test is judged significant;
$\forall \, w \in \mathbf{W}$ (for all window scales)
    · center a square window $w$ at each *relevant* pair. Remove the windows that exceed the boundary of the chromosomes where the probes of the included pairs are located: $\mathbf{X}', \mathbf{Y}' \in \mathbf{X}, \mathbf{Y} \,|\, \forall i \; \mathcal{C}h(l^{\mathbf{X}_i} \pm \frac{w}{2}) = \mathcal{C}h(l^{\mathbf{X}_i}) \,\&\, \mathcal{C}h(l^{\mathbf{Y}_i} \pm \frac{w}{2}) = \mathcal{C}h(l^{\mathbf{Y}_i})$, with $\mathcal{C}h$ a function that return the chromosome of the location $l^X$ of the $X^{th}$ probe.
    · Initialize: $P = \{p_i = 0 \, \forall i\}$ with $i = 1...|\mathbf{X}'|$, stores the p-value of the test,
    · $\forall$ pairs $i$ in $\mathbf{X}', \mathbf{Y}'$
        · $h = \sum_{c=0}^{x} \mathcal{H}(c|m, k, o)$, where $\mathcal{H}$ is the hypergeometric test with:
            $x =$ number of *relevant* pairs in the windows $[l^{\mathbf{X}'_i} - \frac{w}{2}, l^{\mathbf{X}'_i} + \frac{w}{2}]$, $[l^{\mathbf{Y}'_i} - \frac{w}{2}, l^{\mathbf{Y}'_i} + \frac{w}{2}]$ on aCGH and mRNA respectively,
            $m =$ number of pairs in the matrix $M$,
            $k =$ number of *relevant* pairs in the matrix $M$,
            $o =$ number of pairs in the windows $[l^{\mathbf{X}'_i} - \frac{w}{2}, l^{\mathbf{X}'_i} + \frac{w}{2}]$, $[l^{\mathbf{Y}'_i} - \frac{w}{2}, l^{\mathbf{Y}'_i} + \frac{w}{2}]$ on aCGH and mRNA respectively.
        · $\mathbf{P}_i^w = 1 - h$;
    · apply the Bonferroni correction: $\mathbf{P}^w = \mathbf{P}^w \times |\mathbf{X}'|$,
    · $\forall i$
        · if $\mathbf{P}_i^w \leq t_h$
            · $\forall \, a, b \,|\, l_a \in [l^{\mathbf{X}'_i} - \frac{w}{2} ... l^{\mathbf{X}'_i} + \frac{w}{2}] \,\&\, l_b \in [l^{\mathbf{Y}'_i} - \frac{w}{2} ... l^{\mathbf{Y}'_i} + \frac{w}{2}]$
                $C_{a,b} = C_{a,b} + 1$.
Output: all locations $\mathbf{X}, \mathbf{Y}$ with $C_{a,b} \leq s$.
```

this way, we enforce that the largest window does not always cover both the p and q arms of the chromosomes. Consequently, we adopt 13 different window scales, i.e. $W \in \{0.5, 1, 2, 4, 6, ..., 12\}$Mb. For each scale the hypergeometric test is computed, and its p-value is corrected for multiple testing. The genomic regions contained in the windows with p-value smaller the 0.005 are considered enriched, i.e. $t_h = 0.005$.

7.3.2 Experimental results

Figure 7.4 (a), (b) and (c) shows the results of the analysis in three different cases. On the vertical axis of each plot is the copy number measurements located along the chromosomes, while the location of the genes is on the horizontal axis. A generic point in row i and column j of the matrix represents the enrichment of the pair i, j across the scales, i.e. the pair i, j was contained in a significantly enriched windows of dimension

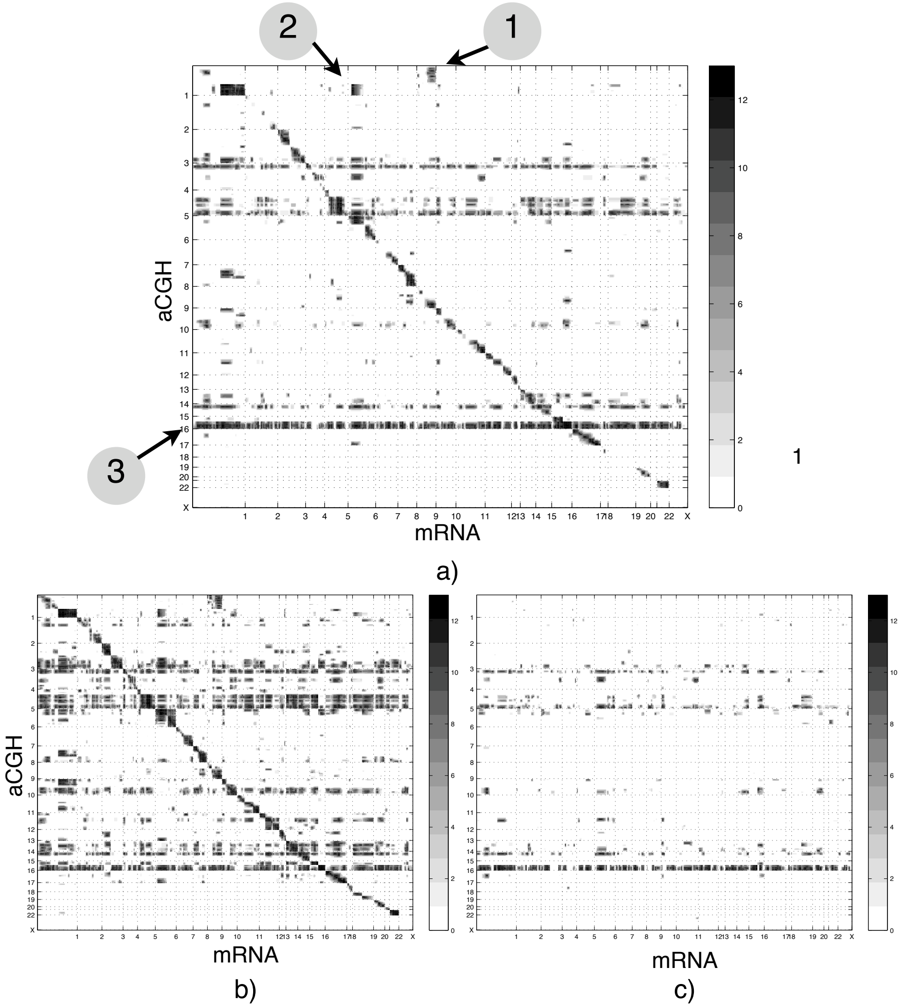
W. The color represents the number of scales in which the pair i, j was part of a window judged significantly enriched. Black means that in all 13 scales there is an enriched window that contains that pair, while white indicates that there was never an enriched window including the pair i, j. In Figure 7.4 (a) the *relevant* pairs used as input for the scale search are the ones with p-value of the Spearman correlation smaller than 10^{-4}. In the plots (b) and (c) the sign of the correlation is taken into account. The input pairs are divided into positive and negative correlation, respectively. Therefore, Figure 7.4 (b) exhibits the enriched genomic regions where the correlation between the aCGH and the expression data is in the same direction. Similarly, Figure 7.4 (c) displays the enriched regions of anti-correlation between the genes and the aCGH probes. Note that there is a significant proportion of negative correlations. In this data the pairs with significant negative correlation account for 37% of the total number of *relevant* pairs.

The first observation that can be made is the evidence of the local correlation, i.e. the copy number and the expression of the genes are locally positively correlated (cis-effects). This is apparent from the fact that the "diagonal" of the matrix that contains the pairs of copy number/gene on the same genomic location, is often enriched in all scales, as can be seen in Figure 7.4 (a) and (b). The negative correlation does not exhibit such enrichment. The black diagonal is, in fact, not present in Figure 7.4 (c). The local correlation was expected, since it as been observed by many authors [Adle 06, Hyma 02, Frid 06, Call 06]. It is, therefore, a positive control of our approach.

Two main types of trans-effects can be observed, a *local* and a *genome-wide* enrichment. A local enrichment of the correlation is visible, for example, on the rows associated with Chromosome 1 in Figure 7.4 (a). The copy number measurements on the p arm of Chromosome 1 are correlated with the expression on Chromosome $9q$ (1 in Figure 7.4). While the q arm of Chromosome 1 is correlated with Chromosome $6p$ (2 in Figure 7.4). No other regions of the genome are significantly enriched along the rows related to the copy number measurements on Chromosome 1. The Spearman correlation that associate with these points is positive. The enrichment is strongly present, in fact, only when the positive highly correlated pairs are considered (see Figure 7.4 (b)), but disappear in the negative correlation plot (Figure 7.4 (c)).

The genome-wide correlation is observable as the horizontal lines of significant enrichment. The most evident is comprised of the copy number measurements located on the q arm of Chromosome 16 (3 in Figure 7.4). This region seems correlated with the expression of many genes along the genome, no matter whether the sign of the Spearman correlation is used in the analysis or not. All three cases ((a), (b) and (c) plots in Figure 7.4) show genome-wide effects on Chromosomes $16q, 15p, 5q$ and $4p$.

At this point, we have not considered any information regarding the copy number aberrations that may be present in the aCGH data. Since we have used all 68 samples, a part of the DNA could be deleted in some samples and amplified in others. In order to evaluate if there is a dependency between the amplifications or deletions and the correlation structures we have observed, we investigate the "potential" amplifications and deletions separately. Instead of using all samples, we selected for each copy number measurement the samples with positive values. The correlation with the expression data is computed using only those samples. The same procedure is done with the negative values of the aCGH probes. The p-values of the correlation is then thresholded and the scale search is performed. The results of the IGDam algorithm with these different *relevant* pairs as input are presented in Figure 7.5. Note that once the Spearman correlation



a)

b)

c)

Figure 7.4: Results on the *NKI* dataset of the window scale search were the input is: a) the *relevant* pairs (p-values of the Spearman correlation $< 10^{-4}$), b) the *relevant* pairs with positive Spearman correlation, c) the *relevant* pairs with negative Spearman correlation.

is computed for the chosen samples, the *relevant* pairs are selected based only on the p-value of the correlation, i.e. regardless of the sign. Figure 7.5 (a) displays the enrichment of the potential amplifications, while Figure 7.5 (b) views the result of the potential deletions. Many of the correlation patterns visible in Figure 7.4, are also present here. However, the frequency of the enrichment across the scales is less pronounced. The diagonal effect is modest, and the number of pairs that are included in enriched windows across all scales is limited. This has to be expected, since now more homogeneous sample groups are considered. The point enrichment between the copy number measurements on Chromosome 1q and the genes on Chromosome 6p is still visible in Figure 7.5 (a)

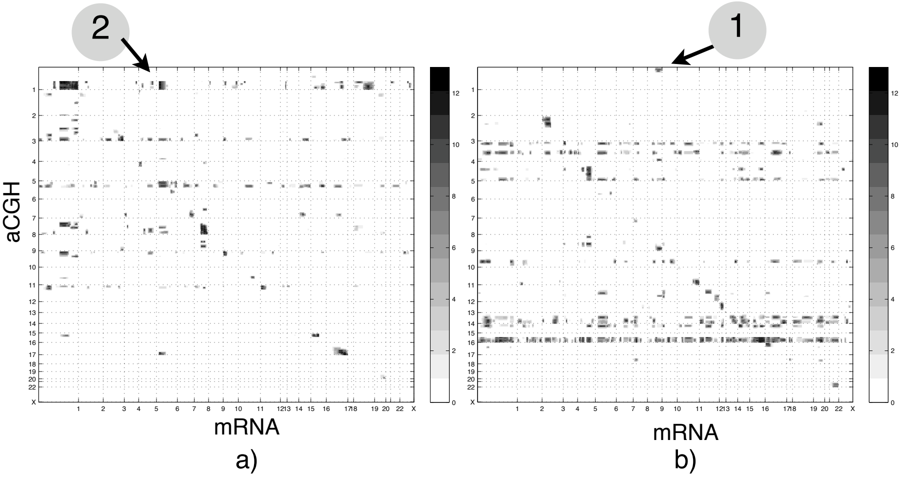


Figure 7.5: Results of the window scales search using as starting pairs the correlation of the samples with a) positive and b) negative values of the copy number measurements.

(2). Here, modest genome-wide effects involve Chromosomes $1q, 3q$ and $6p$, while in Figure 7.5 (b) the genome-wide effects are visible on Chromosomes $4, 5q, 14q$ and especially $16q$.

Interestingly, we do observe a connection between the enrichment and the types of genomic aberrations potentially present in the aCGH data. The local correlation on Chromosome $1q$ is still visible in Figure 7.5 (a), while it is not present in Figure 7.5 (b). This suggests the copy number measurements that have generated it are amplified. The opposite can be observed for the point enrichment between the copy number measurements on Chromosome $1p$ and the gene expression on Chromosome $9q$ (1 in Figure 7.5(b)). It is present when the correlation is computed only between the "potential" deletions. Actually, these genomic aberration are present in the aCGH data, as reported in out previous work [Lai 07]. In the same aCGH dataset we have identified the aberrations for the classes of interest using our supervised algorithm SIRAC. As summarized in [Lai 07] there is, indeed, an amplification on Chromosome $1q$ and a deletion on Chromosome $1p$ in a group of samples, which are predictive of the Luminal A and Luminal B subtypes, respectively. The same findings hold for almost all genome-wide effects identified in Figure 7.5. The ones related with the potential amplifications, in plot (a), involved copy number measurements amplified in one of the subtypes. Similarly, the genome-wide effects detected in the "potential" deletions, see Figure 7.5 (b), are indeed connected to deleted probes.

7.3.3 Discussion

Scale search analysis in a single data type

The results of our algorithm suggest the presence of more than a local correlation between copy number and expression data. Dense regions of significant correlation do occur genome-wide. However, several questions are still open. Is this correlation due to only

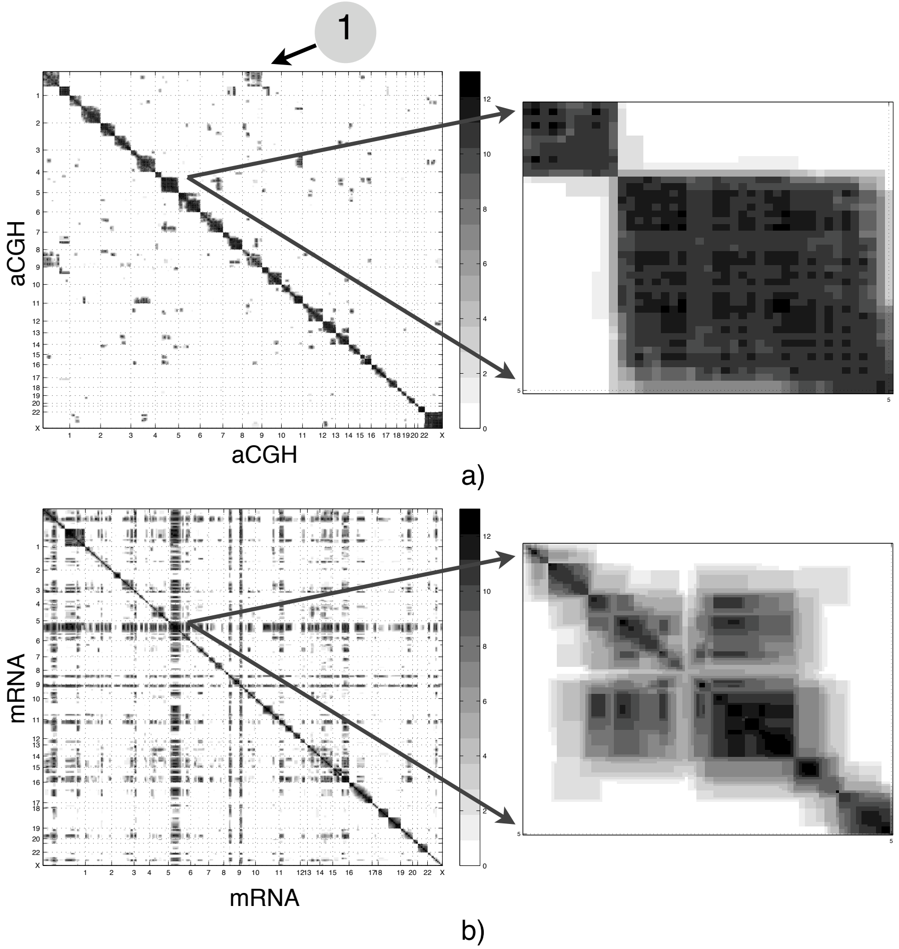


Figure 7.6: Results of the window scale search applied on aCGH data only (a) or on expression data only (b). The thresholded p-value of the Spearman correlation is used as input for the IGDam algorithm.

one type of data, either copy number or expression? Or are the observed effects artifacts?

In order to evaluate the correlation structures within a single type of data, we have applied the scale search analysis to the aCGH and expression data separately. Figure 7.6 presents the results when the thresholded p-values of the Spearman correlation are used as input. Note that the matrices are now symmetric. The enrichment of regions of correlation for the aCGH data is illustrated in Figure 7.6 (a). The highly enriched correlation regions are along the diagonal of the matrix, and often involve the entire chromosome arm. The square black blocks frequently coincide with the boundaries of the chromosome arms, see e.g. the sharp edges on the enlarged Chromosome 5. There is no evidence of correlations as strong as the ones observed in the diagonal, which often involve all window scales (black color). The largest off-diagonal effect is between Chromosomes 1p and 9q (1

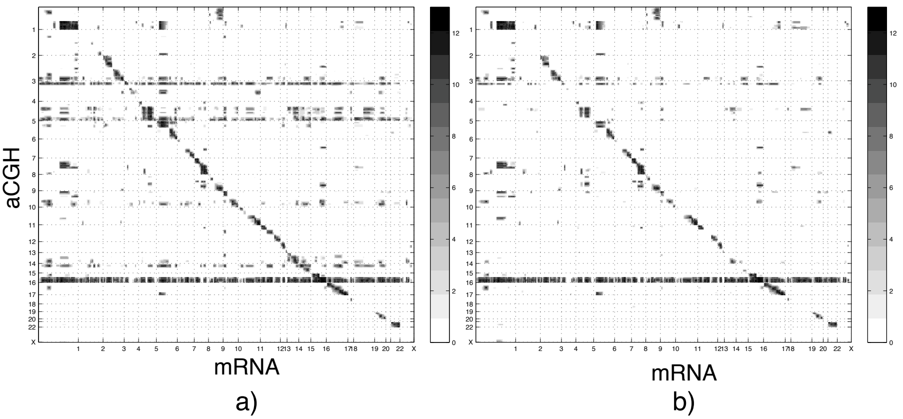


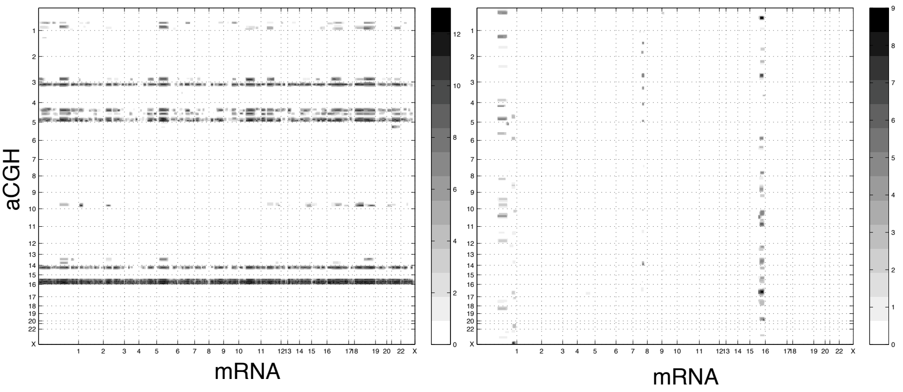
Figure 7.7: Results of the IGDam algorithm using the thresholded p-value only of the probes that have a standard deviation larger than a threshold. The thresholds are determined independently for aCGH and mRNA as the top of the smallest percent of the data. The percentage is fixed in 20% for plot (a) and 50% for plot (b).

in Figure 7.6). The same dependency was present in Figure 7.4 (a) and (b). Note that, however, no local correlation is now present between Chromosomes 1*q* and 6*p*. These results suggest that the aCGH data has very strong local correlation, often involving the entire arm of the chromosomes, but very minor effects can be observed between probes that are not closely located.

Figure 7.6 (b) illustrates the scale search applied to the expression data, with an enlarged plot of Chromosome 5. The spatial correlation is noticeable also in this case. However, the local effects are here much more confined than in the aCGH case. The black diagonal only seldom involves the entire chromosome arm. More genome-wide effects are, instead, visible in the mRNA data, the largest one being on Chromosome 6*p*. Some genome-wide effects were detected also in the combined data, see the mentioned Chromosome 6*p*, others, such as Chromosomes 9*q* and 10*p*, are not present in the results summarized in Figure 7.4. Therefore, the expression data reveals a more complex correlation patterns than the ones observed on the aCGH data, with weaker local dependencies and genome-wide effects.

Effect of the contribution of both data type in the definition of *relevant* pairs

Can the spatial correlation between aCGH and expression achieve significant p-value if either the expression or the copy number have only a small variance across the samples, i.e. the measurements do not contain any information? In order to dispel this doubt we have analyzed the standard deviations (STD) of the probes. First, we compute the STD across samples for expression and aCGH data separately. Second, we set to extract the thresholds to detect which probes/genes have a small variation. The probes/genes with the 20% or 50% smaller STD are considered as having a small variation. Third, we assign the p-value of the Spearman correlation for those pair of probe/genes a value of one. In this way they are not included as *relevant* pairs even if their correlation would be signifi-



a) b)

Figure 7.8: Results of the IGDam algorithm using a randomized matrix of thresholded p-values of the Spearman correlation. The p-values are shuffled independently within each row (a) or within each column (b).

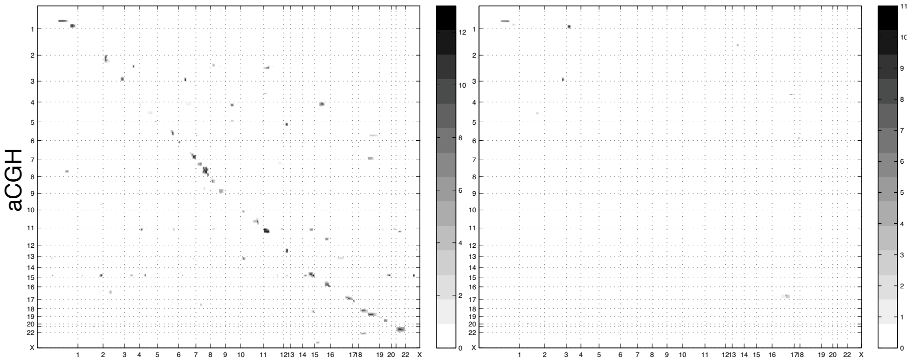
cant. More formally, given the matrix $\mathcal{M}_{p\times g}$ of the p-values of the Spearman correlation, t_p and t_g the thresholds for the copy number and genes measurements respectively:

$$\mathcal{M}_{i,j} = \begin{cases} pvalue_{i,j}, & \text{if } std(\mathbf{p}_i) \geq t_p \,\&\, std(\mathbf{g}_j) \geq t_g \\ 1, & \text{otherwise.} \end{cases} \tag{7.1}$$

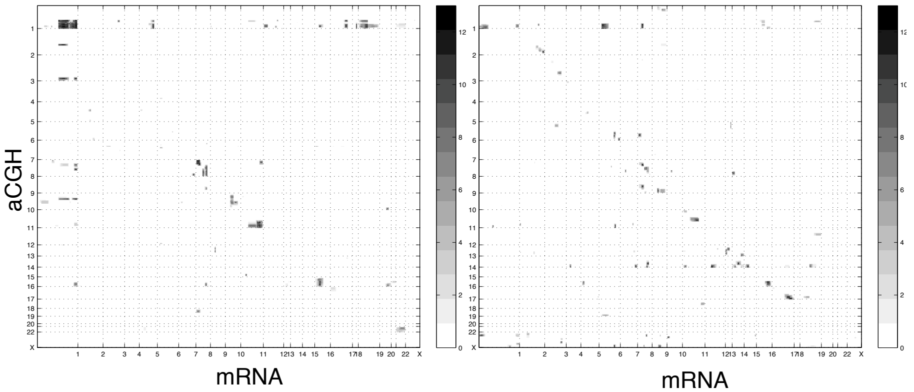
with $\mathbf{p}_i$ and $\mathbf{g}_j$ the vectors of the copy number and expression alterations respectively for the pair i, j. Fourth, the window scale search is performed with the new subset of *relevant* pairs. Figure 7.7 illustrates the results for both threshold values of 20% and 50%, i.e. the *relevant* pairs have STD larger than the thresholds both in the copy number and the expression data. If we compare these results with the corresponding one illustrated in Figure 7.4 (a), we can see that both local and genome-wide regions previously highlighted can still be identified here. When the 20% is used as value to determine t_p and t_g, the percentage of *relevant* pairs is 87% of the total number of pairs selected without any restriction on the STD. When the stringent value of 50% is adopted, 58% of the *relevant* pairs are still selected. This indicates that the pairs selected as input for the scale search have indeed large variation between the samples in both copy number and expression data. Therefore, the high correlation is not only due to a single type of data, but is the result of a variation of both data sources.

Effects of data randomization in the scale search results

In order to investigate whether the patterns of enrichment could be an artifact, we evaluated the scale search applied to randomized pairs of p-values. The randomization is performed on the matrix of the Spearman correlation p-values in two ways, within the rows, by shuffling the gene position of the pairs in each row, or within the columns, by randomizing the copy number measurements within each column independently. The results of the scale search are presented in Figure 7.8. In the left plot the genome location of the copy number measurements is not altered. We do observe genome-wide effects on



a) BASAL b) ERBB2



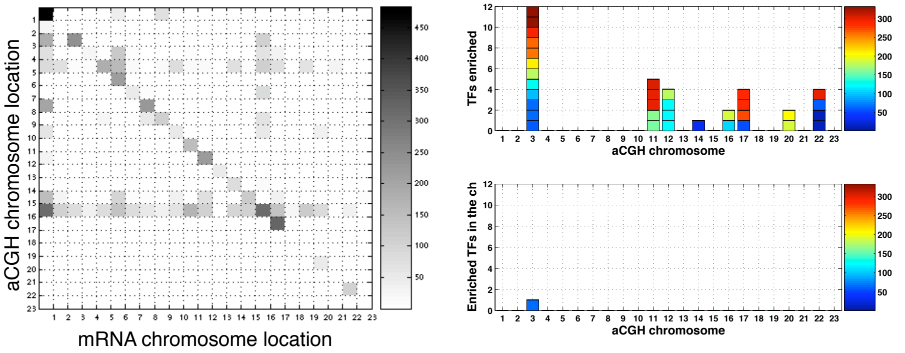
c) LUMINAL A d) LUMINAL B

Figure 7.9: Results of the scale search applied to more homogeneous patient group, i.e. the patients belonging to the same subtype.

Chromosomes 4*p*, 5*q*, 15*p* and most strongly 16*q*. Those locations have been observed already in Figure 7.4 (a) as expressing an enrichment along many chromosomes in the genome. Although changing the locations of the genes, the randomization does not alter the fact that the *relevant* pairs are scattered along the whole genome. Therefore, the rows involving those copy number measurements are still significantly enriched, as expected. When we randomize the structure of the location of the copy number measurements, instead, nothing is enriched anymore. The maximum number of scales reached is 9, and it is obtained in just a few pairs. These results suggest that the correlation patterns observed are not an artifact of the method, but are, indeed, present in the data.

Application of the IGDam algorithm to a homogeneous patients group

In our analysis we have used all samples available, in an unsupervised manner. We hypothesize that the correlation patterns observed are due to the heterogeneity of the patients. Those patterns would disappear if a more homogeneous patient group would be used. In order to verify this hypothesis we apply our algorithm to the patients belonging to the subtypes first proposed by Sorlie *et al.* [Sorl 01, Pero 00, Sorl 03]. We have exten-



a) b)

Figure 7.10: Results of the TFs enrichment analysis for the case summarized in Figure 7.4 (a). The right plot (a) indicate the number of target genes according to their chromosomal location (horizontal axis). The target genes are the genes present in the selected pairs whose aCGH probes are located on the same chromosome. The color codes for the number of genes, ranging from 0 (white) to 483 (black), which is the number of genes on Chromosome 1 present in the relevant pairs whose aCGH is located on Chromosome 1. The left plot (b) is comprised by two panels. In the upper panel are indicated the TFs enriched in the target gene-lists, e.g. the genes present in each row of the matrix in part (a). The lower panel select which TFs of t he upper panel is also located in the same chromosome as the aCGH probes of the relevant pairs originating the list. The color here codes the index that identify each individual TF.

sively analyzed the stratification of the 68 *NKI* patients according to the subtype labels in [Lai 07]. There, we have identified the genomic aberration specific for each class. Now we consider the patients belonging to a single class. First we compute the correlation between copy number and expression, and second we apply the scale search. The results for the Basal, ERBB2, Luminal A and B subtypes are presented in Figure 7.9. No strong correlation patterns can be detected across all window scales. The ERBB2 patients do no have a single pair contained in an enriched window in 12 different scales. The other classes have few scattered pairs, the larger contribution is found in the Luminal A, on Chromosome 1q. Overall, we can confirm that no patterns of correlation can be observed when the sample groups are rather homogeneous, as in case of the four subtypes analyzed.

Transcription factor analysis

We have observed several spatial correlation structures in the combined aCGH and mRNA datasets, which we have referred to as local and genome-wide effects. In order to investigate the possible causes for these dependencies we analyzed whether the genes involved in the mentioned effects are co-regulated by the same transcription factors. We used the TRANSFAC database [Data] to obtain a list of transcription factors (TFs) with their binding sites. The TFs with known genomic location are selected from the database, resulting in a total of 332 TFs. For each gene in the NKI dataset, we determined which TFs can putatively bind their upstream region, allowing us to establish a link between a transcription factor and his putative targets. Let's consider the rows in the matrix depicted in Figure 7.4 (a), and select the genes associated with the (aCGH-mRNA) pairs that were significantly enriched in e.g. 12 window scales. Instead of considering the genes associated with a single DNA-probe, we chose to do the analysis per chromosome, looking at the genes significantly correlated with the entire chromosome. This is because often the correlation effects involve an entire part of a chromosome in a similar manner. Therefore, treating the copy number measurements individually would be redundant, besides being more difficult to interpret. First, the pairs whose aCGH probes are located on the same chromosome and that are significantly enriched in at least 12 window scales are selected. Second, the genes in those pairs are considered. Figure 7.10 (a) depicts the number of those target genes. For example, the first row of the matrix indicates the target genes located in the pairs correlated with Chromosome 1, which are situated on Chromosome 1 itself, and on Chromosomes 6 and 9. As expected, the largest number of genes is found in the row associated with Chromosome 16, which contains genes located on almost all chromosomes. Only Chromosomes 18, 19, 21 and 23 did not have any region with significantly enriched pairs. While some chromosomes shown only pairs with the aCGH and mRNA located in the same chromosome (CIS effect), e.g. see Chromosome 17. The color codes for the number of pairs found per chromosome, depicted according to the location of the mRNA component of the pairs. One could hypothesize that the DNA probes are correlated with the expression since a TF is located in the region where the copy number is altered and that this TF has a set of targets located in the region where the expression is correlated with the DNA copy number. To test this hypothesis we employed the hypergeometric distribution to determine if a particular TF is enriched in the given list of target genes, i.e. whether a TF preferentially binds to these genes. Figure 7.10 (b) (upper panel), depicts the TFs of whom the binding sites are significantly enriched (p-value smaller than 0.005) amongst the target genes associated with the copy number measurements in each chromosome. The color is not important here, since it codes for the index of the TF. Figure 7.10 (b) (lower panel) depicts the TFs that are located on the same chromosome as the copy number measurements of the pairs comprising the target genes. One TF is on the same chromosome as the DNA probes used to generate the target gene-list, i.e. on Chromosome 3. The target gene-list for the pairs whose aCGH is located on Chromosome 3 comprises genes on Chromosomes $1, 2, 3$ and 16, as illustrated by the 3rd row of the matrix in Figure 7.10 (a). Several target gene-lists present an enrichment in some TFs, although the results are depending on the specific genesets. We have repeated the analysis using a less stringent criterion for the selection of the correlated pairs: the required number of scales where pairs should be enriched equaling to only $10, 8$ and 6. The gene-lists become larger, and some TFs are identified as enriched in the target set. However none are located on the same chromosome as the DNA probe correlated with the mRNA expression of the target list. The TFs present change as well, i.e. for a given chromosome, the gene lists obtained with the different values for the scale search enrichment do not always identify the same TFs. Therefore, this analysis does not help us in finding causal relationships which could explain the observed correlation. We do not find evidence that the genome-wide effects are due to the presence of a TF, which, altered by a chromosomal aberration, causes a genome-wide change in expression.

7.4 Conclusions

We have investigated the genome-wide correlation dependencies between copy number and expression datasets. Our goal has been the identification of regions of interdependencies between these two data type. The results have reconfirmed the cis-effect, i.e. the high correlation between the copy number changes and the expression of the genes in the same region. Furthermore, concerning the data type independently, we have highlighted that in the aCGH data only the cis-effects are very strong, involving often the whole chromosome arm, but are less pronounced in the expression data.

Two types of dependencies between copy number and expression datasets have been identified: a local and a genome-wide dependency. We have verified that these findings are not artifacts, since they do disappear when randomized data is used instead. Nor they are caused by one dataset only, since the number of *relevant* pairs does not drop when only the pairs with high STD in both datasets are allowed to be selected, as illustrated in Figure 7.7. These patterns of correlation are influenced by the sample heterogeneity. As shown in Figure 7.9, they are not strongly present when homogeneous sample groups are analyzed.

It should be noted that *correlation does not imply causality*, which is the ultimate goal of our search. In order to better understand the mechanism of cancer, our final aim is, in fact, the identification of the causal relationship between copy number and expression data. In order to further investigate this aspect, we have analyzed whether the correlation patterns are caused by the presence of a transcription factor in the altered copy number measurements, which is significantly enriched in the genes showing correlation with those copy number measurements. We can pinpoint at transcription factors enriched in the gene list generated by the IGDam algorithm. Unfortunately, we cannot propose any TF as a cause for the local or genome-wide effects that we have found, since no TF was both produced by a gene located in the genomic region identified by the copy number measurements, and significantly enriched in the genes correlated with those copy number measurements.

In conclusion, our method makes possible for the fist time to explore genome-wide dependencies between copy number and expression dataset. Several mechanism of interest have been hypothesized. Further work is needed for a validation of the findings and a determination of the causality direction of the identified trans-effect dependencies.

8

Conclusions

8.1 Concluding remarks

Part I: dependencies in gene expression datasets.

The first part of the thesis focused on dependencies within the expression datasets, and addressed the issues of gene selection and predictor building. By the time gene expression data became available a vast body of knowledge on classification methods had already been developed in the fields of machine learning and pattern recognition. The methods developed in those fields were immediately applied, often without the necessary precaution, to this new data type. Therefore, many errors appeared even in studies published in high impact factors journals [Dupu 07,Brag 07]. We have identified these problems at the early stage of our research, and dedicated considerable effort towards performing extensive and systematic comparisons, as presented in Chapter 4. Our major conclusions are outlined in the following paragraphs. Recently several reviews have aimed at summarizing the major limitations and highlight the reached consensus [Alli 06,Dupu 07,Brag 07]. Dupuy and Simon [Dupu 07] provide a useful checklist of "Do's and Don'ts" concerning the statistical analysis of gene expression data. Allison *et al.* [Alli 06] present a constructive view of the consensus reached in what they define as the five components of gene expression analysis (design, preprocessing, inference, classification and validation). They point out that in some areas the need for evaluating existing techniques is more important than the development of new ones. Our findings are in line with this statement.

Simple methods work best.

Our aim has been to compare and clarify the benefits of the different gene selection and classification techniques. Our results suggest that simple methods, e.g. univariate gene selection coupled with the Nearest Mean Classifier, when not outperforming, are as good as more complex ones. In particular, the multivariate selection techniques, which were expected to detect high level dependencies between the genes, did not improve the classification of the samples. Therefore, these procedures did not succeed in detecting higher order dependencies. We want to emphasize that the limited sample size is currently the major constraint for the complexity of the gene selection and classification algorithms. The comparison between multivariate and univariate selection techniques should be performed on larger cohorts that are now starting to be available.

The genes carry redundant information.

We have attempted to identify informative gene sets. Our initial hypothesis has been that only a small number of genes would carry the information of interest, e.g. be predictive of cancer aggressiveness. This was a shared assumption, encouraged also by the small size of the different signatures proposed by many studies, which were often smaller than a hundred genes [Veer 02, Wang 05b, Ma 04, Gema 04]. This hypothesis has been an initial motivation for the development of the Random Subspace Method (RSM) described in Chapter 3. However, the RSM failed to improve the classification performance when applied to gene expression datasets. Since the strength of the algorithm is in the identification of multivariate information present in data with large number of uninformative features, our results suggest that this assumption does not hold for expression datasets. The number of informative genes is not restricted to tens of genes, but it may extend to thousands of genes. This conclusion is also supported by the experiments in Chapter 2. Our results show that the larger the signatures the higher the classification performance, suggesting that increasing information was, indeed, available in the genesets.

The fact that signatures describing the same processes are not overlapping and are sometimes not validated in independent cohorts has casted doubt on their reliability and robustness [Ein 06, Reid 05]. For example, Ein-Dor *et al.* [Ein 05] have shown that multiple signatures achieve the same classification performances. However, the comparison of related signature by Fan *et al.* [Fan 06] have suggested that the different signatures track a common set of biologic characteristics, and therefore, they all have validity.

Part II: dependencies between copy number and expression datasets

In the second part of this thesis we investigated the spatial dependencies within copy number datasets, and between these and the corresponding expression datasets. While there were many available strategies for the construction of classifiers (Part I), the methodologies and tools for the data integration of copy number and expression data had to be developed from scratch.

Copy number data shows spatial local dependencies

Several algorithms specifically developed for copy number data, reveal that genomic aberrations involve regions of the genome that are spatially related [Lai 05, Pica 05, Jong 03, Jong 04, Wang 05a]. Our emphasis has not been on the identification of chromosomal gain or loss on a per sample basis, but we have strived to evaluate what makes two classes different from each other, and what are the aberrations that distinguish them. This is advantageous especially when working with human tumor samples, whose heterogeneity is much larger then the one observed e.g. in mouse datasets. We assumed that the heterogeneity of tumors may lead to signals for the aberrations smaller than the ones expected if the tumors were homogeneous. Therefore, amplifications/deletions with small absolute values may be of interest as well, especially when they discriminate the classes of interest. Several authors (e.g. Saramaki *et al.* [Sara 06], Fridlyand and Chin *et al.* [Frid 06, Chin 06], and Nymark *et al.* [Nyma 06]) have recently pointed out that even low-level copy number aberrations may have significant effects on gene expression and, therefore, on cellular functions and tumor development. These findings support the assumption made in the SIRAC algorithm.

Combining copy number and expression data allows a feasible search of marker cancer genes

The expression data was integrated with the results from the copy number data in a post-processing step. Prioritizing the gene expression according to the correlation with the corresponding copy number data is especially relevant since we expected that, for instance, not all genes in a region of aberration would be active. Some may be silent and not contributing to the mechanism of cancer. In Chapter 6, a selection based on these additional information sources, resulted in smaller lists of potentially interesting genes that were analyzed further.

For classification purposes the expression is to be preferred to the copy number data

The copy number data provides a powerful tool to investigate genomic aberrations. However, expression data turns out to bear much larger variability, since not only genetic but also epigenetic events are responsible for the changes in the expression of the genes. In a preliminary, unpublished study, we have used both copy number and expression of 68 samples for classification purposes. We have observed that prediction based upon expression data provides a lower classification error than prediction based upon copy number data. Although these results are based on a limited sample set, they are an indication that gene expression should be preferred to copy number data for the sake of classification.

Trans-effect dependencies have been observed between expression and copy number data

The study of the dependencies via the combination of copy number and expression data is a challenging task. Distinguishing the epigenetic and genetic dependencies would bring more understanding about the cancer development. In Chapter 7 we have made a first attempt to evaluate, in a statistical way, the genome-wide dependencies between copy number and expression data. The results have confirmed the high correlation between copy number changes and expression of the genes located in the same genomic region. We have highlighted that only the local spatial dependencies are very strong in the copy number data, involving often the whole chromosome arm, but are less pronounced in the expression data. Furthermore, two types of trans-effect dependencies between copy number and expression datasets have been identified: a local and a genome-wide dependency. A local dependency involves two different chromosomal regions in copy number and expression data. A genome-wide effects occurs when a region on the copy number data shows genome-wide correlation with the expression data. However, further validation is needed in order to confirm and understand the causality of the observed effects.

8.2 Open issues

Cancer research using high throughput data is a fast changing and developing field. The advances of the technologies are opening new possibilities and providing new information. Therefore, there are many open issues.

Validation

A convincing validation is a must [Thie 06, Dupu 07, Alli 06], but has been limited by several restraints. A major one is the absence of a ground truth. How can one determine, for example, if the genes identified in a given analysis are related to cancer if their function is still unknown? Another major issue is the heterogeneity in the data collection and processing. The samples may come from a non homogeneous clinical group. Many different platforms and array types are available on the market, and even within the same platform, every lab has its own protocol to perform the analysis. Even when adopting the class labels, it is important to keep in mind that errors can occur in the label process due to several reasons. Errors can occur first, due to inaccuracy in the manual annotation, second, due to the subjectivity of the human evaluation, third, due to inconsistency of the label when two sources are available. This situation happens, e.g. when the label based on immunoistochemisty staining does not concord with the one obtained from the expression value of the gene coding for the same protein. The label uncertainty has motivated several authors to address this issues also on a statistical level, incorporating the label uncertainty in the model used for classification [Zhan 06b, Li 07]. Furthermore, the small sample availability imposes strong assumptions, such as gene independence, which is implied by the univariate gene selection. Therefore, larger cohorts and more extensive experimental validation are needed to reach more generally applicable and reproducible results.

Data integration

Data integration is becoming more and more important, and is attracting increasing efforts of the scientific community [Alli 06, Buss 07]. Initially, the integration of different expression datasets has been investigated, as a possibility to obtain larger cohorts. For example, Segal *et al.* [Sega 04] presented an integrated analysis of published gene expression datasets across 22 tumor types, Hwang *et al.* [Hwan 05] proposed a method to combine different datasets to obtain a unique network model, and applied it to 18 yeast datasets, Teschendorff *et al.* [Tesc 06] built a prognostic gene expression classifier using three independent cohorts of ER positive breast cancer.

The interest for integration is becoming even more relevant with the availability of various data types, which provide different views of the same biological processes [Edgr 06, Alli 06, Buss 07]. Besides expression measurements, other examples are DNA copy number variations [Pink 05], single nucleotide polymorphisms (SNPs) [Shas 03], transcription factors binding sites [Jeff 07], protein levels [Buck 04], and the diverse databases which contain information on gene annotation. Many tools that aim at facilitating an integrated analysis and the interpretation of the results have appeared [Anal, Al S 07, Chan 06, Nam 06]. Recent studies have included Transcription Factors activities [Buss 07, Jeff 07], SNPs [Stra 07], specific sequence of DNA called motifs [Noto 06, Eden 07], and metabolome data [Caki 06]. The aim is to obtain a more complete picture of the biological processes via the combination of different aspects described in the diverse data types. Similarly, our work in Chapters 6 and 7 has investigated the dependencies between copy number and expression datasets, suggesting that data integration allows the identification of potential marker genes, and the generation of novel hypothesis regarding cancer mechanisms. However, it is too early to have a consensus on the methodology used and on the findings obtained.

Afterword

Current cancer research is mainly focused on the molecular mechanisms within cells, reducing the human being to his/her DNA material. However, the human being is not only composed of his/her physical body, but also of his/her emotional and thinking processes. It is becoming a shared consensus that stress, unhealthy habits and environmental influences have a great impact in the development of diseases. In recent years, several medical doctors have promoted a more radical view of the illnesses, cancer included, being a result of emotional stress and unhealthy thinking patterns, which are often not only unsolved, but not even expressed at a conscious level [Deth 02, Bens 96, Chop 91, Weil 04, Hame 02]. The philosophy behind these approaches is that the human being is considered as a whole, not only as the sum of his organs and tissues. The patient is empowered, since the healing is believed to originate within the patient rather than from the physician. The change in perspective and the implications that these approaches require are very radical. Therefore, the integration of these ideas with conventional medicine is just in its infancy and presents many challenges [Bell 02, Rees 01]. However, the awareness towards these approaches is increasing. This is testified, for example, by the growing adoption, besides the conventional therapies, of what is today referred to as complementary and alternative medicine (CAM), which include e.g. traditional Chinese medicine, Ayurveda and homeopathy. From a research view point, several cancer institutes have presented, for example, studies on the effect of diet, stress reduction, meditation and yoga on the development of the diseases or on the symptoms of stress, such as sleep disorders [Cohe 04, Saxe 01, Spec 00, Zama 96]. However, the effort and investments are heavily unbalanced towards conventional medicine. I'm convinced that in order to understand cancer a more holistic approach should be considered, and that cancer research should take on the challenge to incorporate the more subtle components of the human being, such as mind and emotions.

Delft, February 2008 *Carmen Lai*

A

Genome and gene array description

The genome and its annotation

Every eukaryotic cell is defined by a membrane that holds the cytoplasm and the nucleus, which contains the DNA. In humans, the DNA material is organized in two copies of twenty-three chromosomes (from 1 to 22 plus the chromosome Y or X, which characterize the sex of a person). Some parts of the DNA are *coding* regions, these are referred to as genes. These parts may produce RNA copies and release these transcripts in the nucleus. The transcripts are used as templates for the translation into proteins.

In 1982 the sequencing of the human genome started. This huge project is a shared effort of the American GenBank, the European EMBL (European Molecular Biology Laboratory) and the Japanese DDBJ (DNA Data Bank of Japan). The human genome has been nearly fully sequenced, and this information is stored in shared primary databases of the mentioned three initiatives, and is accessible via the world wide web. However, this work has not been completed, the sequence information is constantly improving, and the databases which store it are regularly updated. These primary databases do not contain meta-information, e.g. to which gene the annotated sequence belongs. Therefore, this need is addressed by a second layer of databases (Unigene, TIGR GI, RefSeq). These databases aim to cluster together the sequences into unique genes. In particular, the RefSeq database contains high-quality and well-annotated information [Stek 03]. A third level is provided by the Ensembl database, which builds on both the low level information stored in the primary databases, and the higher level annotations of the sequences contained in the second level databases. The Ensembl database aims at organizing all the sequence information into chromosome sequences in such a way that each chromosomes appears as a long single sequence, the so-called *Golden path*. Therefore, it facilitates efficient querying and mining of the complete genome.

The knowledge about the genome sequence is used to create the arrays, which measure e.g. the quantity of mRNA expressed, or the amount of DNA at a specific genomic loca-

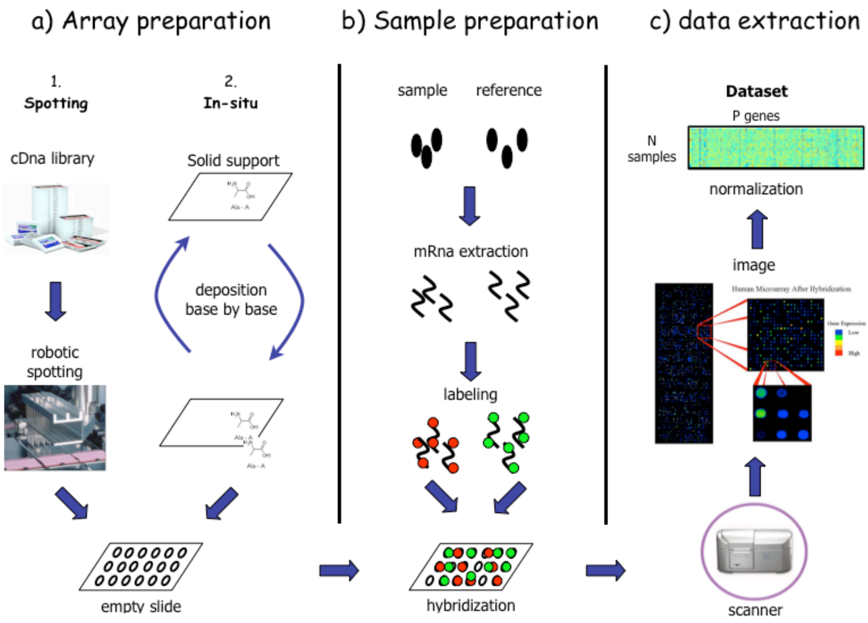


Figure A.1: Illustration of the process that leads to a high throughput dataset. The two-color technique, used e.g. in the Agilent platform, is depicted.

tion in the cells of interest. This data can be employed to study the genetic mechanism of the cell and to learn about the alterations, which lead to and govern the cancer.

Gene array description

The first paper to introduce micro-array technology was published in 1995 by Schena *et al.* [Sche 95]. Since then, many techniques have become available and have improved the reliability and quality of the micro-arrays. Major advances are pushing the high throughput technology towards higher resolution and more reliable data generation. Here, we will briefly summarize the process of fabricating and using the micro-arrays employed to quantify the expression or copy number of thousands of genes simultaneously. For a more detailed description see [Cheu 99, Stek 03, Dugg 99].

The array fabrication process is illustrated in Figure A.1. The first stage consists of the creation of the microscope slide onto which the DNA molecules have been chemically bound. To this end, the main techniques are robotic spotting and in-situ synthesis.

The robotic spotting consists of three steps (Figure A.1 (a)). A library of DNA clones is prepared. The probes can be either polymerase chain reaction (PCR) products or oligonucleotides. The spotting is done by a robot, using a set of pins. The pins are first dipped in the wells containing the DNA, so that the clones are collected. Then the pins are moved onto the glass, and the DNA is spotted. Finally, the pins are washed to remove all residual material. The process is repeated on the successive set of wells till the array is complete. The array is then *fixed* so that no other DNA material can attach to it.

The in-situ synthesis is schematized in Figure A.1 (a). Instead of presynthesising the oligonucleotides, the procedure consists of growing the oligos, base after base, on the surface of the array. There are three main technologies which differ in the way the oligos are grown. The Affymetrix technology uses photolithography to glue the desired bases together. The light passing through particular masks directs the reaction only to desired spots in the array. Other technologies, such as Nimblegen, use micromirror arrays instead of masks to direct the light and consequently the position where a base is attached. The third technique uses chemical reactions with synthesis via ink-jet technology. This is the strategy adopted by Agilent. The advantage of the micro-mirroring and ink-jet techniques is the high versatility. The operator can fully control the process via the computer, choosing to make the array with any oligo that he/she wishes. Although limited in flexibility by the fixed masks, the Affymetrix technique is very efficient in large scale production of identical arrays. Concerning the synthesis effectiveness, the techniques which make use of light have an efficiency of 95%, while the chemical ink-jet synthesis has a 98% efficiency. The diffraction of light through the masks limits the quality of the spots of the Affymetrix arrays. This problem is addressed in their specific image-processing software. The ink-jet approach gives the best spot quality [Stek 03].

Both the spotted and the in-situ synthesis technologies provide as an outcome an empty array with the DNA probes attached to it. The goal is now to obtain and prepare the DNA or mRNA material to be measured. Figure A.1 (b) schematically depicts the necessary steps. The DNA or mRNA material has to be extracted from the samples, purified and amplified. When the interest is in the expression of the genes, the messenger RNA (mRNA molecules) are collected and reverse transcribed into the corrispondent DNA nucleotides. If our focus is on the genomic aberrations, we want to measure the copy number of the genes. Therefore, the DNA from the chromosomes is collected. The extraction is a complex process, that is performed in various ways in different laboratories. When the interest is in the amount of DNA material present in the cell, as is the case in the Affymetrix arrays, we measure each sample separately. A reference sample is used when we want to compare the sample with it, e.g. to measure the relative difference of the sample versus a normal tissue, or versus a pool of tumor samples. The DNA material, from the sample and the reference, are then labeled with different fluorescent dyes. Two dyes are mainly used: Cy3 (that emits green light when excited with the corresponding laser wavelength) and Cy5 (red light emission). The hybridization is the step in which the labeled DNA binds the specific complementary DNA probe on the array. This process generally takes place over a period of 12 to 24 hours. The slides are then washed. In this way the excess DNA is removed from the slide. Another result of the washing is to reduce the cross-hybridizations, i.e. the bonds between not complementary pieces of DNA, which are generally weaker then the matching ones.

The final step in Figure A.1 (c) consists of obtaining the measurements from the hybridized array. The slide is put into a scanner, where a laser with the proper wavelength (green or red light) focuses on a point in the array, and excites the dyes present at that spot. After excitation, the red (or green) fluorofors emit the light, which is collected by a photomultiplier tube in the scanner. The quantity of the signal collected is proportional to the DNA material present in the sample. The scanner focuses on every point of the array, by moving either the laser or the slide, so that an intensity image of the array is obtained. If the interest is in an absolute measurement then a single channel is collected. For a two-color measurement the green and red channel are acquired separately and then combined to produce a ratio of intensities. The image is finally stored as a tagged image file format (TIFF). The array image provide us with the per pixel intensity of the dye emissions. The image needs to be processed in order to obtain the quantitative measure of each probe. The information of different samples is combined together in a dataset to be used for the statistical analysis. Since there are many sources of systematic variation in microarray experiments, several normalization procedures are needed to correct for it [Yang 02, Stek 03, Spee 03]. A necessary normalization is done *within* a slide to correct for spatial and intensity dependent dye biases. Another normalization can be applied if a dye-swap experiment is performed. Two arrays are hybridized using two different dyes for the same sample. Since the efficiency of the dyes is different, the normalization conducted on the results of the combined arrays will correct for this effect. However, a dye-swap experiment requires double the number of arrays. A third type of normalization maybe performed *between* slides. The aim is to adjust for sample variances in the intensities across slides. However, the risk is to artificially increase variability in the data. Therefore, if the difference between samples are fairly small, it is advisable to perform only a *within* sample normalization [Yang 02].

After the normalization step, the data from all samples is organized in a matrix $N \times P$, with N samples and P probes. This is the starting data that was employed in the research presented in this thesis.

References

[Abul 05] O. Abul, R. Alhajj, F. Polat, and K. Barker. “Finding differentially expressed genes for pattern generation”. *Bioinformatics*, Vol. 21, No. 4, pp. 445–450, 2005.

[Adel 07] J. Adelaide, P. Finetti, I. Bekhouche, L. Repellini, J. Geneix, and F. e. a. Sircoulomb. “Integrated profiling of basal and luminal breast cancers”. *Canceer Research*, Vol. 67, No. 24, pp. 11565–75., 2007.

[Adle 06] A. Adler, M. Lin, H. Horlings, D. Nuyten, M. van de Vijver, and H. Chang. “Genetic regulators of large-scale transcriptional signatures in cancer.”. *Nature Genetics*, Vol. 38, No. 4, 2006.

[Al K 04] K. Al Kuraya, P. Schraml, J. Torhorst, C. Tapia, B. Zaharieva, and H. e. a. Novotny. “Prognostic relevance of gene amplifications and coamplifications in breast cancer”. *Canceer Research*, Vol. 64, No. 23, pp. 8534–40, 2004.

[Al S 07] F. Al-Shahrour, L. Arbiza, H. Dopazo, J. Huerta-Cepas, P. Minguez, D. Montaner, and J. Dopazo. “From genes to functional classes in the study of biological systems”. *BMC Bioinformatics*, Vol. 8, No. 114, 2007.

[Albe 02] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular biology of the cell.* Garland Science, 2002.

[Aliz 00] A. Alizadeh, M. Eisen, R. Davis, and M. e. a. Chi. “Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling”. *Nature*, Vol. 403, pp. 503–511, 2000.

[Alli 06] D. Allison, X. Cui, G. Page, and M. Sabripout. “Microarray data analysis: from disarray to consolidation and consensus”. *Nature Reviews*, Vol. 7, pp. 55–65, 2006.

[Allw 00] E. Allwein, R. Schapire, and Y. Singer. “Reducing multiclass to binary: A unifying approach for margin classifiers.”. In: M. Kaufmann, Ed., *In Proc. 17th International Conf. on Machine Learning*, pp. 9–16, 2000.

[Alon 99] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine. “Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays”. *PNAS*, Vol. 96, No. 12, pp. 6745–6750, 1999.

[Ambr 02] C. Ambroise and G. McLachlan. “Selection bias in gene extraction on the basis of microarray gene-expression data”. *PNAS*, Vol. 99, No. 10, pp. 6562–6566, 2002.

[Anal] I. P. Analysis. http://www.ingenuity.com/.

[Ayer 04] M. Ayers, W. Symmans, J. Stec, A. Damokosh, K. Clark, E.and Hess, M. Lecocke, J. Metivier, D. Booser, N. Ibrahim, V. Valero, M. Royce, B. Arun, G. Whitman, J. Ross, N. Sneige, G. Hortobagyi, and L. Pusztai. "Gene Expression Profiles Predict Complete Pathologic Response to Neoadjuvant Paclitaxel and Fluorouracil, Doxorubicin, and Cyclophosphamide Chemotherapy in Breast Cancer". *Journal of Clinical Oncology*, Vol. 22, No. 12, pp. 2284–2293, 2004.

[Beer 05] E. van Beers, T. van Welsem, L. Wessels, Y. Li, R. Oldenburg, P. Devilee, C. Cornelisse, S. Verhoef, F. Hogervorst, and P. van't Veer, L.J.and Nederlof. "Comparative Genomic Hybridization Profiles in Human BRCA1 and BRCA2 Breast Tumors Highlight Differential Sets of Genomic Aberrations". *Cancer Research*, Vol. 65, No. 3, pp. 822–827, 2005.

[Bell 02] I. Bell, O. Caspi, G. Schwartz, K. Grant, T. Gaudet, D. Rychener, V. Maizes, and A. Weil. "Integrative Medicine and Systemic Outcomes Research". *Archives of Internal Medicine*, Vol. 162, pp. 133–140, 2002.

[Ben 00] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. "Tissue classification with gene expression profiles.". In: *Proceedings of the fourth annual international conference on Computational molecular biology*, pp. 54–64, ACM Press, Tokyo, Japan, 2000.

[Ben 99] A. Ben-Dor, R. Shamir, and R. Yakhini. "Clustering Gene Expression Patterns". *Journal of Computational Biology*, Vol. 6, pp. 281–297, 1999.

[Bens 96] H. Benson. *Timeless Healing.* Wheeler Pub Inc, 1996. ISBN 9781568953663.

[Berg 06] A. Bergamaschi, Y. Kim, P. Wang, T. Sorlie, T. Hernandez-Boussard, P. Lonning, R. Tibshirani, A. Borresen-Dale, and J. Pollack. "Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer". *Genes, Chromosomes and Cancer*, Vol. 45, No. 11, pp. 1033–1040, 2006.

[Bert 03] F. Bertucci, P. Viens, P. Hingamp, V. Nasser, R. Houlgatte, and D. Birnbaum. "Breast Cancer Revisited Using Dna Array-Based Gene Expression Profiling". *International Journal of Cancer*, Vol. 103, pp. 565–571, 2003.

[Bhat 03] C. Bhattacharyya, L. R. Grate, A. Rizki, D. Radisky, F. J. Molina, M. I. Jordan, M. J. Bissell, and I. S. Mian. "Simultaneous classification and relevant feature identification in high-dimensional spaces: application to molecular profiling data". *Signal Processing*, Vol. 83, No. 4, pp. 729–743, 2003.

[Bicc 03] S. Bicciato, A. Luchini, and C. Di Bello. "PCA disjoint models for multiclass cancer analysis using gene expression data.". *Bioinformatics*, Vol. 19, No. 5, pp. 571–578, 2003.

[Blam 07] R. Blamey, I. Ellis, S. Pinder, A. Lee, R. Macmillan, and D. e. a. Morgan. "Survival of invasive breast cancer according to the Nottingham Prognostic Index in cases diagnosed in 1990-1999.". *European Journal of Cancer*, Vol. 43, No. 10, pp. 1548–1555, 2007.

[Blan 04] R. Blanco, P. Larranaga, I. Inza, and B. Sierra. "Gene selection for cancer classification using wrapper approaches.". *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 18, No. 8, pp. 1373–1390, 2004.

[Blen 07] C. Blenkiron, L. Goldstein, N. Thorne, I. Spiteri, S. Chin, and M. e. a. Dunning. “MicroRNA expression profiling of human breast cancer identifies new markers of tumour subtype”. *Genome Biology*, Vol. 8, No. 10, p. R214, 2007.

[Bo 02] T. Bo and I. Jonassen. “New feature subset selection procedures for classification of expression profiles”. *Genome biology*, Vol. 3, 2002.

[Brag 07] U. Braga-Neto. “Fads and fallacies in the name of small-sample microarray classification”. *IEEE Signal Processing Magazine*, Vol. 24, No. 1, pp. 91–99, 2007.

[Buck 04] M. Buck and J. Lieb. “ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments”. *Genomics*, Vol. 83, No. 3, pp. 349–360, 2004.

[Buss 07] H. Bussemaker, B. Foat, and L. Ward. “Predictive Modeling of Genome-Wide mRNA Expression: From Modules to Molecules”. *Annual Review of Biophysics and Biomolecular Structure*, Vol. 36, pp. 329–347, 2007.

[Caki 06] T. Cakir, K. Patil, Z. Onsan, K. Ulgen, B. Kirdar, and J. Nielsen. “Integration of metabolome data with metabolic networks reveals reporter reactions”. *Molecular Systems Biology*, Vol. 2, No. 50, 2006.

[Call 05] G. Callagy, P. Pharoah, S. Chin, T. Sangan, Y. Daigo, L. Jackson, and C. Caldas. “Identification and validation of prognostic markers in breast cancer with the complementary use of array-CGH and tissue microarrays”. *Journal of Pathology*, Vol. 205, No. 3, pp. 388–396, 2005.

[Call 06] A. Callegaro, D. Basso, and S. Bicciato. “A locally adaptive statistical procedure (LAP) to identify differentially expressed chromosomal regions”. *Bioinformatics*, Vol. 22, No. 21, pp. 2658–2666, 2006.

[Capo 06] E. Capobianco. “Statistical Embedding in Complex Biosystems”. *Journal of Integrative Bioinformatics*, Vol. 3, No. 2, 2006.

[Cart 06] S. Carter, A. Eklund, I. Kohane, L. Harris, and Z. Szallasi. “A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers.”. *Nature Genetics*, Vol. 38, No. 9, pp. 1043–1048, 2006.

[Chan 05a] H. Chang, D. Nuyten, J. Sneddon, T. Hastie, R. Tibshirani, T. Sorlie, H. Dai, Y. He, L. van’t Veer, H. Bartelink, M. van de Rijn, P. Brown, and M. van de Vijver. “Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival”. *PNAS*, Vol. 102, No. 10, pp. 3738–3743, 2005.

[Chan 05b] J. Chang, E. Wooten, A. Tsimelzon, S. Hilsenbeck, M. Gutierrez, Y. Tham, M. Kalidas, R. Elledge, S. Mohsin, C. Osborne, G. Chamness, D. Allred, M. Lewis, H. Wong, and P. O’Connell. “Patterns of Resistance and Incomplete Response to Docetaxel by Gene Expression Profiling in Breast Cancer Patients”. *International Journal of Oncology*, Vol. 23, No. 6, pp. 1169–1177, 2005.

[Chan 06] J. Chang and J. Nevins. “Gather: A System approach to Interpreting Genomic Signatures”. *Bioinformatics*, Vol. 22, No. 23, pp. 2926–2933, 2006.

[Char 05] E. Charafe-Jauffret, C. Ginestier, F. Monville, S. Fekairi, J. Jacquemier, D. Birnbaum, and F. Bertucci. “How to best classify breast cancer: conventional and novel classifications (review).”. *International Journal of Oncology*, Vol. 27, No. 5, pp. 1307–1313, 2005.

[Cheu 99] V. Cheung, M. Morley, F. Aguilar, A. Massimi, R. Kucherlapati, and G. Childs. “Making and reading microarrays”. *Nature genetics*, Vol. 21, pp. 15–19, 1999.

[Chil 02] A. Chilingaryan, N. Gevorgyan, A. Vardanyan, D. Jones, and A. Szabo. “A multivariate approach for selecting sets of differentially expressed genes”. *Mathematical Biosciences*, Vol. 176, pp. 59–69, 2002.

[Chin 06] K. Chin, S. DeVries, J. Fridlyand, P. Spellman, R. Roydasgupta, W. Kuo, A. Lapuk, R. Neve, Z. Qian, T. Ryder, F. Chen, H. Feiler, T. Tokuyasu, C. Kingsley, S. Dairkee, Z. Meng, K. Chew, D. Pinkel, A. Jain, B. Ljung, L. Esserman, D. Albertson, F. Waldman, and J. Gray. “Genomic and transcriptional aberrations linked to breast cancer pathophysiologies”. *Cancer Cell*, Vol. 10, pp. 529–541, 2006.

[Chin 07a] S. Chin, A. Teschendorff, J. Marioni, Y. Wang, N. Barbosa-Morais, and N. e. a. Thorne. “High-resolution array-CGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer”. *Genome Biology*, Vol. 8, No. 10, p. R215, 2007.

[Chin 07b] S. Chin, Y. Wang, N. Thorne, A. Teschendorff, S. Pinder, and M. e. a. Vias. “Using array-comparative genomic hybridization to define molecular portraits of primary breast cancers”. *Oncogene*, Vol. 26, No. 13, pp. 1959–70, 2007.

[Cho 03] S. Cho and H. Won. “Machine learning in DNA microarray analysis for cancer classification”. In: *Proceedings of the First Asia-Pacific bioinformatics conference*, 2003.

[Chop 91] D. Chopra. *Creating health.* Houghton Mifflin Books, 1991. ISBN 0395755158.

[Chow 01] M. Chow and I. Moler, E.J.and Mian. “Identifying marker genes in transcription profiling data using a mixture of feature relevance experts.”. *Physiological Genomics*, Vol. 5, pp. 99–111, 2001.

[Cohe 04] L. Cohen, C. Warneke, R. Fouladi, M. Rodriguez, and A. Chaoul-Reich. “Psychological adjustment and sleep quality in a randomized trial of the effects of a Tibetan yoga intervention in patients with lymphoma”. *Cancer*, Vol. 100, No. 10, pp. 2253–2260, 2004.

[Dai 03] D. Dai and P. Yuen. “Regularized Disciminant Analysis and Its Application to Face Recognition”. *Pattern Recognition*, Vol. 36, pp. 845–847, 2003.

[Data] T. Database. http://www.gene-regulation.com/.

[Dell 02] A. Dellas, J. Torhorst, E. Schultheiss, M. Mihatsch, and H. Moch. “DNA sequence losses on chromosomes 11p and 18q are associated with clinical outcome in lymph node-negative ductal breast cancer”. *Clinical Cancer Research*, Vol. 8, No. 5, pp. 1210–6, 2002.

[Deth 02] T. Dethlefsen and D. Dahlke. *The Healing Power of Illness.* Vega Books, 2002. ISBN 1843330482.

[Diet 95] T. Dietterich and G. Bakiri. "Solving multiclass learning problems via error-correcting output codes". *Journal of Artificial Intel ligence Research*, pp. 263–286, 1995.

[Ding 03] C. Ding and H. Peng. "Minimum Redundancy Feature Selection from Microarray Gene Expression Data". In: *Proceedings of the Computational Systems Bioinformatics*, 2003.

[Disk 06] T. Diskin, S.J.and Eck, J. Greshock, Y. Mosse, T. Naylor, C. Stoeckert Jr, J. WeberB.L. and. Maris, and G. Grant. "STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments". *Genome Research*, Vol. 16, No. 9, pp. 1149–1158, 2006.

[Dres 03] M. Dressman, A. Baras, R. Malinowski, L. Alvis, I. Kwon, T. Walz, and M. Polymeropoulos. "Gene expression profiling detects gene amplification and differentiates tumor types in breast cancer.". *Cancer Research*, Vol. 63, pp. 2194–2199, 2003.

[Duda 01] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* John Wiley & Sons, Inc., second Ed., 2001.

[Dudo 02] S. Dudoit, J. Fridlyand, and T. Speed. "Comparison of discrimination methods for the classification of tumors using gene expression data". *Journal of the American Statistical Association*, Vol. 97, No. 457, pp. 77–87, 2002.

[Dudo 03] S. Dudoit and J. Fridlyand. *Statistical analysis of gene expression microarray data*, Chap. 3. 2003.

[Dugg 99] D. Duggan, M. Bittner, Y. Chen, P. Meltzer, and J. Trent. "Expression profiling using cDNA microarrays". *Nature genetics*, Vol. 21, pp. 10–14, 1999.

[Duin 04] R. P. W. Duin, P. Juszczak, D. de Ridder, P. Paclík, E. Pekalska, and D. M. J. Tax. "PR-Tools 4.0, a Matlab toolbox for pattern recognition". Tech. Rep., ICT Group, TU Delft, The Netherlands, January 2004. `http://www.prtools.org`.

[Duin 95] R. Duin. "Small sample size generalization". In: *9th Scandinavian Conf. on Image Analysis*, pp. 957–964, 1995.

[Dupu 07] A. Dupuy and R. Simon. "Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting". *Journal of the National Cancer Institute*, Vol. 99, No. 2, pp. 147–157, 2007.

[Eden 07] E. Eden, D. Lipson, S. Yogev, and Z. Yakhini. "Discovering Motifs in Ranked List of DNA Sequences". *PLoS computational biology*, Vol. 3, No. 3, pp. 508–522, 2007.

[Edgr 06] H. Edgren and O. Kallioniemi. "Integrated breast cancer genomics". *Cancer Cell*, Vol. 10, No. 6, pp. 453–454, 2006.

[Ein 05] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany. "Outcome signature genes in breast cancer: is there a unique set?". *Bioinformatics*, Vol. 21, No. 2, pp. 171–178, 2005.

[Ein 06] L. Ein-Dor, O. Zuk, and E. Domany. “Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer”. *PNAS*, Vol. 103, No. 15, pp. 5923–5928, 2006.

[Fan 06] C. Fan, D. Oh, L. Wessels, B. Weigelt, D. Nuyten, A. Nobel, L. van’t Veer, and C. Perou. “Concordance among gene-expression-based predictors for breast cancer.”. *The New England Journal of Medicine*, Vol. 355, No. 6, pp. 560–569, 2006.

[Fish 36] R. Fisher. “The use of multiple measurements in taxonomic problems”. *Ann. Eugenics*, Vol. 7, pp. 179–188, 1936.

[Frid 06] J. Fridlyand, A. Snijders, and B. Ylstra et al. “Breast tumor copy number aberration phenotypes and genomic instability”. *BMC Cancer*, Vol. 6, No. 96, 2006.

[Frie 89] J. Friedman. “Regularized Discriminant Analysis”. *Am. Statistical Assoc.*, 1989.

[Fure 00] T. Furey, N. Christianini, N. Duffy, D. Bednarski, M. Schummer, and D. Hauessler. “Support vector machine classification and validation of cancer tissue samples using microarray expression data”. *Bioinformatics*, Vol. 16, No. 10, pp. 906–914, 2000.

[Furg 05] K. Furge, K. Dykema, C. Ho, and X. Chen. “Comparison of array-based comparative genomic hybridization with gene expression-based regional expression biases to identify genetic abnormalities in hepatocellular carcinoma”. *BMC Genomics*, Vol. 6, No. 67, 2005.

[Geis 01] S. Geisler, P. Lonning, T. Aas, H. Johnsen, O. Fluge, and D. e. a. Haugen. “Influence of TP53 gene alterations and c-erbB-2 expression on the response to treatment with doxorubicin in locally advanced breast cancer”. *Cancer Research*, Vol. 61, No. 6, pp. 2505–12, 2001.

[Gema 04] D. Geman, C. d’Avignon, D. Naiman, and R. Winslow. “Classifying Gene Expression Profiles from Pairwise mRNA Comparisons”. *Statistical Applications in Genetics and Molecular Biology*, Vol. 3, No. 1, 2004.

[Gold 07] A. Goldhirsch, W. Wood, R. Gelber, A. Coates, B. Thurlimann, and H. Senn. “Progress and promise: highlights of the international expert consensus on the primary therapy of early breast cancer”. *Ann Oncol 2007*, Vol. 18, No. 7, pp. 1133–1144, 2007.

[Golu 99] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander. “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring”. *Science*, Vol. 286, pp. 531–537, 1999.

[Gong 07] Y. Gong, K. Yan, F. Lin, K. Anderson, C. Sotiriou, and F. e. a. Andre. “Determination of oestrogen-receptor status and ERBB2 status of breast carcinoma: a gene-expression profiling study”. *Lancet Oncology*, Vol. 8, No. 3, pp. 203–11, 2007.

[Grat 02] L. Grate, C. Bhattacharyya, M. Jordan, and I. Mian. “Simultaneous classification and relevant feature identification in high-dimensional spaces”. In: *Workshop on Algorithms in Bioinformatics*, 2002.

[Guan 05] Z. Guan and H. Zhao. “A semiparametric approach for marker gene selection based on gene expression data”. *Bioinformatics*, Vol. 21, No. 4, pp. 529–536, 2005.

[Guo 02] X. Guo and W. e. a. Lui. “Identifying cancer-related genes in nasopharyngeal carcinoma cell lines using DNA and mRNA expression profiling analyses.”. *International Journal of Oncology*, Vol. 21, pp. 1197–1204, 2002.

[Guyo 02] I. Guyon, J. Weston, and S. Barnhill. “Gene Selection for Cancer Classification using Support Vector Machines”. *Machine Learning*, No. 46, pp. 389–422, 2002.

[Guyo 03] I. Guyon, J. Weston, and S. Barnhill. “Gene Selection for Cancer Classification using Support Vector Machines”. 2003. http://www.clopinet.com/isabelle/Papers/ RFE-erratum.html.

[Hame 02] R. Hamer. *Summary of the New Medicine*. Editiones de la Nueva Medicina S.L., 2002.

[Han 06] W. Han, M. Han, J. Kang, J. Bae, J. Lee, and Y. e. a. Bae. “Genomic alterations identified by array comparative genomic hybridization as prognostic markers in tamoxifen-treated estrogen receptor-positive breast cancer”. *BMC Cancer*, Vol. 6, No. 92, 2006.

[Hann 05] J. Hannemann, H. Oosterkamp, C. Bosch, A. Velds, L. Wessels, C. Loo, E. Rutgers, S. Rodenhuis, and M. van de Vijver. “Changes in Gene Expression Associated With Response to Neoadjuvant Chemotherapy in Breast Cancer”. *International Journal of Oncology*, Vol. 23, No. 15, pp. 3331–3342, 2005.

[Hick 06] J. Hicks, A. Krasnitz, B. Lakshmi, N. Navin, M. Riggs, and E. e. a. Leibu. “Novel patterns of genome rearrangement and their association with survival in breast cancer”. *Genome Research*, Vol. 16, No. 22, pp. 1465–79, 2006.

[Ho 95] T. K. Ho. “Random Decision Forests”. In: *3rd Int’l Conference on Document Analysis and Recognition*, pp. 278–282, 1995.

[Ho 98] T. K. Ho. “The Random Subspace Method for Constructing Decision Forests”. *IEEE Trans. Pattern Analysis and Machine Inteligence*, Vol. 20, No. 8, pp. 832–844, 1998.

[Horl ed] H. Horlings, C. Lai, P. Kristel, E. van Beers, C. Klijn, S. Joosse, F. Reyal, D. Nuyten, C. Froyland, A. Borresen-Dale, M. Reinders, P. Nederlof, L. Wessels, and M. van de Vijver. “Integration of DNA copy number alterations and prognostic gene signatures to predict prognosis of patients with breast cancer”. *Cancer Research*, submitted.

[Hu 06] Z. Hu, C. Fan, D. Oh, J. Marron, X. He, and B. e. a. Qaqish. “The molecular portraits of breast tumors are conserved across microarray platforms”. *BMC Genomics*, Vol. 7, No. 96, 2006.

[Hube 03] W. Huber, A. von Heydebreck, H. Sueltmann, A. Poustka, and M. Vingron. “Parameter estimation for the calibration and variance stabilization of microarray data”. *Statistical Applications in Genetics and Molecular Biology*, Vol. 2, 2003.

[Hwan 05] D. Hwang, A. Rust, S. Ramsey, J. Smith, D. Leslie, A. Weston, P. de Atauri, J. Aitchison, L. Hood, A. Siegel, and H. Bolouri. “A data integration methodology for systems biology”. *PNAS*, Vol. 102, No. 48, pp. 17296–17301, 2005.

[Hyma 02] E. Hyman and P. Kauraniemi. “Impact of DNA Amplification on Gene Expression Patterns in Breast Cancer”. *Cancer Research*, Vol. 62, pp. 6240–6245, 2002.

[Jaeg 03] J. Jaeger, R. Sengupta, and W. Ruzzo. “Improved Gene Selection For Classification Of Microarrays”. In: *Pacific Symposium on Biocomputing*, 2003.

[Jain 97] A. Jain and D. Zongker. “Feature Selection: Evaluation, Application, and Small Sample Performance”. *IEEE Trans. Pattern Analysis and Machine Inteligence*, Vol. 19, No. 2, 1997.

[Jeff 07] I. Jeffery, S. Madden, P. McGettigan, G. Perriere, A. Culhane, and D. Higgins. “Integrating transcription factor binding site information with gene expression datasets.”. *Bioinformatics*, Vol. 23, No. 3, pp. 298–305, 2007.

[Jian] D. Jiang, C. Tang, and A. Zhang. “Cluster Analysis for Gene Expression Data: A Survey”.

[Jone 02] P. Jones and S. Baylin. “The fundamental role of epigenetic events in cancer.”. *Nature ReviewsGenetics*, Vol. 3, No. 6, pp. 415–428, 2002.

[Jong 03] K. Jong, E. Marchiori, A. van der Vaart, B. Ylstra, M. Weiss, and G. Meijer. “Chromosomal Breakpoint Detection in Human Cancer”. In: SPRINGER, Ed., *Applications of Evolutionary Computing. EvoBIO: Evolutionary Computation and Bioinformatics*, pp. 54–65, 2003.

[Jong 04] K. Jong, E. Marchiori, G. Meijer, A. van der Vaart, and B. Ylstra. “Breakpoint Identification and Smoothing of array Comparative Genomic Hybridization data”. *Bioinformatics*, Vol. 20, No. 18, pp. 3636–3637, 2004.

[Joos 07] S. Joosse, E. van Beers, and P. Nederlof. “Automated array-CGH optimized for archival formalin-fixed, paraffin-embedded tumor material”. *BMC Cancer*, Vol. 7, No. 43, 2007.

[Kall 92] A. Kallioniemi, O. Kallioniemi, D. Sudar, D. Rutovitz, J. Gray, F. Waldman, and D. Pinkel. “Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors.”. *Science*, Vol. 258, pp. 818–821, 1992.

[Kaur 01] P. Kauraniemi, M. Barlund, O. Monni, and A. Kallioniemi. “New Amplified and Highly Expressed Genes Discovered in the ERBB2 Amplicon in Breast Cancer by cDNA Microarrays”. *Cancer Research*, Vol. 61, pp. 8235–8240, 2001.

[Khan 01] J. Khan, J. Wei, M. Ringner, L. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. Antonescu, C. Peterson, and P. Meltzer. “Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks”. *Nature Medicine*, Vol. 7, No. 6, pp. 673–79, 2001.

[Kiku 03] S. Kikuchi, D. Tominaga, M. Arita, K. Takahashi, and M. Tomita. “Dynamic modeling of genetic networks using genetic algorithm and S-system”. *Bioinformatics*, Vol. 19, No. 5, pp. 643–50, 2003.

[Koha 95] R. Kohavi. "The Power of Decision Tables". In: *Proceedings of the European Conference on Machine Learning*, 1995.

[Koha 97] G. Kohavi, R.and John. "Wrappers for Feature Subset Selection". *Artificial Intelligence*, Vol. 97, pp. 273–324, 1997.

[Kors 05] E. Korsching, J. Packeisen, C. Liedtke, D. Hungermann, P. Wulfing, and P. e. a. van Diest. "The origin of vimentin expression in invasive breast cancer: epithelial-mesenchymal transition, myoepithelial histogenesis or histogenesis from progenitor cells with bilinear differentiation potential?". *Journal of Pathology*, Vol. 206, No. 4, pp. 451–7, 2005.

[Kuma 07] M. Kumar, J. Lu, K. Mercer, T. Golub, and T. Jacks. "Impaired microRNA processing enhances cellular transformation and tumorigenesis". *Nature Genetics*, Vol. 39, No. 5, pp. 673–7, 2007.

[Lai 04] C. Lai, M. Reinders, and L. Wessels. "On univariate selection methods in gene expression datasets". In: *Tenth Annual Conference of the Advanced School for Computing and Imaging.*, pp. 335–341, The Netherlands, 2004.

[Lai 05] W. Lai, M. Johnson, R. Kucherlapati, and P. Park. "Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data". *Bioinformatics*, Vol. 21, No. 19, pp. 3763–3770, 2005.

[Lai 06a] C. Lai, M. Reinders, L. van't Veer, and L. Wessels. "A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets". *BMC Bioinformatics*, Vol. 7, No. 235, 2006.

[Lai 06b] C. Lai, M. Reinders, and L. Wessels. "Random Subspace Method for multivariate feature selection". *Patter Recognition Letters*, Vol. 27, No. 10, pp. 1067–1076, 2006.

[Lai 07] C. Lai, H. Horlings, M. van de Vijver, E. van Beers, P. Nederlof, L. Wessels, and M. Reinders. "SIRAC: Supervised Identification of Regions of Aberration in aCGH datasets". *BMC Bioinformatics*, Vol. 8, p. 422, 2007.

[Lee 03] S. Lee and S. Batzoglou. "Application of independent component analysis to microarrays.". *Genome Biology*, Vol. 4, No. 11, 2003.

[Leng 98] C. Lengauer, K. Kinzler, and B. Vogelstein. "Genetic instabilities in human cancer". *Nature*, Vol. 396, No. 6712, pp. 643–649, 1998.

[Lete 06] A. Letessier, F. Sircoulomb, C. Ginestier, N. Cervera, F. Monville, and V. e. a. Gelsi-Boyer. "Frequency, prognostic impact, and subtype association of 8p12, 8q24, 11q13, 12p13, 17q12, and 20q13 amplifications in breast cancers". *BMC Cancer*, Vol. 6, No. 245, 2006.

[Levi 05] D. Levin, A.M.and Ghosh, K. Cho, and S. Kardia. "A model-based scan statistic for identifying extreme chromosomal regions of gene expression in human tumors.". *Bioinformatics*, Vol. 21, No. 12, pp. 2867–2874, 2005.

[Li 01] L. Li, C. Weinberg, T. Darden, and L. Pedersen. "Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method". *Bioinformatics*, Vol. 17, No. 12, pp. 1131–42, 2001.

[Li 07] Y. Li, L. Wessels, D. de Ridder, and M. Reinders. “Classification in the Presence of Class Noise Using a Probabilistic Kernel Fisher Method”. *Pattern Recognition*, Vol. 40, No. 12, pp. 3349–3357, 2007.

[Lieb 02] W. Liebermeister. “Linear modes of gene expression determined by independent component analysis.”. *Bioinformatics*, Vol. 18, No. 1, pp. 51–60, 2002.

[Loo 04] L. Loo, D. Grove, E. Williams, C. Neal, L. Cousens, E. Schubert, I. Holcomb, H. Massa, J. Glogovac, C. Li, K. Malone, J. Daling, J. Delrow, B. Trask, L. Hsu, and P. Porter. “Array comparative genomic hybridization analysis of genomic alterations in breast cancer subtypes”. *Cancer Research*, Vol. 64, pp. 8541–8549, 2004.

[Ma 04] X. Ma, Z. Wang, P. Ryan, S. Isakoff, A. Barmettler, A. Fuller, B. Muir, G. Mohapatra, R. Salunga, and J. Tuggle. “A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen.”. *Cancer Cell*, Vol. 5, No. 6, pp. 607–616, 2004.

[Mart 03] J. Martinez-Climent, A. Alizadeh, R. Segraves, D. Blesa, F. Rubio-Moscardo, D. Albertson, J. Garcia-Conde, M. Dyer, R. Levy, D. Pinkel, and I. Lossos. “Transformation of follicular lymphoma to diffuse large cell lymphoma is associated with a heterogeneous set of DNA copy number and gene expression alterations.”. *Neoplasia*, Vol. 101, No. 8, pp. 3109–3117, 2003.

[Maty 03] V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, and R. e. a. Hehl. “TRANSFAC(R): transcriptional regulation, from patterns to profiles”. *Nucleic Acids Research*, Vol. 31, No. 1, pp. 374–8, 2003.

[Mele 05] B. Melendez, B. Martinez-Delgado, M. Cuadros, V. Fernandez, R. Diaz-Uriarte, and J. Benitez. “Identification of amplified and highly expressed genes in amplicons of the T-cell line huT78 detected by cDNA microarray CGH.”. *Molecular Cancer*, Vol. 4, No. 5, 2005.

[Mich 05] S. Michiels, S. Koscielny, and C. Hill. “Prediction of cancer outcome with microarrays: a multiple random validation strategy”. *The Lancet*, Vol. 365, pp. 488–92, 2005.

[Mill 05] L. Miller, J. Smeds, J. George, V. Vega, L. Vergara, A. Ploner, Y. Pawitan, P. Hall, S. Klaar, E. Liu, and J. Bergh. “An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival”. *PNAS*, 2005.

[Monn 01] O. Monni, M. Barlund, S. Mousses, J. Kononen, G. Sauter, and M. e. a. Heiskanen. “Comprehensive copy number and gene expression profiling of the 17q23 amplicon in human breast cancer”. *PNAS*, Vol. 98, No. 10, pp. 5711–6, 2001.

[Nade 07] A. Naderi, A. Teschendorff, N. Barbosa-Morais, S. Pinder, A. Green, D. Powe, J. Robertson, S. Aparicio, I. Ellis, J. Brenton, and C. Caldas. “A gene-expression signature to predict survival in breast cancer across independent data sets”. *Oncogene*, Vol. 26, pp. 1507–1516, 2007.

[Nam 06] D. Nam, S. Kim, S. Kim, S. Yang, S. Kim, and I. Chu. “ADGO: analysis of differentially expressed gene sets using composite GO annotation”. *Bioinformatics*, Vol. 22, No. 18, pp. 2249–2253, 2006.

[Nayl 05] T. Naylor, J. Greshock, Y. Wang, T. Colligon, Q. Yu, V. Clemmer, T. Zaks, and B. Weber. "High resolution genomic analysis of sporadic breast cancer using array-based comparative genomic hybridization". *Breast Cancer Research*, Vol. 7, No. 6, pp. R1186–R1198, 2005.

[Ness 05] M. Nessling, K. Richter, C. Schwaenen, P. Roerig, G. Wrobel, S. Wessendorf, B. Fritz, M. Bentz, H. Sinn, B. Radlwimmer, and P. Lichter. "Candidate genes in breast cancer revealed by microarray-based comparative genomic hybridization of archived tissue". *Cancer Research*, Vol. 65, pp. 439–447, 2005.

[Newt 01] M. Newton, C. Kendziorski, C. Richmond, F. Blattner, and K. Tsui. "On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data". *Journal of Computational Biology*, Vol. 8, No. 1, pp. 37–52, 2001.

[Noto 06] K. Noto and M. Craven. "Learning probabilistic models of cis-regulatory modules that represent logical and spatial aspects". *Bioinformatics*, Vol. 23, pp. e156–e162, 2006.

[Nova 06] P. Novak, T. Jensen, M. Oshiro, R. Wozniak, M. Nouzova, and G. e. a. Watts. "Epigenetic inactivation of the HOXA gene cluster in breast cancer". *Cancer Research*, Vol. 66, No. 22, pp. 10664–70, 2006.

[Nyma 06] P. Nymark, H. Wikman, S. Ruosaari, G. Hollmen, E. Vanhala, A. Karjalainen, and S. Anttila, S.and Knuutila. "Identification of Specific Gene Copy Number Changes in Asbestos-Related Lung Cancer". *Cancer Research*, Vol. 66, No. 11, pp. 5737–5743, 2006.

[Olsh 04] A. Olshen, E. Venkatraman, R. Lucito, and M. Wigler. "Circular binary segmentation for the analysis of array-based DNA copy number data". *Biostatistics*, Vol. 5, No. 4, pp. 557–572, 2004.

[Orse 04] B. Orsetti, M. Nugoli, N. Cervera, L. Lasorsa, P. Chuchana, and L. e. a. Ursule. "Genomic and expression profiling of chromosome 17 in breast cancer reveals complex patterns of alterations and novel candidate genes". *Cancer Research*, Vol. 64, No. 18, pp. 6453–60, 2004.

[Orse 05] B. Orsetti, M. Nugoli, N. Cervera, L. Lasorsa, P. Chuchana, and C. e. a. Rouge. "Genetic profiling of chromosome 1 in breast cancer: mapping of regions of gains and losses and identification of candidate genes on 1q". *British Journal of Cancer*, Vol. 95, No. 10, pp. 1439–47, 2005.

[Pacl 05] P. Paclík, T. C. Landgrebe, and R. P. Duin. "PRExp 2.0, a Matlab toolbox for evaluation of pattern recognition experiment". Tech. Rep., ICT Group, TU Delft, The Netherlands, Dec. 2005.

[Pero 00] C. Perou, T. Sorlie, M. Eisen, M. van de Rijn, S. Jeffrey, C. Rees, J. Pollack, D. Ross, H. Johnsen, L. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. Zhu, P. Lonning, A. Borresen-Dale, P. Brown, and D. Botstein. "Molecular portraits of human breast tumours.". *Nature*, Vol. 406, No. 6797, pp. 747–752, 2000.

[Pero 99] C. Perou, S. Jeffrey, M. van de Rijn, C. Rees, M. Eisen, D. Ross, A. Pergamenschikov, C. Williams, S. Zhu, J. Lee, D. Lashkari, D. Shalon, P. Brown, and D. Botstein. “Distinctive gene expression patterns in human mammary epithelial cells and breast cancers.”. *PNAS*, Vol. 96, No. 16, pp. 9212–9217, 1999.

[Pica 05] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J. Daudin. “A statistical approach for array CGH data analysis”. *BMC Bioinformatics*, Vol. 6, No. 27, 2005.

[Pink 05] D. Pinkel and D. Albertson. “Array comparative genomic hybridization and its applications to cancer”. *Nature Genetics*, Vol. 37, pp. s11–s17, 2005.

[Poll 02] T. Pollack, J.R.ăand Sorlie, C. Perou, and C. Rees. “Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors”. *PNAS*, Vol. 99, No. 20, pp. 12963–12968, 2002.

[Pome 02] S. Pomeroy, P. Tamayo, M. Gaasenbeek, L. Sturla, M. Angelo, M. McLaughlin, J. Kim, L. Goumnerova, P. Black, J. Lau, C. Allen, D. Zagzag, J. Olson, T. Curran, C. Wetmore, J. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. Louis, J. Mesirov, E. Lander, and T. Golub. “Prediction of central nervous system embryonal tumour outcome based on gene expression”. *Nature*, Vol. 415, pp. 436–442, 2002.

[Pudi 94] P. Pudil, J. Novovicova, and J. Kittler. “Floating search methods in feature selection”. *Pattern Recognition Letters*, Vol. 15, pp. 1119–1125, 1994.

[Raap 04] A. Raap, M. van der Burg, J. Knijnenburg, E. Meershoek, C. Rosenberg, and J. e. a. Gray. “Array comparative genomic hybridization with cyanin cis-platinum-labeled DNAs”. *Biotechniques*, Vol. 37, No. 1, pp. 130–4, 2004.

[Rako 03] A. Rakotomamonjy. “Variable Selection using SVM-based criteria”. *Journal of Machine Learning Research, Special Issue on Variable Selection*, Vol. 3, pp. 1357–1370, 2003.

[Rama 03] S. Ramaswamy, K. Ross, E. Lander, and T. Golub. “A molecular signature of metastasis in primary solid tumors”. *Nature*, Vol. 33, pp. 49–54, 2003.

[Raud 91] S. Raudys and A. Jain. “Small sample size effect in statistical pattern recognition: recommendations for practitioners”. *IEEE Trans. Pattern Analysis and Machine Inteligence*, Vol. 13, No. 3, pp. 252–264, 1991.

[Ravd 01] P. Ravdin, L. Siminoff, G. Davis, M. Mercer, J. Hewlett, and N. e. a. Gerson. “Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer.”. *Journal of Clinical Oncology*, Vol. 19, No. 4, pp. 980–991, 2001.

[Rees 01] L. Rees and A. Weil. “Imbues orthodox medicine with the values of complementary medicine”. *British Medical Journal*, Vol. 322, pp. 119–120, 2001.

[Reid 05] J. Reid, L. Lusa, L. De Cecco, D. Coradini, S. Veneroni, M. Daidone, M. Gariboldi, and M. Pierotti. “Limits of predictive models using microarray data for breast cancer clinical treatment outcome”. *Journal of the National Cancer Institute*, Vol. 97, No. 12, pp. 927–930, 2005.

[Reya 05] F. Reyal, N. Stransky, I. Bernard-Pierrot, A. Vincent-Salomon, Y. de Rycke, P. Elvin, A. Cassidy, A. Graham, C. Spraggon, Y. Desille, A. Fourquet, C. Nos, P. Pouillart, H. Magdelenat, D. Stoppa-Lyonnet, J. Couturier, B. Sigal-Zafrani, B. Asselain, X. Sastre-Garau, O. Delattre, J. Thiery, and F. Radvanyi. "Visualizing Chromosomes as Transcriptome Correlation Maps: Evidence of Chromosomal Domains Containing Co-expressed Genes. A Study of 130 Invasive Ductal Breast Carcinomas". *Cancer Research*, Vol. 65, No. 4, pp. 1376–1383, 2005.

[Roep 05] L. Roepman, P.and Wessels, N. Kettelarij, P. Kemmeren, A. Miles, M. Lijnzaad, P.and Tilanus, R. Koole, G. Hordijk, P. Van der Vliet, M. Reinders, P. Slootweg, and F. Holstege. "An expression profile for diagnosis of lymph node metastases from primary head and neck squamous cell carcinomas". *Nature Genetics*, Vol. 37, pp. 182–186, 2005.

[Sanc 03] D. Sanchez-Izquierdo, G. Buchonnet, R. Siebert, R. Gascoyne, J. Climent, L. Karran, M. Marin, D. Blesa, D. Horsman, A. Rosenwald, L. Staudt, D. Albertson, M. Du, H. Ye, P. Marynen, J. Garcia-Conde, D. Pinkel, M. Dyer, and J. Martinez-Climent. "MALT1 is deregulated by both chromosomal translocation and amplification in B-cell non-Hodgkin lymphoma". *Blood*, Vol. 101, pp. 4539 – 4546, 2003.

[Sara 06] O. Saramaki, K. Porkka, R. Vessella, and T. Visakorpi. "Genetic aberrations in prostate cancer by microarray analysis". *International Journal of Cancer*, Vol. 119, pp. 1322–1329, 2006.

[Saxe 01] G. Saxe, J. Hebert, J. Carmody, J. Kabat-Zin, P. Rosenzweig, D. Jarzobski, G. Reed, and R. Blute. "Can Diet in Conjunction with Stress reduction Affect the Rate of Increase in Prostate-specific Antigen After Biochemical Recurrence of Prostate Cancer?". *Journal of Urology*, Vol. 166, pp. 2202–2207, 2001.

[Sche 95] M. Schena, D. Shalon, R. Davis, and P. Brown. "Quantitative monitoring of gene expression patterns with a complementary DNA microarray". *Science*, Vol. 270, No. 5235, pp. 467–470, 1995.

[Schw 04] C. Schwaenen, M. Nessling, S. Wessendorf, T. Salvi, G. Wrobel, B. Radlwimmer, H. Kestler, C. Haslinger, S. Stilgenbauer, H. Dohner, M. Bentz, and P. Lichter. "Automated array-based genomic profiling in chronic lymphocytic leukemia: Development of a clinical tool and discovery of recurrent genomic alterations". *PNAS*, Vol. 101, No. 4, pp. 1039–1044, 2004.

[Sega 04] E. Segal, N. Friedman, D. Koller, and A. Regev. "A module map showing conditional activity of expression modules in cancer". *Nature genetics*, Vol. 36, No. 10, pp. 1090–1098, 2004.

[Shas 03] B. Shastry. "SNPs and haplotypes: Genetic markers for disease and drug response.". *International Journal of Molecular Medicine*, Vol. 11, pp. 379–382, 2003.

[Silv 05] P. Silva, R. Hashimoto, S. Kim, J. Barrera, L. Brandao, E. Suh, and E. Dougherty. "Feature selection algorithms to find strong genes". *Pattern Recognition Letters*, Vol. 26, No. 10, pp. 1444–1453, 2005.

[Sing 02] D. Singh, P. Febbo, K. Ross, D. Jackson, J. Manola, C. Ladd, P. Tamayo, A. Renshaw, A. D'Amico, J. Richie, E. Lander, M. Loda, P. Kantoff, T. Golub, and W. Sellers. "Gene expression correlates of clinical prostate cancer behavior". *Cancer Cell.*, Vol. 1, pp. 203–209, 2002.

[Skur 01] M. Skurichina. *Stabilizing weak classifiers.* PhD thesis, Delft,Technical University, 2001.

[Skur 02] M. Skurichina and R. Duin. "Bagging, Boosting and the Random Subspace Method for Linear Classifiers". *Pattern Analysis and Applications*, Vol. 5, pp. 121–135, 2002.

[Sorl 01] T. Sorlie, C. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. Eisen, M. van de Rijn, S. Jeffrey, T. Thorsen, H. Quist, J. Matese, P. Brown, D. Botstein, E. Lonning, and A. Borresen-Dale. "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications". *PNAS*, Vol. 98, No. 19, pp. 10869–10864, 2001.

[Sorl 03] T. Sorlie, R. Tibshirani, J. Parker, T. Hastie, J. Marron, S. Nobel, A.and Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C. Perou, P. Lonning, P. Brown, A. Borresen-Dale, and D. Botstein. "Repeated observation of breast tumor subtypes in independent gene expression data sets". *PNAS*, Vol. 100, No. 14, pp. 8418–8423, 2003.

[Soti 03] C. Sotiriou, S. Neo, L. McShane, E. Korn, P. Long, and A. e. a. Jazaeri. "Breast cancer classification and prognosis based on gene expression profiles from a population-based study.". *PNAS*, Vol. 100, No. 18, pp. 10393–8, 2003.

[Soti 07] C. Sotiriou and M. Piccart. "Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care?". *Nature Reviews Cancer*, Vol. 7, No. 7, pp. 545–553, 2007.

[Spec 00] M. Speca, L. Carlson, E. Linda, E. Goodey, and M. Angen. "A Randomized Wait-List Controlled Trial: The Effect of a Mindfulness Meditation-Based Program on Mood and Symptoms of Stress in Cancer Outpatients". *Psychosomatic Medicine*, Vol. 62, pp. 613–622, 2000.

[Spee 03] T. Speed, Ed. *Statistical analysis of gene expression microarray data.* 2003.

[Stat 05a] A. Statnikov, C. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy. "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis". *Bioinformatics*, Vol. 21, No. 5, pp. 631–643, 2005.

[Stat 05b] A. Statnikov, Y. Tsamardinos, I.and Dosbayev, and C. Aliferis. "GEMS: A system for automated cancer diagnosis and biomarker discovery from microarray gene expression data.". *International Journal of Medical Informatics*, Vol. 74, pp. 491–503, 2005.

[Stek 03] D. Stekel. *Microarray Bioinformatics.* Cambridge University Press, 2003.

[Stra 06] N. Stransky, C. Vallot, F. Reyal, I. Bernard-Pierrot, S. Diez de Medina, R. Segraves, Y. de Rycke, P. Elvin, A. Cassidy, C. Spraggon, A. Graham, J. Southgate, B. Asselain, Y. Allory, C. Abbou, D. Albertson, J. Thiery, D. Chopin, D. Pinkel, and F. Radvanyi. "Regional copy number■independent deregulation of transcription in cancer". *Nature Genetics*, Vol. 38, pp. 1386–1396, 2006.

[Stra 07] B. Stranger, M. Forrest, M. Dunning, C. Ingle, C. Beazley, N. Thorne, R. Redon, C. Bird, A. de Grassi, C. Lee, C. Tyler-Smith, N. Carter, S. Scherer, S. Tavare, P. Deloukas, M. Hurles, and E. Dermitzakis. "Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes". *Science*, Vol. 315, No. 5813, pp. 848–853, 2007.

[Tax 02] D. Tax and R. Duin. "Using Two-Class Classifiers for Multiclass Classification". In: *16th International Conference on Pattern Recognition*, 2002.

[Tesc 06] A. Teschendorff, A. Naderi, N. Barbosa-Morais, S. Pinder, I. Ellis, S. Aparicio, J. Brenton, and C. Caldas. "A consensus prognostic gene expression classifier for ER positive breast cancer". *Genome Biology*, Vol. 7, No. 10, p. R101, 2006.

[Thie 06] J. Thiery, X. Sastre-Garau, B. Vincent-Salomon, X. Sigal-Zafrani, J. Pierga, C. Decraene, J. Meyniel, E. Gravier, B. Asselain, Y. De Rycke, P. Hupe, E. Barillot, S. Ajaz, M. Faraldo, M. Deugnier, M. Glukhova, and D. Medina. "Challenges in the stratification of breast tumors for tailored therapies". *Electronic Journal of Oncology*, 2006.

[Tibs 02] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. "Diagnosis of multiple cancer types by shrunken centroids of gene expression". *PNAS*, Vol. 99, No. 10, pp. 6567–6572, 2002.

[Trun 79] G. Trunk. "A problem af dimensionality: a simple example". *IEEE Trans. Pattern Analysis and Machine Inteligence*, Vol. 1, No. 3, July 1979.

[Tsam 03] C. Tsamardinos, I.and Aliferis. "Towards Principled Feature Selection: Relevancy, Filters and Wrappers". In: *in Ninth International Workshop on Artificial Intelligence and Statistics.*, 2003.

[Tush 01] V. Tusher, R. Tibshirani, and G. Chu. "Significance analysis of microarrays applied to the ionizing radiation response". *PNAS*, Vol. 98, No. 9, pp. 5116–5121, 2001.

[Veer 02] L. van 't Veer, H. Dai, M. van de Vijver, D. H. Yudong, A. Hart, M. Mao, H. Peterse, K. van der Kooy, M. Marton, A. Witteveen, G. Schreiber, R. Kerkhoven, C. Roberts, P. Linsley, R. Bernards, and S. Friend. "Gene expression profiling predicts clinical outcome of breast cancer". *Nature*, Vol. 415, pp. 530–536, 2002.

[Veld 04] D. de Veld, M. Skurichina, M. Witjes, R. Duin, D. Sterenborg, W. Star, and J. Roodenburg. "Clinical study for classification of benign, dysplastic, and malignant oral lesions using autofluorescence spectroscopy". *Journal of Biomedical Optics*, Vol. 9, No. 5, pp. 940–950, 2004.

[Velt 03] J. Veltman, J. Fridlyand, S. Pejavar, A. Olshen, J. Korkola, S. DeVries, W. Carroll, P.and Kuo, D. Pinkel, D. Albertson, C. Cordon-Cardo, A. Jain, and F. Waldman. "Array-based Comparative Genomic Hybridization for Genome-Wide Screening of DNA Copy Number in Bladder Tumors". *Cancer Research*, No. 63, pp. 2872–2880, 2003.

[Vijv 02] M. van de Vijver, Y. He, L. van t Veer, H. Dai, A. Hart, D. Voskuil, G. Schreiber, J. Peterse, C. Roberts, M. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards. "A Gene-Expression

Signature as a Predictor of Survival in Breast Cancer". *The New England Journal of Medicine*, Vol. 347, No. 25, pp. 1999–2009, 2002.

[Visv 03] J. Visvader and G. Lindeman. "Transcriptional regulators in mammary gland development and cancer". *Int J Biochem Cell Biol*, Vol. 35, No. 7, pp. 1034–51, 2003.

[Wang 05a] P. Wang, Y. Kim, J. Pollack, B. Narasimhan, and R. Tibshirani. "A method for calling gains and losses in array CGH data". *Biostatistics*, Vol. 6, No. 1, pp. 45–58, 2005.

[Wang 05b] 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*, Vol. 265, No. 9460, pp. 671–679, 2005.

[Weil 04] A. Weil. *Health and Healing.* Mariner Books, 2004. ISBN 0618479082.

[Wess 05] L. Wessels, M. Reinders, A. Hart, C. Veenman, H. Dai, Y. He, and L. van 't Veer. "A protocol for building and evaluating predictors of disease state based on microarray data.". *Bioinformatics*, Vol. 21, No. 19, pp. 3755–3762, 2005.

[West 00] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. "Feature Selection for SVMs". In: *Proc of NIPS*, pp. 668–674, 2000.

[Wids 02] M. Widschwendter and P. Jones. "DNA methylation and breast carcinogenesis". *Oncogene*, Vol. 21, No. 35, pp. 5462–82, 2002.

[Xing 01] E. Xing, M. Jordan, and R. Karp. "Feature selection for high-dimensional genomic microarray data". In: *International Conference on Machine Learning*, 2001.

[Xion 01a] M. Xiong, X. Fang, and J. Zhao. "Biomarker Identification by Feature Wrappers". *Genome Research*, Vol. 11, No. 11, pp. 1878–1887, 2001.

[Xion 01b] M. Xiong, W. La, J. Zhao, L. Jin, and E. Boerwinkle. "Feature (Gene) Selection in Gene Expression-Based Tumor Classification". *Molecular Genetics and Metabolism*, Vol. 73, pp. 239–247, 2001.

[Xu 05] L. Xu, A. Tan, D. Naiman, D. Geman, and R. Winslow. "Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data". *Bioinformatics*, Vol. 21, No. 20, pp. 3905–3911, 2005.

[Yang 02] Y. Yang, S. Dudoit, Luu, D. Lin, V. Peng, G. Ngai, and T. Speed. "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation". *Nucleic Acids Research*, Vol. 30, 2002.

[Yeun 01] K. Yeung and W. Ruzzo. "Principal component analysis for clustering gene expression data". *Bioinformatics*, Vol. 17, No. 9, pp. 763–774, 2001.

[Yi 05] Y. Yi, J. Mirosevich, Y. Shyr, R. Matusik, and A. George. "Coupled analysis of gene expression and chromosomal location.". *Genomics*, Vol. 85, pp. 401–412, 2005.

[Zama 96] J. Zamarra, R. Schneider, I. Besseghini, D. Robinson, and J. Salerno. “Usefulness of the transcendental meditation program in the treatment of patients with coronary artery disease.”. *American Journal of Cardiology*, Vol. 77, No. 10, pp. 867–70, 1996.

[Zhan 06a] L. Zhang, J. Huang, N. Yang, J. Greshock, M. Megraw, and A. e. a. Giannakakis. “microRNAs exhibit high frequency genomic alterations in human cancer”. *PNAS*, Vol. 103, No. 24, pp. 9136–41, 2006.

[Zhan 06b] W. Zhang, R. Rekeya, and K. Bertrand. “A method for predicting disease subtypes in presence of misclassification among training samples using gene expression: application to human breast cancer.”. *Bioinformatics*, Vol. 22, No. 3, pp. 317–325, 2006.

Summary

Cancer consists of cells of the body which proliferate in an uncontrolled fashion. Understanding of the genetic mechanisms of the disease would greatly improve its treatment. The advent of high throughput biomolecular measurements, such as gene expression arrays, allows a close look at the molecular mechanisms of cancer. This thesis studies data analysis procedures tailored to analyze and learn from high throughput data, finalized both to clinical applications such as cancer classification, and to build knowledge of biological mechanisms. A gene expression array measures the expression of thousands of genes simultaneously, via the quantification of the mRNA present in the samples. Since it is hypothesized that cancer is reflected in changed mRNA expression levels, analyzing this data potentially gives insights into cancer development or might enable to predict therapies. One reason for altered mRNA expression stems from genetic mutations. These can be measured through DNA copy number arrays . Since human cells are diploid, aberrations in the amount of DNA copies pinpoint towards genetic mutations related to the initiation and development of cancer.

The underlying theme throughout this thesis is the investigation of dependencies in gene expression and/or copy number measurements. Since cancer is a complex disease, one expects that multiple genes are affected simultaneously when a cell becomes tumorous. This in turn implies that the mRNA expression or the DNA copy number of multiple genes will change in concordance with each other, or, in other words, that the change in these genes are dependent on each other. One prominent clinical application that is being studied in this thesis is that of classifying a patient's tumor into one of the categories of interest, e.g. aggressive/non-aggressive cancer, or a tumor that will/will not respond to a certain treatment. This can be achieved by learning from examples, i.e. array data of patients for which the classes of interest (labels) are known. A statistical model, called a classifier, is built (trained) to discriminate the classes of interest. This model should be able to generalize to data unseen during the training process, i.e. to tumor material of new patients.

Part I of this thesis is dedicated to obtain and evaluate a reliable supervised classification procedure using gene expression measurements only. A gene expression array produces measurements for a large number of genes, but not all of them are thought to be involved in the development of cancer. Identifying a limited number of genes compared with the number of genes on the array has several benefits. An informative set of genes would increase the classification performance and it provides the biologists with a tractable number of variables to be evaluated in order to gain understanding of the cancer mechanisms. Moreover, a small number of genes, e.g. in the order of tens, would allow cheaper tests to be used routinely in the clinic. In order to reduce the dimensionality of the original datasets, this thesis focuses on gene selection procedures. New ways to perform outcome prediction have been investigated and these have been compared to

state of the art gene selection and classification procedures in a rigorous framework. The major contribution on this work is the first consistent evaluation study on univariate and multivariate selection techniques, in order to identify the strong and weak characteristics of both approaches.

Supervised classification based on gene expression data provides the possibility to construct generalizing classifiers using gene subsets. However, in order to gain biological insight into the mechanisms of cancer, the statistical analysis of gene expression arrays alone is not sufficient. These insights could be revealed by integrating different sources of information; in particular copy number data, expression data, and the genomic location of alterations in these measurements. This is the theme of Part II of this thesis. First, the focus is on analyzing DNA copy number data on itself. Here the questions is whether there are genomic aberrations that define the classes of interest. For that a systematic search across the complete genome has been built that identifies copy number aberrations specific to the problem under study, e.g. cancer stratification and clinical outcome. Then, this thesis continues to investigate the influence of genome aberrations on alterations of gene expression and proposes procedures to identify genes that are affected by the aberrations in the DNA copy number. The final contribution investigates local and genome-wide spatial relationships between DNA alteration and changes in gene expression. This study pinpoints genome-wide dependencies via the identification of the correlation between a chromosomal aberration in a region and the expression on other locations on the genome.

Samenvatting

De ziekte kanker refereert naar de ongeremde en ongecontroleerde vermenigvuldiging van lichaamscellen. Het verkrijgen van inzicht in de genetische mechanismes die aan de basis staan van deze ziekte, zou een enorme stap voorwaarts betekenen in de behandeling ervan. De komst van biomoleculaire meetapparatuur met grote verwerkingscapaciteit (”high-throughput”), zoals genexpressie-arrays, faciliteert de gedetailleerde studie naar de moleculaire mechanismes van kanker. Dit proefschrift beschrijft het onderzoek naar methodes die zijn gericht op het leren en analyseren van de high-throughput data met als tweeledig doel tumor classificatie in klinische toepassingen en het verzamelen van fundamentele kennis over biologische mechanismes.

Een genexpressie-array kwantificeert de mRNA concentraties van duizenden genen in een monster van cellen en geeft hiermee een indicatie van de expressie van alle genen in een genoom. Aangezien kanker onder andere wordt gekarakteriseerd door veranderde mRNA expressiewaardes, verschaft de analyse van deze data inzicht in tumorontwikkeling en mogelijke therapieën. Eén verklaring voor veranderde expressiewaardes kan worden gevonden in mutaties in het DNA. Deze mutaties kunnen worden gemeten met zogenaamde ”DNA copy number arrays”. Omdat menselijke cellen diploïde zijn, kunnen afwijkingen in de hoeveelheid kopieën van bepaalde stukken DNA verwijzen naar bepaalde genetische mutaties, die gerelateerd zijn aan de initiatie en ontwikkeling van kanker.

Het onderliggende thema van dit proefschrift is het onderzoek naar de afhankelijkheden binnen genexpressie en/of copy number metingen. Aangezien kanker een complexe ziekte is, kan men verwachten dat meerdere genen tegelijkertijd worden beïnvloed, wanneer een cel tumorachtig wordt. Op zijn beurt betekent dit dat de mRNA expressie of het DNA copy number van verschillende genen eendrachtig zal veranderen, of, met andere woorden, dat de verandering in deze genen afhankelijk is van elkaar. Eén belangrijke klinische toepassing, die wordt bestudeerd in dit proefschrift is de classificatie van de tumor van een patiënt in één van een aantal voorbestemde categorieën (of klasses), bv. kwaadaardige/goedaardige kanker, of een tumor die wel/niet reageert op een bepaalde behandeling. Dit doel kan worden bereikt door te leren van voorbeelden, d.w.z. data van patiënten, waarvan de tumorklasses bekend zijn. Een statistisch model, genaamd een klassificator, wordt gebouwd (getraind) om onderscheid te kunnen maken tussen de klasses. Dit model moet daarnaast in staat zijn om data, die niet in het trainingsproces is gebruikt, bv. tumormateriaal van nieuwe patiŚnten, te classificeren.

Deel 1 van dit proefschrift is gewijd aan het verkrijgen en evalueren van een betrouwbaar ”supervised” classificatie systeem, dat enkel gebruik maakt van genexpressiemetingen. Een genexpressie-array produceert metingen voor een groot aantal genen. Echter, niet al deze genen zullen betrekking hebben op de ontwikkeling van kanker. De identificatie van een beperkte groep genen heeft verschillende voordelen. Het gebruik van een informatieve set van genen leidt tot verbeterde prestaties van de klassificator en het voorziet de bioloog van een handelbaar aantal genen, dat kan worden onderzocht om de mechanismes van kanker te leren begrijpen. Daarnaast kan een klein aantal genen (in de ordergrootte van tientallen) leiden tot goedkopere testen, die als gebruikelijke procedures in de kliniek kunnen worden ingezet. Om de dimensionaliteit van de oorspronkelijke datasets te verkleinen, concentreert dit proefschrift zich op genselectieprocedures. Nieuwe manieren om de tumorklasse te voorspellen worden onderzocht en vergeleken met de laatste geavanceerde genselectie- en classificatieprocedures. Dit alles gebeurt binnen een nauwgezet en zorgvuldig gedefinieerd kader.

Classificatie gebaseerd op genexpressie verschaft de mogelijkheid om klassificatoren te maken met subsets van genen. Echter, voor het verkrijgen van biologisch inzicht in de mechanismes van kanker is de statistische analyse van genexpressie-arrays niet genoeg. Deze inzichten kunnen worden verkregen door de integratie van verschillende informatiebronnen; in het bijzonder copy number data, genexpressiedata en de locaties in het genoom waar deze metingen betrekking op hebben. Dit is het thema van Deel 2 van dit proefschrift. In de eerste plaats is er de analyse van de DNA copy number data zelf. Hier wordt gekeken naar de vraag of er op basis van DNA veranderingen onderscheid kan worden gemaakt tussen verschillende tumorklasses. Hiervoor is een systematische zoekstrategie ontwikkeld, die over het gehele genoom zoekt naar DNA copy number veranderingen, die specifiek zijn voor een bepaald probleem, bv. kankerclassificatie of klinische uitkomst. Het proefschrift wordt vervolgd met het onderzoek naar de invloed van veranderingen in het genoom op veranderingen in genexpressie. De ontwikkelde procedures voor de identificatie van genen, die op deze manier worden beïnvloed, worden uitvoerig besproken. De afsluitende bijdrage van deze thesis is het onderzoek naar lokale en genoombrede relaties tussen DNA verandering en verandering van de genexpressie. Hier worden afhankelijkheden binnen het genoom onderzocht door te kijken naar de correlatie tussen chromosomale afwijkingen in een bepaalde regio en de expressie op andere lokaties in het genoom.

Acknowledgments

A PhD project is a challenging learning process. Now that the process described in this thesis has reached the end, it is with great pleasure that I look back at the many things I have learned and experienced. And it is with even a greater pleasure that I want to thank here the many people that in one way or another have contributed to it.

A first thought goes to my promotor Marcel and my supervisor Lodewyk. I'd like to thank for their support and guidance throughout all four years, especially for being always responsive and ready to help, I could really count on you! The Bioinformatics group has been growing these years, becoming a nice platform for sharing of expertise and discussions, not restricted to the biolearn and biotalk meetings. A big thanks goes to Chris, Dick, Domenico, Eugene, Marco, Martin, Miranda (very happy for you and wish you wisdom and strength in the new direction you have chosen), Oliver (thanks for your help with the TFs even in what was a very busy time for you), Rogier, Wouter, Yunlei. A special thanks to my roommates Jeroen also for translating the propositions, and Theo for our discussions, taking good care of the plants and translating the Dutch summary of this thesis. I'd like to acknowledge the entire ICT group in Delft, it has been a nice environment, with lot of friendly people. In particular I'd like to thank the "Danish team" Jan and Jesper (also for bearing the outcome of couple of TV soccer matches ;)), Ioana (beside other things, I'm happy we shared the work with the students), Andrei, Peter Jan, Kosmas, Leo (you really nurtured my interest in wooden puzzles ;)), Richard, Zeki, Jun, Jenneke (you were always able to give a cheerful smile!), Ronald, Gineke, Maarten, Mark, Bartek, Umut, Hasan, Jan, Pim, Omar, Alan (your Italian has improved over these years ;)), Emile, Richard, Inald and Jan. The support staff plays an important role in making things working as smoothly as possible, thanks to Anja, Ben, Hans, Robbert and Saskia. I'd like to thank Bob and Artsiom, I enjoyed our collaboration on feature selection issues. Besides science, the consciousness discussions has been an occasion to question and share view points and experiences on consciousness related issues. It has been also very interesting to witness the shifts and development of the topics we have explored. I'm grateful to Bob, David, Dick, Ela & Andrzej, Jeroen, Marina, Mauricio, Pavel, Piotr, Sergey, Theo, Wan-Jui and Yunlei.

One of the most enjoyable aspects of my PhD research was its inter-disciplinary nature. I'm really grateful for the collaboration with the Nederlands Kanker Instituut (NKI). It has given me the opportunity to look from the inside at how the genetic cancer research is performed, something that I would not consider possible when I graduated as electronic engineer. One of the first persons I met at the NKI was Laura, whose curiosity, sharp thinking and interest inspired me to push further, thank you! In those early times I have good memories of my roommates Tako and Annuska, thank you!. I'm especially indebted to Hugo, with whom we shared the search for genomic aberrations. Besides for our long discussions, thank you for explaining and showing me many aspects of the lab work. Your openness and friendliness made it a real pleasure to collaborate with you and to share views besides our research interests. A big thanks goes to my roommates at the H6, besides Hugo, Fabien, Richard, Stephanie, Abdel and Ferdi. Fabien, you brought a warm Mediterranean atmosphere! Thanks for your enthusiasm and the valuable input to my work. I'd like to acknowledge Eric, who has been always ready to explain aCGH issues and discuss with me. Many thanks also to Petra and Marc for their interest in my research. The bioinformatics and statistics group has promoted interesting Wednesday discussions, providing a meeting place for several research groups in the institute. I acknowledge in particular Lodewyk for his coordination efforts. Besides many nice persons I met at the NKI, I'd like to thank Michael, Nicola, Arno for his help with the Perl scripts, Saske for her kindness and also her patience in the tour in one of the mouse lab, Britta, Marleen and Xiaoling for their friendliness. Dorien and Alina, it has been *very* interesting to learn about the subject of your research, I wish you find ways to continue these studies further.

To live in a different country than your own is a great experience that I would recommend to everybody, at least for a short period. Life in Delft has been an inspiring experience for me as it gave me the opportunity to meet and share ideas, experiences and view points with people from all over the world.

I'm grateful to the International Student Church, which is a living example of an ecumenical community, focused on the values and needs of the students, and tangibly open to each person as a human being, regardless of his/her religious background. In particular I'd like to acknowledge the leaders, Ben, your positive attitude and cheerfulness create constructive, inspiring and pleasant gatherings; Waltraut, your creativity, passion and sharing attitude always foster new opportunities; Avin, your cheerfulness and enthusiasm is an unexpected and most welcome gift. Many thanks to Yusuf, Anna, Julius, Nicolo‘, Arlina, Rose, Daniela, Fabio, Nelson and all the choirs members that greatly contribute to lively Sunday services. Many thanks especially to Mieke and Reini, that provide a cheerful meeting time on Thursday evening in the choir practice.

The Interfaith Sharing Group has been a great occasion for sharing views on many issues of religion and culture between people of all corners of the world. My deep gratitude goes to the “steering committee” active in 2004: Dedy, Firas, Raji, Riccardo and Yadira, for your warm and sincere friendship matured out of the many open, kind and frank meetings, especially those on very sensitive topics. Many thanks for the commitment of the new persons that are today bringing forward this opportunity: Waltraut, Mahdi, Frans, Amir, Mehdi... just to mention a few.

Developing a story, meditating on a thought patiently using colorful dots has been the lesson of the dot-painting sessions. My gratitude goes to Ben, the founder, to Sinar, Maria, Paty & Christian, Yenory, Ada, Sandrita, Susi, Susanna, Vilda, Anna, Yana, Elena, Eka & Dela, Julie-Ann, Enny ... and all the other dotters, for sharing their stories and for the cheerful atmosphere of our painting Fridays.

I'm grateful to ALL participants of the meditation sessions, some of which are Avin, Mary, Vilda, Diana and Stan, Natia, Wouter, Rebekka, Marieke, Walter, Dimostenis, Xesc, Robin, Kasia, Maja, Andres, Alessio, Carolyn, Kari & Paul, Mohana, Hari, Ladslaus, Rajesh, Navchaa... It has been a great privilege for me to share this experience with you. A special thank goes to Marisa. Working with you on the retreat has been an amazing experience of truly questioning and learning to synthesize in a constructive and joyful way our different perspective towards a common goal. How great would it be to repeat the experience!

I'm in-debt with Heleen for her teaching of yoga and for sharing with me much more then the asanas. I'm grateful I had the opportunity to meet Rolf, which is a beautiful example of a teacher and a leader, whose strengths are his gentleness and humbleness.

Despite the unreliability of the Dutch weather, we were brave enough to enjoy the Dutch nature even in the winter! My thanks goes to Ludvik, Gianluca & Diana (we shared the most rainy trekking of these last years ;)) Fatemeh, Roderik, Francesca, Denny, Iwa, Dwi, Monica, Mario and Michela, Helena, Mary, Daniele, Alessandro and all the other participants. I discovered in The Netherlands a nice Dutch custom: the sponsored walks, meant to collect money for charity purposes. The fun for me has been not only the walk itself, but also observing the reactions of people when searching for sponsorship. I've realized that the attitude with which you give is more important than what is given. I'm grateful for all those who enthusiastically sponsored me for the yearly 40 km walk in Oldenzaal.

I'd like to mentioned few more persons with whom I shared friendly dinners and pleasant moments, my thanks go to Christian, Yuval & Nili, Paul (has been a pleasure to meet you again in Vienna), Ruben, Henk, Daniela & Walter, Carlos, Yovita & Caterinus, Anna & Alex, Fabrice, Ronald, Chris, Cristina & Cris, Wouter and Jurien. A special thought goes to Jaggi, I always remember your words on responsibility, and Olga, for your openness and hospitality.

Although living in Holland, Sardinia has been always present in me, with its trees, mountains, beaches and especially people. Un grazie di cuore va ad alcuni buoni amici che sento molto vicini e presenti nonostante la distanza geografica. Grazie per la qualita' del tempo passato insieme che, nonostante limitato, non ci ha impedito di condividere i nostri percorsi. Penso a Sabrina e Salvatora, con cui abbiamo in comune i 4 anni di "specializzazione" sebbene in campi diversi ;) ; Laura con cui e' sempre un piacere condividere quello che capita: thee/natura/pizza... M.Rosaria, con cui siamo sempre riuscite a incontrarci fosse anche solo per un caffe al volo a ridosso della 131; Davide, con cui abbiamo un interesse comune che sarebbe bellissimo approfondire insieme; Massi, sempre pronto ad organizzare un trekking nei miei rietri in Sardegna; Sergio e Dany, anche se siete dall'altra parte del mondo (bellissimi gli sms dalla Nuova Zelanda!); Maria, con cui potremmo discutere per ore...; Barbara, cui auguro un mondo di bene nel nuovo capitolo appena cominciato; Aurora e Maria Grazia "disperse" per il meglio a Rimini e Londra ;); Giulio & Ilaria. Un grazie in particolare a p.Piras e p.Neudeker che nonostante i rari incontri sento presenti nel mio percorso, e non solo per scambio di e-mail!

Vivere in Olanda ha significato scegliere di stare geograficamente lontano dalla mia famiglia. Ringrazio i miei genitori Mario e Elena per il vostro supporto, vicinanza telefonica e pazienza nell' accettare le mie scelte che non sempre hanno un senso per voi...;) Un grazie speciale a nonna Zelinda per essere, dall'alto dei suoi 92 anni, un esempio di saggezza e apertura al nuovo. Ringrazio i miei fratellini, Roberto, sempre disponibile a raccattarci dall'aeroporto, Antonello, con le sue e-mail incredibilmente loquaci ma efficaci e precise ;) e Paolo, cui auguro tanta gioia e serenita' con Marta. Un grazie a Franca, per la tua generosita' e il tuo ottimismo particolare, Maria e Giovanna, che ultimamente hanno fatto i piccioni viaggiatori piu' di me ;), Caterina, con le sue torte virtuali e reali ;), Paolo, Biagina, Teresa e Tore, Luciana e Marco, Salvatore e Elisabetta, Antonello, Vincenzo e Lina, e tutti i cuginetti.

I've been appreciating more and more along these years Czechia, with its sweet hills, forests and castles. My gratitude goes to Věra and Vladimír for their kindness, interest and support, děkuji, děkuji! I'm especially grateful to Pavel for his love, patience, trust, unconditional support and smiles; it has been a pleasure to witness how our relashionship has been changing and improving with us along those years... Siamo un team ;)

Finally, I'm grateful to Life and her teachers, present and future ...

Publications

Journal papers

H. Horlings, C. Lai, P. Kristel, E. van Beers, C. Klijn, S. Joosse, F. Reyal, D. Nuyten, C.J. Froyland, A. Borresen-Dale, M.J.T. Reinders, P. Nederlof, L.F.A. Wessels, and M.J. van de Vijver. *Integration of dna copy number alterations and prognostic gene signatures to predict prognosis of patients with breast cancer.* manuscript in preparation.

C. Lai, H.M. Horlings, M.J. van de Vijver, E.H. van Beers, P.M. Nederlof, L.F.A. Wessels, and M.J.T. Reinders. *Sirac: Supervised identification of regions of aberration in acgh datasets.* BMC Bioinformatics, 8:422, 2007.

A. Harol, C. Lai, E. Pekalska, and R.P.W. Duin. *Pairwise evaluation for the selection of features and prototypes.* Pattern Analysis and Applications, 10(1):55-67, 2007.

C. Lai, M.J.T. Reinders, L.J. van't Veer, and L.F.A. Wessels. *A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets.* BMC Bioinformatics, 7(235), 2006.

C. Lai, M.J.T. Reinders, and L.F.A. Wessels. *Random subspace method for multivariate feature selection.* Patter Recognition Letters, 27(10):1067-1076, 2006.

C. Lai, D.M.J. Tax, R.P.W. Duin, E. Pekalska, and P. Paclik. *On combining image representations for image classification and retrieval.* International Journal of Pattern Recognition and Artificial Intelligence, 18(5):867-890, 2004.

Conference papers

C. Lai, M.J.T. Reinders, and L. Wessels.*A study on multivariate gene selection for cancer classification.* In 12th Annual Conference of the Advanced School for Computing and Imaging., pages 236-244, The Netherlands, 2006.

C. Lai, M.J.T. Reinders, and L.F.A. Wessels. *Multivariate gene selection: Does it help?* In IEEE Computational Systems Biology Conference, Stanford, California, 2005.

C. Lai, M.J.T. Reinders, and L. Wessels. *Random subspace method for multivariate feature selection.* In 11th Annual Conference of the Advanced School for Computing and Imaging., pages 328-335, The Netherlands, 2005.

E. Pekalska, A. Harol, C. Lai, and R.P.W. Duin. *Pairwise selection of features and prototypes.* In Proc. of 4th Int. Conf. on Computer Recognition Systems CORES 05, pages 271-278, Poland, 2005. Springer.

C. Lai, M.J.T. Reinders, and L. Wessels. *On univariate selection methods in gene expression datasets.* In Tenth Annual Conference of the Advanced School for Computing and Imaging., pages 335-341, The Netherlands, 2004.

C. Lai, D.M.J. Tax, R.P.W. Duin, E. Pekalska, and P. Paclik. *Database retrieval: the use of combined dissimilaities.* In 9th Annual Conference of the Advanced School for Computing and Imaging., pages 177-184, The Netherlands, 2003.

C. Lai, D.M.J. Tax, R.P.W. Duin, E. Pekalska, and P. Paclik. *On combining one-class classifiers for image database retrieval.* In 3rd international workshop on multiple classifier systems, pages 212-221, Cagliari, Italy, 2002.

Curriculum Vitae

Carmen Lai was born on 17 February 1973 in Nuoro, Sardegna, Italy. She obtained the diploma from the Scientific Lyceum "Enrico Fermi", Nuoro in 1992, after which she started her studies in Electronic Engineering at Cagliari University. From September 2001 till April 2002 she worked for her graduation project, in the context of an Erasmus exchange, in the Pattern Recognition Group, at TU Delft, under supervision of Dr.ir. R.P.W. Duin and Dr. D.M.J. Tax. The topic of the research was *image database retrieval using user feedback*. In June 2002 she obtained her M.Sc. degree from Cagliari University. She came back to The Netherlands where she worked as a research fellow, from September 2002 till March 2003, in the Pattern Recognition Group (today Quantitative Imaging Group), at TU Delft. In April 2003 she started her PhD study in the ICT group on pattern recognition applications in cancer research. Her PhD was sponsored and performed in collaboration with the Molecular Biology and Pathology divisions of the Dutch Cancer Institute in Amsterdam. She has been working as a reviewer for Pattern Recognition Letters, Pattern Analysis and Applications, The British Journal of Cancer. She has recently joined PR Sys Design consulting company, developing custom pattern recognition systems for industrial applications.